Statistical Learning Theory

Collection Editor:
Robert Nowak

Statistical Learning Theory

Collection Editor:
Robert Nowak

Authors:
Rui Castro
Robert Nowak
Aarti Singh

Online:
< http://cnx.org/content/col10532/1.3/ >

C O N N E X I O N S

Rice University, Houston, Texas

Collection structure revised: April 10, 2009
PDF generated: February 4, 2011
For copyright and attribution information for the modules contained in this collection, see p. 163.

Table of Contents

Chapter 1

Basic Elements of Statistical Decision Theory and Statistical Learning Theory[1]

Throughout this module, let X denote the **input** to a decision-making process and Y denote the correct response or **output** (e.g., the value of a parameter, the label of a class, the signal of interest). We assume that X and Y are random variables or random vectors with joint distribution $P_{X,Y}(x,y)$, where x and y denote specific values that may be taken by the random variables X and Y, respectively. The observation X is used to make decisions pertaining to the quantity of interest. For the purposes of illustration, we will focus on the task of determining the value of the quantity of interest. A decision rule for this task is a function f that takes the observation X as input and outputs a prediction of the quantity Y. We denote a decision rule by $\hat{Y}$ or $f(X)$, when we wish to indicate explicitly the dependence of the decision rule on the observation. We will examine techniques for designing decision rules and for analyzing their performance.

1.1 Measuring Decision Accuracy: Loss and Risk Functions

The accuracy of a decision is measured with a loss function. For example, if our goal is to determine the value of Y, then a loss function takes as inputs the true value Y and the predicted value (the decision) $\hat{Y} = f(X)$ and outputs a non-negative real number (the "loss") reflective of the accuracy of the decision. Two of the most commonly encountered loss functions include:

1. 0/1 loss: $\ell_{0/1}\left(\hat{Y}, Y\right) = \mathbf{I}_{\hat{Y} \neq Y}$, which is the indicator function taking the value of 1 when $\hat{Y} \neq Y$ and taking the value 0 when $\hat{Y}(X) = Y$.
2. squared error loss: $\ell_2\left(\hat{Y}, Y\right) = \|\hat{Y} - Y\|_2^2$, which is simply the sum of squared differences between the elements of $\hat{Y}$ and Y.

The 0/1 loss is commonly used in detection and classification problems, and the squared error loss is more appropriate for problems involving the estimation of a continuous parameter. Note that since the inputs to the loss function may be random variables, so is the loss.

A risk $R(f)$ is a function of the decision rule f, and is defined to be the expectation of a loss with respect to the joint distribution $P_{X,Y}(x,y)$. For example, the expected 0/1 loss produces the **probability of error**

[1] This content is available online at <http://cnx.org/content/m16263/1.3/>.

risk function; i.e., a simply calculation shows that $R_{0/1}(f) = E\left[\left(\mathbf{I}_{f(X)\neq Y}\right] = \Pr(f(X) \neq Y)\right.$. The expected squared error loss produces the **mean squared error** MSE risk function, $R_2(f) = E\left[\| f(X) - Y \|_2^2\right]$.

Optimal decisions are obtained by choosing a decision rule f that minimizes the desired risk function. Given complete knowledge of the probability distributions involved (e.g., $P_{X,Y}(x,y)$) one can explicitly or numerically design an optimal decision rule, denoted f^*, that minimizes the risk function.

1.2 The Maximum Likelihood Principle

The conditional distribution of the observation X given the quantity of interest Y is denoted by $P_{X|Y}(x|y)$. The conditional distribution $P_{X|Y}(x|y)$ can be viewed as a generative model, probabilistically describing the observations resulting from a given value, y, of the quantity of interest. For example, if y is the value of a parameter, the $P_{X|Y}(x|y)$ is the probability distribution of the observation X when the parameter value is set to y. If X is a continuous random variable with conditional density $p_{X|Y}(x|y)$ or a discrete random variable with conditional probability mass function (pmf) $p_{X|Y}(x|y)$, then given a value y we can assess the probability of a particular measurment value y by the magnitude of either the conditional density or pmf.

In decision making problems, we know the value of the observation, but do not know the value y. Therefore, it is appealing to consider the conditional density or pmf as a function of the unknown values y, with X fixed at its observed value. The resulting function is called the likelihood function. As the name suggests, values of y where the likelihood function is largest are intuitively reasonable indicators of the true value of the unknown quantity, which we will denote by y^*. The rationale for this is that these values would produce conditional densities or pmfs that place high probability on the observation $X = x$.

The Maximum Likelihood Estimator (MLE) is defined to be the value of y that maximizes the likelihood function; i.e., in the continuous case

$$\hat{y}(X) = \underset{y}{argmax} p_{X|Y}(X|y) \tag{1.1}$$

with an analogous definition for the discrete case by replacing the conditional density with the conditional pmf. The decision rule $\hat{y}(X)$ is called an "estimator," which is common in decision problems involving a continuous parameter. Note that maximizing the likelihood function is equivalent to minimizing the negative log-likelihood function (since the logarithm is a monotonic transformation). Now let y^* denote the true value of Y. Then we can view the negative log-likelihood as a loss function

$$\ell_L(y, y^*) = -\log p_{X|Y}(X|y) \tag{1.2}$$

where the dependence on y^* on the right hand side is embodied in the observation X on the left. An interesting special case of the MLE results when the conditional density $P_{X|Y}(X|y)$ is a Gaussian, in which case the negative log-likelihood corresponds to a squared error loss function.

Now let us consider the expectation of this loss, with respect to the conditional distribution $P_{X|Y}(X|y^*)$:

$$-E\left[\log p_{X|Y}(X|y)\right] = \int \log\left(\frac{1}{p_{X|Y}(x|y)}\right) p_{X|Y}(x|y^*)\, dx \tag{1.3}$$

The true value y^* minimizes the expected negative log-likelihood (or, equivalently, maximizes the expected log-likelihood). To see this, compare the expected log-likelihood of y^* with that of any other value y:

$$\begin{aligned} E\left[\log p_{X|Y}(X|y^*) - \log p_{X|Y}(X|y)\right] &= E\left[\log\left(\frac{p_{X|Y}(X|y^*)}{p_{X|Y}(X|y)}\right)\right] \\ &= \int \log\left(\frac{p_{X|Y}(x|y^*)}{p_{X|Y}(x|y)}\right) p_{X|Y}(x|y^*)\, dx \\ &= \mathrm{KL}\left(p_{X|Y}(x|y^*), p_{X|Y}(x|y)\right) \end{aligned} \tag{1.4}$$

The quantity $\mathrm{KL}\left(p_{X|Y}(x|y^*), p_{X|Y}(x|y)\right)$ is called the Kullback-Leibler (KL) divergence between the conditional density function $p_{X|Y}(x|y^*)$ and $p_{X|Y}(x|y)$. The KL divergence is non-negative, and zero if and

only if the two densities are equal [1]. So, we see that the KL divergence acts as a sort of risk function in the context of Maximum Likelihood Estimation.

1.3 The Cramer-Rao Lower Bound

The MLE is based on finding the value for Y that maximizes the likelihood function. Intuitively, if the maximum point is very distinct, say a well isolated peak in the likelihood function, then the easier it will be to distinguish the MLE from alternative decisions. Consider the case in which Y is a scalar quantity. The "peakiness" of the log-likelihood function can be gauged by examining its curvature, $-\frac{\partial^2 log p_{X|Y}(x|y)}{\partial y^2}$, at the point of maximum likelihood. The higher the curvature, the more peaky is the behavior of the likelihood function at the maximum point. Of course, we hope that the MLE will be a good predictor (decision) for the unknown true value y^*. So, rather than looking at the curvature of the log-likelihood function at the maximum likelihood point, a more appropriate measure of how easily it will be to distinguish y^* from the alternatives is the expected curvature of the log-likelihood function evaluated at the value y^*. The expectation taken over all possible observations with respect to the conditional density $p_{X|Y}(x|y^*)$. This quantity, denoted $I(y^*) = E\left[-\frac{\partial^2 log p_{X|Y}(x|y)}{\partial y^2}\right]|_{y=y^*}$, is called the Fisher Information (FI). In fact, the FI provides us with an important performance bound known as the Cramer-Rao Lower Bound (CRLB).

The CRLB states that under some mild regularity assumptions about the conditional density function $p_{X|Y}(x|y)$, the variance of any unbiased estimator is bounded from below by the inverse of the $I(y^*)$[5], [4], [3]. Recall that an unbiased estimator is any estimator $\hat{Y}$ that satisfies $E\left[\hat{Y}\right] = y^*$. The CRLB tells us is that

$$\text{var}\left(\hat{Y}\right) \geq \frac{1}{I(y^*)}. \tag{1.5}$$

If Y is a vector-valued quantity, then the expected negative Hessian matrix (matrix of partial second derivatives) of the log-likelihood function is called the Fisher Information Matrix (FIM), and a similar inequality tells us that the variance of each component of any unbiased estimator of y^* is bounded below by the corresponding diagonal element of the inverse of the FIM. Since the MSE of an unbiased estimator is equal to its variance, we see that the CRLB provides a very useful lower bound on the best MSE performance that we can hope to achieve. Thus, the CRLB is often used as a comparison point for evaluating estimators. It may or may not be possible to achieve the CRLB, but if we find a decision rule that does, we know that it also minimizes the MSE risk among all possible unbiased estimators. In general, it may be difficult to compute the CRLB, but in certain important cases it is possible to find closed-form or computational solutions.

1.4 Bayesian Decision Theory

Bayesian Decision Theory provides a formal system for integrating prior knowledge and observed observations. For the purposes of illustration we will focus on problems involving continuous variables and observations, but extensions to discrete cases are straightforward (simple replace probability densities with probability mass functions, and integrals with summations). The key elements of Bayesian methods are:

1. a prior probability density function $p_Y(y)$ describing a priori knowledge of probable states for the quantity Y;
2. the likelihood function $p_{X|Y}(x|y)$, as described above;
3. the posterior density function $p_{Y|X}(y|x)$.

The posterior density is a function of the prior and likelihood, obtained according to Bayes rule:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)\, p_Y(y)}{\int p_{X|Y}(x|y)\, p_Y(y)\, dy}. \tag{1.6}$$

The posterior is an indicator of probable values for Y, based on the prior knowledge and the observation. Several options exist for deriving a specific estimate of Y using the posterior. The mean value of the posterior density is one common choice (commonly called the **posterior mean**). The posterior mean is the decision rule that minimizes the expected squared error loss (MSE risk) function. The value y where the posterior density is maximized is another popular estimator (commonly called the **Maximum A Posteriori** (MAP) estimator). Note that the denominator of the posterior is independent of y, so the MAP estimator is simply the maximizer of the product of the likelihood and the prior. Therefore, if the prior is a constant function, the MAP estimator and MLE coincide.

1.5 Statistical Learning

In all of the methods described above, we assumed some amount of knowledge about the distributions of the observation X and quantity of interest Y. Such knowledge can come from a careful analysis of the physical characteristics of the problem at hand, or it can be gleaned from previous experience. However, there are situations where it is difficult to model the physics of the problem and we may not have enough experience to develop complete and accurate probability models. In such cases, it is natural to adopt a **statistical learning** approach [2], [7].

Statistical learning methods are based on developing decision rules or estimators based only on a collection of training examples, rather than predetermined probability models. Statistical learning methods are often said to be **distribution-free**, since they do not assume particular probability models. The canonical set-up for statistical learning is as follows. We begin with a collection of training examples, $\{(X_i, Y_i)\}_{i=1}^n$, which are assumed to be independently and identically distributed according to an **unknown** probability distribution $P_{X,Y}(x, y)$. If we knew $P_{X,Y}(x, y)$, then we could compute a desired risk function and design an optimal decision rule using the methods described above. In essence, the training examples give us a glimpse at the underlying distribution, but our knowledge of it is far from complete. We cannot exactly compute a risk function, and therefore we cannot derive a corresponding optimal decision rule.

There are at least two ways to proceed at this point. One possibility is to use the training examples to estimate the joint probability distribution, and then use this estimate to derive an decision rule. Unfortunately, the (general-purpose) problem of estimating a distribution is often more difficult from a limited pool of data than is the problem of designing a specific-purpose decision rule. For this reason, a second possibility is more commonly favored in practice. Rather than estimating the complete distribution, one can use the training examples to directly design a decision rule. More precisely, perhaps the most common approach is to use the training examples to compute an estimate of the desired risk function.

Suppose that we are interested in minimizing a particular risk function. Recall that the risk is the expected value of a chosen loss function. Let $\ell\left(\hat{Y}, Y\right)$ denote the loss, and let $f(X)$ denote a candidate decision function, mapping observations to predictions about Y (i.e., $\hat{Y} = f(X)$). The **empirical risk function** is constructed from the training examples as follows:

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f(X_i), Y_i\right). \tag{1.7}$$

This is simply the average loss of the decision rule f over the set of training examples. Note that since the training examples are independent and identically distributed, the expected value of the empirical risk is equal to the true risk $R(f) = E\left[\ell\left(f(X), Y\right)\right]$. Moreover, we known (according to the law of large numbers)

that the empirical risk tends to the true risk as the size of the training sample increases. These facts lend support to the idea of choosing a decision rule to minimize the empirical risk.

Empirical risk minimization (ERM) is just this process. Given a collection of possible decision rules, say $\mathcal{F}$, ERM selects a decision rule according to

$$\hat{f}_n \;=\; \underset{f\in\mathcal{F}}{argmin}\,\hat{R}\,(f)\,. \tag{1.8}$$

The selected rule, $\hat{f}_n$, obviously depends on the given set of training examples, and therefore it is itself a random quantity. The theoretically optimal counterpart to $\hat{f}_n$ is the decision rule that minimizes the true risk

$$f^* \;=\; \underset{f\in\mathcal{F}}{argmin} R\,(f)\,. \tag{1.9}$$

The central problem in statistical learning is to quantify how close $\hat{f}_n$ performs relative to f^*. Note that $R\,(f^*) \leq R\left(\hat{f}_n\right)$, since f^* minimizes the true risk. Thus, one way to gauge the performance of $\hat{f}_n$ relative to f^* is to show that there exists small positive values ε and δ such that with probability at least $1-\delta$ we have

$$R\left(\hat{f}_n\right) \;\leq\; R\,(f^*) + \varepsilon. \tag{1.10}$$

If an inequality of this form holds, then we say that $\hat{f}_n$ is a **Probability Approximately Correct** (PAC) decision rule [6].

To show that the empirical risk minimizer is a PAC decision rule, we first must understand how closely the empirical risk matches the true risk. First, let us consider the empirical and true risk of the decision rule f. Assume that the loss function is bounded between 0 and 1 (possibly after a suitable normalization). Then the empirical risk function is a sum of independent random variables bounded between 0 and 1. Hoeffding's inequality is a bound on the deviations of such random sums from their corresponding mean values [2]. In this case, the mean value is the true risk of f, and Hoeffding's inequality states that

$$P\left(\mid \hat{R}\,(f) - R\,(f)\mid > \varepsilon\right) \;\leq\; 2e^{-2n\varepsilon^2}. \tag{1.11}$$

Another equivalent statement is that the inequality $\mid \hat{R}\,(f) - R\,(f)\mid \leq \varepsilon$ holds with probability at least $1-2e^{-2n\varepsilon^2}$. Thus, the two risks are probably close together, and the greater the number of training examples, n, the closer they are.

Now we would like a similar condition to hold for all $f \in \mathcal{F}$, since ERM optimizes over the entire collection $\mathcal{F}$. Suppose that $\mathcal{F}$ is a finite collection of decision rules. Let $|\mathcal{F}|$ denote the number of rules in $\mathcal{F}$. The probability that the difference between the true and empirical risks, of one or more of the decision rules, exceeds ε is bounded by the sum of the probabilities of each individual event of the form $\mid \hat{R}\,(f) - R\,(f)\mid > \varepsilon$, the so-called **Union of Events** bound. Therefore, with probability at least $1 - |\mathcal{F}|2e^{-2n\varepsilon^2}$ we have that

$$\mid \hat{R}\,(f) - R\,(f)\mid \leq \varepsilon \tag{1.12}$$

for all $f \in \mathcal{F}$. Equivalently, setting $\delta = 2|\mathcal{F}|e^{-2n\varepsilon^2}$, we have that with probability at least $1-\delta$ and for all $f \in \mathcal{F}$

$$| \hat{R}(f) - R(f) | \leq \sqrt{\frac{log|\mathcal{F}| + log(2/\delta)}{2n}}. \tag{1.13}$$

Notice that the two risks are uniformly close together, and the closeness indicated by the bound increases as n increases and decreases as the number of decision rules in $\mathcal{F}$ increases. In fact, the bound scales with $log|\mathcal{F}|$, and so it is reasonable to interpret the logarithm of the number of decision rules under consideration as a measure of the **complexity** of the class.

Now using this bound, we can show that $\hat{f}_n$ is a PAC decision rule as follows. Note that with probability at least $1-\delta$

$$\begin{aligned} R\left(\hat{f}_n\right) &\leq \hat{R}\left(\hat{f}_n\right) + \sqrt{\tfrac{log|\mathcal{F}|+log(2/\delta)}{2n}} \\ &\leq \hat{R}(f^*) + \sqrt{\tfrac{log|\mathcal{F}|+log(2/\delta)}{2n}} \\ &\leq R(f^*) + 2\sqrt{\tfrac{log|\mathcal{F}|+log(2/\delta)}{2n}} \end{aligned} \tag{1.14}$$

where the first inequality follows since the true and empirical risks are close for all $f \in \mathcal{F}$, and in particular for $\hat{f}_n$, the second inequality holds since by definition $\hat{f}_n$ minimizes the empirical risk, and the third inequality holds again since the empirical risk is close to the true risk for all f, in this case for f^* in particular. So, we have shown that $\hat{f}_n$ is PAC.

PAC bounds of this form can be extended in many directions, for example to infinitely large or uncountable classes of decision rules, but the basic ingredients of the theory are essentially like those demonstrated above. The bottom line is that empirical risk minimization is a reasonable approach, provided one has access to a sufficient number of training examples and the number, or more generally the complexity, of the class of decision rules under consideration is not too great.

1.6 Further reading

Excellent treatments of classical decision and estimation theory can be found in a number of textbooks [5], [4], [3], [1]. For references on statistical learning theory, outstanding textbooks are also available [2], [7], [6] for further reading.

Chapter 2

Elements of Statistical Learning Theory[1]

2.1 Three Elements of Statistical Data Analysis

1. **Probabilistic Formulation:** of learning from data and prediction problems.
2. **Performance Characterization::**
   - concentration inequalities
   - uniform deviation bounds
   - approximation theory
   - rates of convergence
3. **Practical Algorithms:** that run in polynomial time (e.g., decision trees, wavelet methods, support vector machines).

2.2 Learning from Data

To formulate the basic learning from data problem, we must specify several basic elements: data spaces, probability measures, loss functions, and statistical risk.

2.2.1 Data Spaces

Learning from data begins with a specification of two spaces:

$$\mathcal{X} \equiv \text{Input Space} \tag{2.1}$$

$$\mathcal{Y} \equiv \text{Output Space.} \tag{2.2}$$

The input space is also sometimes called the "feature space" or "signal domain." The output space is also called the "class label space," "outcome space," "response space," or "signal range."

Example 2.1

$$\mathcal{X} = \mathbf{R}^d \quad d\text{-dimensional Euclidean space of "feature vectors"} \tag{2.3}$$

$$\mathcal{Y} = \{0, 1\} \quad \text{two classes or "class labels"} \tag{2.4}$$

[1] This content is available online at <http://cnx.org/content/m16269/1.2/>.

Example 2.2

$$\mathcal{X} = \mathbf{R} \quad \text{one-dimensional signal domain (e.g., time-domain)} \tag{2.5}$$

$$\mathcal{Y} = \mathbf{R} \quad \text{real-valued signal} \tag{2.6}$$

A classic example is estimating a signal f in noise:

$$Y = f(X) + W \tag{2.7}$$

where X is a random sample point on the real line and W is a noise independent of X.

2.2.2 Probability Measure and Expectation

Define a joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ denoted $P_{X,Y}$. Let (X, Y) denote a pair of random variables distributed according to $P_{X,Y}$. We will also have use for marginal and conditional distributions. Let P_X denote the marginal distribution on X, and let $P_{Y|X}$ denote the conditional distribution of Y given X. For any distribution P, let p denote its density function with respect to the corresponding dominating measure; e.g., **Lebesgue measure** for continuous random variables or **counting measure** for discrete random variables.

Define the expectation operator:

$$E_{X,Y}[f(X,Y)] \equiv \int f(x,y)\, dP_{X,Y}(x,y) = \int f(x,y)\, p_{X,Y}(x,y)\, dxdy. \tag{2.8}$$

We will also make use of corresponding marginal and conditional expectations such as E_X and $E_{Y|X}$.

Wherever convenient and obvious based on context, we may drop the subscripts (e.g., E instead of $E_{X,Y}$) for notational ease.

2.2.3 Loss Functions

A loss function is a mapping

$$\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbf{R}. \tag{2.9}$$

Example 2.3
In binary classification problems, $\mathcal{Y} = \{0, 1\}$. The 0/1 loss function is usually used: $\ell(y_1, y_2) = 1_{y_1 \neq y_2}$, where 1_A is the indicator function which takes a value of 1 if condition A is true and zero otherwise. We typically will compare a true label y with a prediction $\hat{y}$, in which case the 0/1 loss simply counts misclassifications.

Example 2.4
In regression or estimation problems, $\mathcal{Y} = R$. The squared error loss function is often employed: $\ell(y_1, y_2) = (y_1 - y_2)2$, the square of the difference between y_1 and y_2. In application, we are interested in a true value y in comparison to an estimate $\hat{y}$.

2.2.4 Statistical Risk

The basic problem in learning is to determine a mapping $f : \mathcal{X} \mapsto \mathcal{Y}$ that takes an input $x \in \mathcal{X}$ and predicts the corresponding output $y \in \mathcal{Y}$. The performance of a given map f is measured by its expected loss or **risk**:

$$R(f) \equiv E_{X,Y}\left[\ell\left(f\left(X\right), Y\right)\right]. \tag{2.10}$$

The risk tells us how well, on average, the predictor f performs with respect to the chosen loss function. A key quantity of interest is the mininum risk value, defined as

$$R^* = \inf_f R(f) \tag{2.11}$$

where the infinum is taking over all measurable functions.

2.2.5 The Learning Problem

Suppose that (X, Y) are distributed according to $P_{X,Y}$ ($(X, Y) \sim P_{X,Y}$ for short). Our goal is to find a map so that $f(X) \approx Y$ with high probability. Ideally, we would chose f to minimize the risk $R(f) = E\left[\ell\left(f\left(X\right), Y\right)\right]$. However, in order to compute the risk (and hence optimize it) we need to know the joint distribution $P_{X,Y}$. In many problems of practical interest, the joint distribution is unknown, and minimizing the risk is not possible.

Suppose that we have some exemplary samples from the distribution. Specifically, consider n samples $X_i, Y_i{}_{i=1}^n$ distributed independently and identically (iid) according to the otherwise unknown $P_{X,Y}$. Let us call these samples **training data**, and denote the collection by $D_n \equiv X_i, Y_i{}_{i=1}^n$. Let's also define a collection of candidate mappings $\mathcal{F}$. We will use the training data D_n to pick a mapping $f_n \in \mathcal{F}$ that we hope will be a good predictor. This is sometimes called the **Model Selection** problem. Note that the selected model f_n is a function of the training data:

$$f_n(X) = f\left(X; D_n\right), \tag{2.12}$$

which is what the subscript n in f_n refers to. The risk of f_n is given by

$$R(f_n) = E_{X,Y}\left[\ell\left(f_n\left(X\right), Y\right)\right]. \tag{2.13}$$

Note that since f_n depends on D_n in addition to a new random pair (X, Y), the risk is a random variable (i.e., a function of the training data D_n). Therefore, we are interested in the **expected risk**, computed over random realizations of the training data:

$$E_{D_n}\left[R(f_n)\right]. \tag{2.14}$$

We hope that f_n produces a small expected risk.

The notion of expected risk can be interpreted as follows. We would like to define an algorithm (a model selection process) that performs well on average, over any random sample of n training data. The expected risk is a measure of the expected performance of the algorithm with respect to the chosen loss function. That is, we are not gauging the risk of a particular map $f \in \mathcal{F}$, but rather we are measuring the performance of the algorithm that takes any realization of training data and selects an appropriate model in $\mathcal{F}$.

This course is concerned with determining "good" model spaces $\mathcal{F}$ and useful and effective model selection algorithms.

Chapter 3

Introduction to Classification and Regression[1]

3.1 Pattern Classification

Recall that the goal of classification is to learn a mapping from the feature space, $\mathcal{X}$, to a label space, $\mathcal{Y}$. This mapping, f, is called a **classifier**. For example, we might have

$$\begin{aligned} \mathcal{X} &= \mathbf{R}^d \\ \mathcal{Y} &= \{0,1\}. \end{aligned} \tag{3.1}$$

We can measure the loss of our classifier using $0-1$ loss; **i.e.,**

$$\ell\left(\hat{y},y\right)=\mathbf{1}_{\{\hat{y}\neq y\}}=\begin{cases} 1, & \hat{y}\neq y \\ 0, & \hat{y}=y \end{cases}. \tag{3.2}$$

Recalling that risk is defined to be the expected value of the loss function, we have

$$R(f)=E_{XY}\left[\ell\left(f\left(X\right),Y\right)\right]=E_{XY}\left[\mathbf{1}_{\{f(X)\neq Y\}}\right]=P_{XY}\left(f\left(X\right)\neq Y\right). \tag{3.3}$$

The performance of a given classifier can be evaluated in terms of how close its risk is to the Bayes' risk.

Definition 3.1: (Bayes' Risk)
The Bayes' risk is the infimum of the risk for all classifiers:

$$R^*=\inf_f R(f). \tag{3.4}$$

We can prove that the Bayes risk is achieved by the Bayes classifier.

Definition 3.2: Bayes Classifier
The Bayes classifier is the following mapping:

$$f^*(x)=\begin{cases} 1, & \eta(x)\geq 1/2 \\ 0, & otherwise \end{cases} \tag{3.5}$$

where

$$\eta(x)\equiv P_{Y|X}\left(Y=1|X=x\right). \tag{3.6}$$

[1] This content is available online at <http://cnx.org/content/m16272/1.2/>.

Note that for any x, $f^*(x)$ is the value of $y \in \{0,1\}$ that maximizes $P_{XY}(Y=y|X=x)$.

Theorem 3.1: Risk of the Bayes Classifier

$$R(f^*) = R^*. \tag{3.7}$$

Proof:
Let $g(x)$ be any classifier. We will show that

$$P(g(X) \neq Y|X=x) \geq P(f^*(x) \neq Y|X=x). \tag{3.8}$$

For any g,

$$\begin{aligned} P(g(X) \neq Y|X=x) &= 1 - P(Y=g(X)|X=x) \\ &= 1 - [P(Y=1,g(X)=1|X=x) + P(Y=0,g(X)=0|X=x)] \\ &= 1 - \left[E\left[\mathbf{1}_{\{Y=1\}}\mathbf{1}_{\{g(X)=1\}}|X=x\right] + E\left[\mathbf{1}_{\{Y=0\}}\mathbf{1}_{\{g(X)=0\}}|X=x\right]\right] \\ &= 1 - \left[\mathbf{1}_{\{g(x)=1\}}E\left[\mathbf{1}_{\{Y=1\}}|X=x\right] + \mathbf{1}_{\{g(x)=0\}}E\left[\mathbf{1}_{\{Y=0\}}|X=x\right]\right] \\ &= 1 - \left[\mathbf{1}_{\{g(x)=1\}}P(Y=1|X=x) + \mathbf{1}_{\{g(x)=0\}}P(Y=0|X=x)\right] \\ &= 1 - \left[\mathbf{1}_{\{g(x)=1\}}\eta(x) + \mathbf{1}_{\{g(x)=0\}}(1-\eta(x))\right] \end{aligned}. \tag{3.9}$$

Next consider the difference

$$\begin{aligned} &P(g(x) \neq Y|X=x) - P(f^*(x) \neq Y|X=x) \\ &= \\ &\eta(x)\left[\mathbf{1}_{\{f^*(x)=1\}} - \mathbf{1}_{\{g(x)=1\}}\right] + (1-\eta(x))\left[\mathbf{1}_{\{f^*(x)=0\}} - \mathbf{1}_{\{g(x)=0\}}\right] \\ &= \\ &\eta(x)\left[\mathbf{1}_{\{f^*(x)=1\}} - \mathbf{1}_{\{g(x)=1\}}\right] - (1-\eta(x))\left[\mathbf{1}_{\{f^*(x)=1\}} - \mathbf{1}_{\{g(x)=1\}}\right] \\ &= \\ &(2\eta(x)-1)\left(\mathbf{1}_{\{f^*(x)=1\}} - \mathbf{1}_{\{g(x)=1\}}\right), \end{aligned} \tag{3.10}$$

where the second equality follows by noting that $\mathbf{1}_{\{g(x)=0\}} = 1 - \mathbf{1}_{\{g(x)=1\}}$. Next recall

$$f^*(x) = \begin{cases} 1, & \eta(x) \geq 1/2 \\ 0, & otherwise \end{cases}. \tag{3.11}$$

For x such that $\eta(x) \geq 1/2$, we have

$$\underbrace{(2\eta(x)-1)}_{\geq 0}\underbrace{\left(\underbrace{\mathbf{1}_{\{f^*(x)=1\}}}_{1} - \underbrace{\mathbf{1}_{\{g(x)=1\}}}_{0 or 1}\right)}_{\geq 0} \tag{3.12}$$

and for x such that $\eta(x) < 1/2$, we have

$$\underbrace{(2\eta(x)-1)}_{<0}\underbrace{\left(\underbrace{\mathbf{1}_{\{f^*(x)=1\}}}_{0} - \underbrace{\mathbf{1}_{\{g(x)=1\}}}_{0 or 1}\right)}_{\leq 0}, \tag{3.13}$$

which implies

$$(2\eta(x)-1)\left(\mathbf{1}_{\{f^*(x)=1\}}-\mathbf{1}_{\{g(x)=1\}}\right)\geq 0 \tag{3.14}$$

or

$$P(g(X)\neq Y|X=x)\geq P(f^*(x)\neq Y|X=x). \tag{3.15}$$

Note that while the Bayes classifier achieves the Bayes risk, in practice this classifier is not realizable because we do not know the distribution P_{XY} and so cannot construct $\eta(x)$.

3.2 Regression

The goal of regression is to learn a mapping from the input space, $\mathcal{X}$, to the output space, $\mathcal{Y}$. This mapping, f, is called a **estimator**. For example, we might have

$$\begin{aligned}\mathcal{X} &= \mathbf{R}^d\\ \mathcal{Y} &= \mathbf{R}.\end{aligned} \tag{3.16}$$

We can measure the loss of our estimator using squared error loss; **i.e.,**

$$\ell\left(\hat{y},y\right)=\left(y-\hat{y}\right)^2. \tag{3.17}$$

Recalling that risk is defined to be the expected value of the loss function, we have

$$R(f)=E_{XY}\left[\ell\left(f(X),Y\right)\right]=E_{XY}\left[\left(f(X)-Y\right)^2\right]. \tag{3.18}$$

The performance of a given estimator can be evaluated in terms of how close the risk is to the infimum of the risk for all estimator under consideration:

$$R^*=\inf_f R(f). \tag{3.19}$$

Theorem 3.2: Minimum Risk under Squared Error Loss (MSE)
Let $f^*(x)=E_{Y|X}[Y|X=x]$

$$R(f^*)=R^*. \tag{3.20}$$

Proof:

$$\begin{aligned}
R(f) &= E_{XY}\left[(f(X)-Y)^2\right] \\
&= E_X\left[E_{Y|X}\left[(f(X)-Y)^2|X\right]\right] \\
&= E_X\left[E_{Y|X}\left[\left(f(X)-E_{Y|X}[Y|X]+E_{Y|X}[Y|X]-Y\right)^2|X\right]\right] \\
&= \begin{array}{l} E_X[\quad E_{Y|X}\left[\left(f(X)-E_{Y|X}[Y|X]\right)^2|X\right] \\ \quad +2E_{Y|X}\left[\left(f(X)-E_{Y|X}[Y|X]\right)\left(E_{Y|X}[Y|X]-Y\right)|X\right] \\ \quad +E_{Y|X}\left[\left(E_{Y|X}[Y|X]-Y\right)^2|X\right] \end{array} \qquad (3.21) \\
&= \begin{array}{l} E_X[\quad E_{Y|X}\left[\left(f(X)-E_{Y|X}[Y|X]\right)^2|X\right] \\ \quad +2\left(f(X)-E_{Y|X}[Y|X]\right)\times 0 \\ \quad +E_{Y|X}\left[\left(E_{Y|X}[Y|X]-Y\right)^2|X\right] \end{array} \\
&= E_{XY}\left[\left(f(X)-E_{Y|X}[Y|X]\right)^2\right]+R(f^*).
\end{aligned}$$

Example
Thus if $f^*(x)=E_{Y|X}[Y|X=x]$, then $R(f^*)=R^*$, as desired.

3.3 Empirical Risk Minimization

Definition 3.3: Empirical Risk
Let $\{X_i,Y_i\}_{i=1}^n \overset{iid}{\sim} P_{XY}$ be a collection of training data. Then the empirical risk is defined as

$$\hat{R}_n(f)=\frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i),Y_i). \qquad (3.22)$$

Empirical risk minimization is the process of choosing a learning rule which minimizes the empirical risk; **i.e.,**

$$\hat{f}_n = \underset{f\in\mathcal{F}}{argmin}\,\hat{R}_n(f). \qquad (3.23)$$

Example 3.1: Pattern Classification
Let the set of possible classifiers be

$$\mathcal{F}=\{x\mapsto \text{sign}\left(w^{'}x\right): w\in\mathbf{R}^d\} \qquad (3.24)$$

and let the feature space, $\mathcal{X}$, be $[0,1]^d$ or $\mathbf{R}^d$. If we use the notation $f_w(x)\equiv\text{sign}\left(w^{'}x\right)$, then the set of classifiers can be alternatively represented as

$$\mathcal{F}=\{f_w: w\in\mathbf{R}^d\}. \qquad (3.25)$$

In this case, the classifier which minimizes the empirical risk is

$$\begin{aligned} \hat{f}_n &= \underset{f \in \mathcal{F}}{argmin} \hat{R}_n(f) \\ &= \underset{w \in \mathbf{R}^d}{arg\,min} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\text{sign}(w^{'} X_i) \neq Y_i\}}. \end{aligned} \tag{3.26}$$

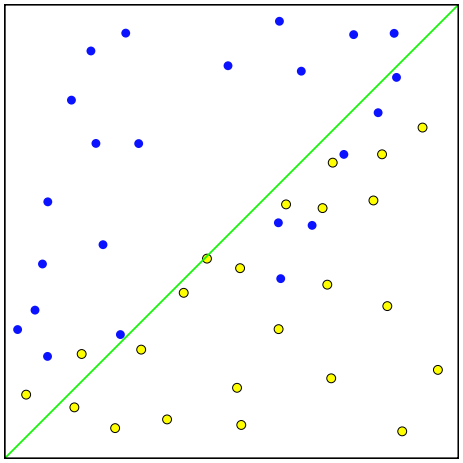


Figure 3.1: Example linear classifier for two-class problem.

Example 3.2: Regression
Let the feature space be

$$\mathcal{X} = [0, 1] \tag{3.27}$$

and let the set of possible estimators be

$$\mathcal{F} = \{\text{degree } d \text{ polynomials on } [0, 1]\}. \tag{3.28}$$

In this case, the classifier which minimizes the empirical risk is

$$\begin{aligned} \hat{f}_n &= \underset{f \in \mathcal{F}}{argmin} \hat{R}_n(f) \\ &= \underset{f \in \mathcal{F}}{argmin} \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) - Y_i\right)^2. \end{aligned} \tag{3.29}$$

Alternatively, this can be expressed as

$$\begin{aligned} \hat{w} &= \underset{w \in \mathbf{R}^{d+1}}{arg\,min} \frac{1}{n} \sum_{i=1}^{n} \left(w_0 + w_1 X_i + ... + w_d X_i^d - Y_i\right)^2 \\ &= \underset{w \in \mathbf{R}^{d+1}}{arg\,min} \parallel Vw - Y \parallel^2 \end{aligned} \tag{3.30}$$

where V is the Vandermonde matrix

$$V = \begin{bmatrix} 1 & X_1 & ... & X_1^d \\ 1 & X_2 & ... & X_2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & ... & X_n^d \end{bmatrix}. \tag{3.31}$$

The pseudoinverse can be used to solve for $\hat{w}$:

$$\hat{w} = \left(V^{'}V\right)^{-1}V^{'}Y. \tag{3.32}$$

A polynomial estimate is displayed in Figure 3.2.

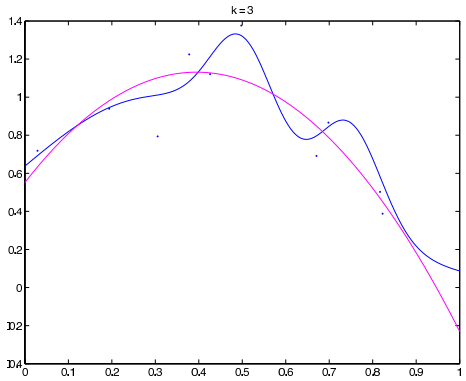


Figure 3.2: Example polynomial estimator. Blue curve denotes f^*, magenta curve is the polynomial fit to the data (denoted by dots).

3.4 Overfitting

Suppose $\mathcal{F}$, our collection of candidate functions, is very large. We can always make

$$\min_{f \in \mathcal{F}} \hat{R}_n(f) \tag{3.33}$$

smaller by increasing the cardinality of $\mathcal{F}$, thereby providing more possibilities to fit to the data.

Consider this extreme example: Let $\mathcal{F}$ be all measurable functions. Then every function f for which

$$f(x) = \begin{cases} Y_i, & x = X_i \text{ for } i = 1, ..., n \\ \text{any value}, & \text{otherwise} \end{cases} \tag{3.34}$$

has zero empirical risk ($\hat{R}_n(f) = 0$). However, clearly this could be a very poor predictor of Y for a new input X .

Example 3.3: Classification Overfitting

Consider the classifier in Figure 3.3; this demonstrates overfitting in classification. If the data were in fact generated from two Gaussian distributions centered in the upper left and lower right quadrants of the feature space domain, then the optimal estimator would be the linear estimator in Figure 3.1; the overfitting would result in a higher probability of error for predicting classes of future observations.

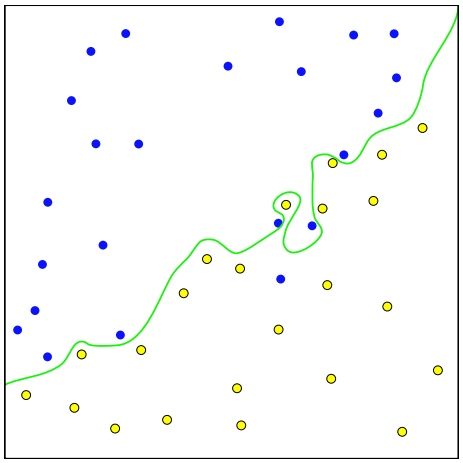


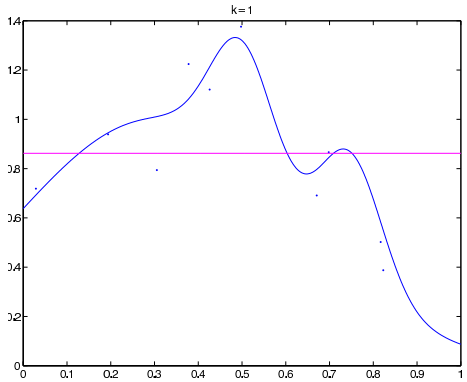
Figure 3.3: Example of overfitting classifier. The classifier's decision boundary wiggles around in order to correctly label the training data, but the optimal Bayes classifier is a straight line.

Example 3.4: Regression Overfitting

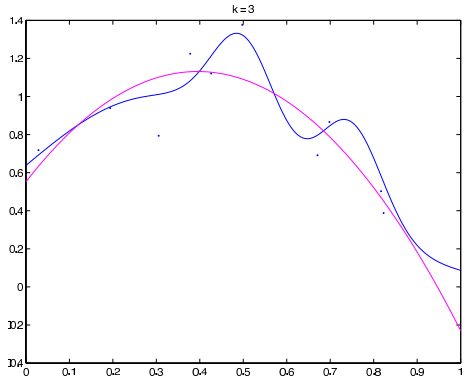
Below is an m-file that simulates the polynomial fitting. Feel free to play around with it to get an idea of the overfitting problem.

```
    %~poly~fitting
%~rob~nowak~~1/24/04
clear
close~all
~
%~generate~and~plot~"true"~function
t~=~(0:.001:1)';
f~=~exp(-5*(t-.3).^2)+.5*exp(-100*(t-.5).^2)+.5*exp(-100*(t-.75).^2);
figure(1)
plot(t,f)
~
%~generate~n~training~data~&~plot
n~=~10;
sig~=~0.1;~%~std~of~noise
x~=~.97*rand(n,1)+.01;
y~=~exp(-5*(x-.3).^2)+.5*exp(-100*(x-.5).^2)+.5*exp(-100*(x-.75).^2)+sig*randn(size(x));
figure(1)
clf
plot(t,f)
hold~on
```

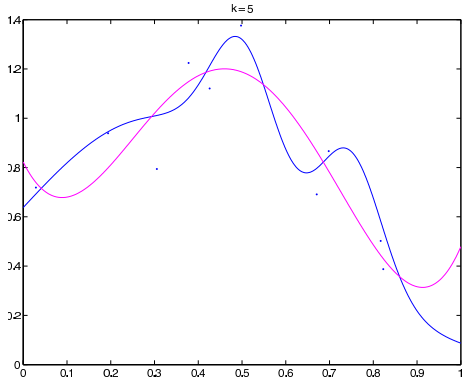
```
plot(x,y,'.')
~
%~fit~with~polynomial~of~order~k~~(poly~degree~up~to~k-1)
k=3;
for~i=1:k
~~~~V(:,i)~=~x.^(i-1);
end
p~=~inv(V'*V)*V'*y;
~
for~i=1:k
~~~~Vt(:,i)~=~t.^(i-1);
end
yh~=~Vt*p;
figure(1)
clf
plot(t,f)
hold~on
plot(x,y,'.')
plot(t,yh,'m')
~
```



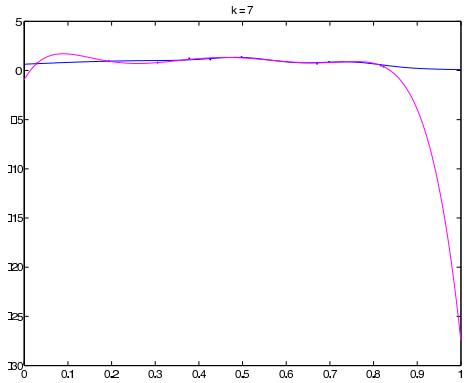
(a)



(b)



(c)



(d)

Figure 3.4: Example polynomial fitting problem. Blue curve is f^*, magenta curve is the polynomial fit to the data (dots). (a) Fitting a polynomial of degree $d = 0$: This is an example of underfitting (b)$d = 2$ (c) $d = 4$ (d) $d = 6$: This is an example of overfitting. The empirical loss is zero, but clearly the estimator

Chapter 4

Introduction to Complexity Regularization[1]

4.1 Competing Goals: The Bias-Variance Tradeoff

We ended the previous lecture (Chapter 3) with a brief discussion of overfitting. Recall that, given a set of n data points, D_n, and a space of functions (or **models**) $\mathcal{F}$, our goal in solving the learning from data problem is to choose a function $\hat{f}_n \in \mathcal{F}$ which minimizes the expected risk $E\left[R\left(\hat{f}_n\right)\right]$, where the expectation is being taken over the distribution P_{XY} on the data points D_n. One approach to avoiding overfitting is to restrict $\mathcal{F}$ to some subset of all measurable function. To gauge the performance of a given f in this case, we examine the difference between the expected risk of f and the Bayes' risk (called the **excess risk**).

$$E\left[R\left(\hat{f}_n\right)\right] - R^* = \underbrace{\left(E\left[R\left(\hat{f}_n\right)\right] - inf_{f\in\mathcal{F}} R(f)\right)}_{\text{estimation error}} + \underbrace{\left(inf_{f\in\mathcal{F}} R(f) - R^*\right)}_{\text{approximation error}} \tag{4.1}$$

The **approximation error** term quantifies the performance hit incurred by imposing restrictions on $\mathcal{F}$. The **estimation error** term is due to the randomness of the training data, and it expresses how well the chosen function $\hat{f}_n$ will perform in relation to the best possible f in the class $\mathcal{F}$. This decomposition into stochastic and approximation errors is similar to the bias-variance tradeoff which arises in classical estimation theory. The approximation error is like a bias squared term, and the estimation error is like a variance term. By allowing the space $\mathcal{F}$to be large[2] we can make the approximation error as small as we want at the cost of incurring a large estimation error. On the other hand, if $\mathcal{F}$is very small then the approximation error will be large, but the estimation error may be very small. This tradeoff is illustrated in Figure 4.1.

[1]This content is available online at <http://cnx.org/content/m16274/1.2/>.
[2]When we say $\mathcal{F}$is large, we mean that $|\mathcal{F}|$, the number of elements in $\mathcal{F}$, is large.

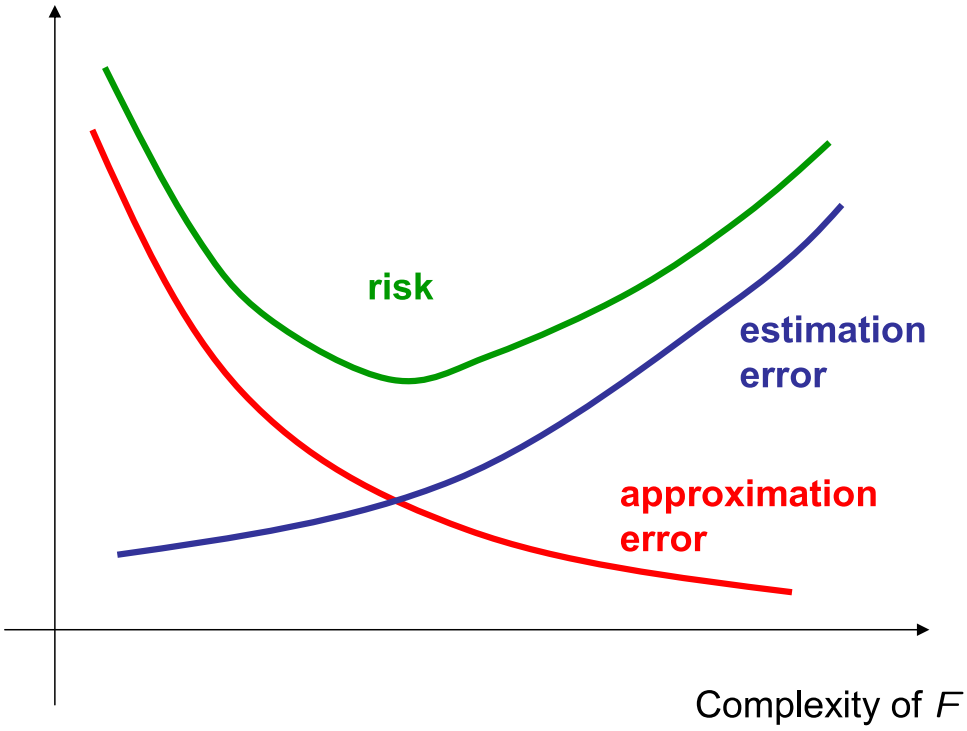


Figure 4.1: Illustration of tradeoff between estimation and approximation errors as a function of the size (complexity) of the $\mathcal{F}$.

Why is this the case? We do not know the true distribution P_{XY} on the data, so instead of minimizing the expected risk of we design a predictor by minimizing the empirical risk:

$$\begin{aligned} \hat{f}_n &= \underset{f \in \mathcal{F}}{argmin} \hat{R}_n(f), \\ \hat{R}_n(f) &= \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i). \end{aligned} \tag{4.2}$$

If $\mathcal{F}$is very large then $\hat{R}_n(f)$ can be made arbitrarily small and the resulting $\hat{f}_n$ can "overfit" to the data since $\hat{R}_n(f)$ is not a good estimator of the true risk $R\left(\hat{f}_n\right)$.

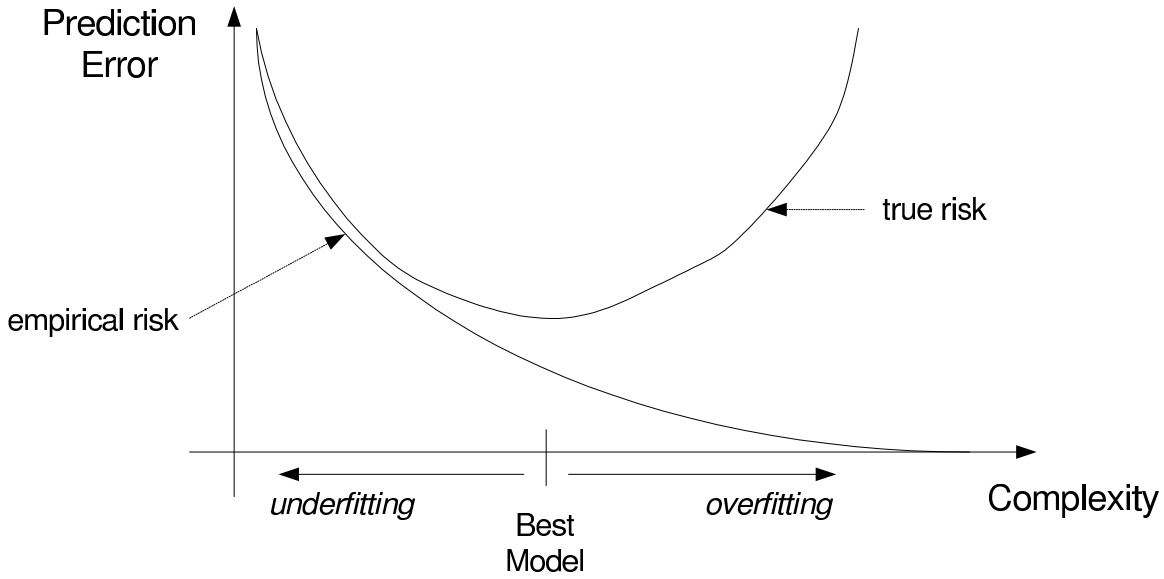


Figure 4.2: Illustration of empirical risk and the problem of overfitting to the data.

The behavior of the true and empirical risks, as a function of the size (or **complexity**) of the space $\mathcal{F}$, is illustrated in Figure 4.2. Unfortunately, we can't easily determine whether we are over or underfitting just by looking at the empirical risk.

4.2 Strategies To Avoid Overfitting

Picking

$$\hat{f}_n = \underset{f \in \mathcal{F}}{argmin} \hat{R}_n(f) \tag{4.3}$$

is problematic if $\mathcal{F}$is large. We will examine two general approaches to dealing with this problem:

1. Restrict the size or dimension of $\mathcal{F}$(e.g., restrict $\mathcal{F}$to the set of all lines, or polynomials with maximum degree d). This effectively places an upper bound on the estimation error, but in general it also places a lower bound on the approximation error.
2. Modify the empirical risk criterion to include an extra cost associated with each model (e.g., higher cost for more complex models):

$$\hat{f}_n = \underset{f \in \mathcal{F}}{argmin} \{\hat{R}_n(f) + C(f)\}. \tag{4.4}$$

The cost is designed to mimic the behavior of the estimation error so that the model selection procedure avoids models with a estimation error. Roughly this can be interpreted as trying to balance the tradeoff illustrated in Figure 4.1. Procedures of this type are often called complexity penalization methods.

Example 4.1
Revisit the polynomial regression example (Lecture 2, Ex. 4) (Example 3.4: Regression Overfitting), and incorporate a penalty term $C(f)$ which is proportional to the degree of f, or the derivative of f. In essence, this approach penalizes for functions which are too "wiggly", with the intuition being that the true function is probably smooth so a function which is very wiggly will overfit the data.

How do we decide how to restrict or penalize the empirical risk minimization process? Approaches which have appeared in the literature include the following.

4.2.1 Method of Sieves

Perhaps the simplest approach is to try to limit the size of $\mathcal{F}$in a way that depends on the number of training data n. The more data we have, the more complex the space of models we can entertain. Let the class of candidate functions grow with n. That is, take

$$F_1, \mathcal{F}_2, \cdots, \mathcal{F}_n, \cdots \tag{4.5}$$

where $|\mathcal{F}_i|$ grows as $i \to \infty$. In other words, consider a sequence of spaces with increasing complexity or degrees of freedom depending on the number of training data samples, n.

Given samples $\{X_i, Y_i\}_{i=1}^n$ i.i.d. distributed according to P_{XY}, select $f \in \mathcal{F}_n$ to minimize the empirical risk

$$\hat{f}_n = arg\min_{f \in \mathcal{F}_n} \hat{R}_n(f). \tag{4.6}$$

In the next lecture (Chapter 5) we will consider an example using the method of sieves. The basic idea is to design the sequence of model spaces in such a way that the excess risk decays to zero as $n \to \infty$. This sort of idea has been around for decades, but Grenander's method of sieves is often cited as a nice formalization of the idea: **Abstract Inference**, Wiley, New York.

4.2.2 Complexity Penalization Methods

4.2.2.1 Bayesian Methods

In certain cases, the empirical risk happens to be a (log) likelihood function, and one can then interpret the cost $C(f)$ as reflecting prior knowledge about which models are more or less likely. In this case, $e^{-C(f)}$ is like a prior probability distribution on the space $\mathcal{F}$. The cost $C(f)$ is large if f is highly improbable, and $C(f)$ is small if f is highly probable.

Alternatively, if we restrict $\mathcal{F}$to be small, and denote the space of all measurable functions as $\mathbb{F} = \mathcal{F} \cup \mathcal{F}^c$, then it is essentially as if we have placed a uniform prior over all functions in $\mathcal{F}$, and zero prior probability on the functions in $\mathcal{F}^c$.

4.2.2.2 Description Length Methods

Description length methods represent each f with a string of bits. More complicated functions require more bits to represent. Accordingly, we can then set the cost $c(f)$ proportional to the number of bits needed to describe f (the **description length**). This results in what is known as the minimum description length (MDL) approach where the minimum description length is given by

$$\min_{f \in \mathcal{F}} \{\hat{R}_n(f) + C(f)\}. \tag{4.7}$$

In the Bayesian setting, $p(f) \propto e^{-C(f)}$ can be interpreted as a prior probability density on $\mathcal{F}$, with more complex models being less probable and simpler models being more probable. In that sense, both the Bayesian and MDL approaches have a similar spirit.

4.2.2.3 Vapnik-Cervonenkis Dimension

The Vapnik-Cervonenkis (VC) dimension measures the complexity of a class $\mathcal{F}$relative to a random sample of n training data. For example, take $\mathcal{F}$to be all linear classifiers in 2-dimensional feature space. Clearly, the space of linear classifiers is infinite (there are an infinite number of lines which can be drawn in the plane). However, many of these linear classifiers would assign the same labels to the training data.

The number of unique labellings of the training data that can be achieved with linear classifiers is, in fact, finite. A line can be defined by picking **any** pair of training points, as illustrated in Figure 4.3. Two classifiers can be defined from each such line: one that outputs a label "1" for everything on or above the line, and another that outputs "0" for everything on or above. There exist $\binom{n}{2}$ such pairs of training points, and these define all possible unique labellings of the training data. Therefore, there are at most $2\binom{n}{2}$ unique linear classifiers for any random set of n 2-dimensional features (the factor of 2 is due to the fact that for each linear classifier there are 2 possible assignments of the labelling).

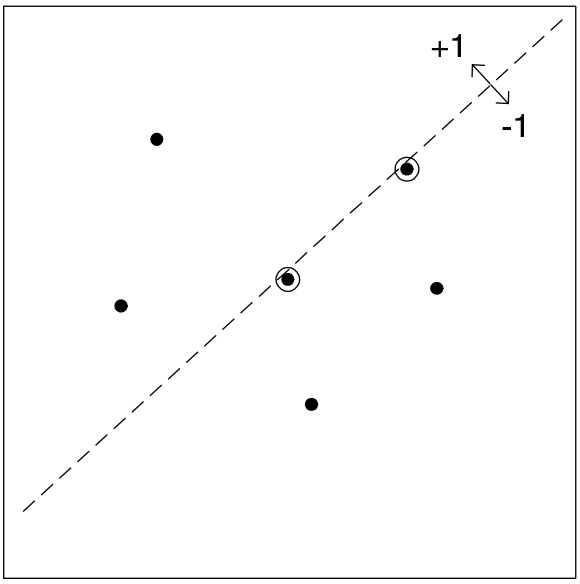


Figure 4.3: Fitting a linear classifier to 2-dimensional data. There are an infinite number of such classifiers. We can generate a linear classifier by choosing two data points, drawing a line with both points on one side, and declaring all points on or above the line to be "+1" (or "−1") and all points below the line to be "−1" (or "+1").

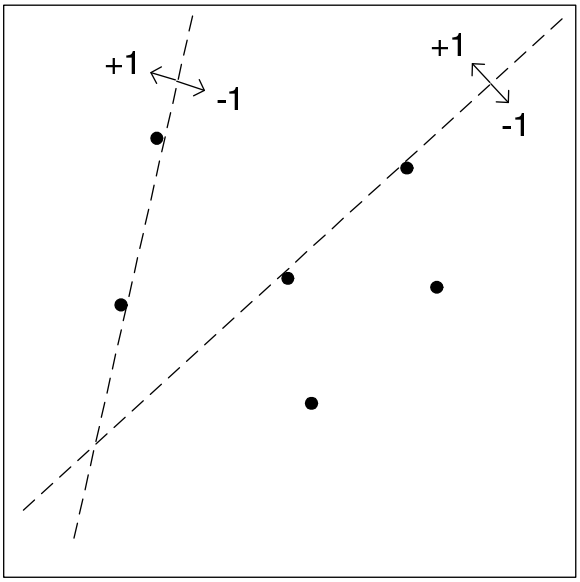


Figure 4.4: From the discussion in the previous figure, we see that the two linear classifiers depicted in this figure are equivalent for this set of data points, and hence relative to the set of n training data there are only on the order of n^2 unique linear classifiers.

Thus, instead of infinitely many linear classifiers, we realize that as far as a random sample of n training data is concerned, there are at most

$$\begin{aligned} 2\binom{n}{2} &= \tfrac{2n!}{(n-2)!2!} \\ &= n(n-1) \end{aligned} \tag{4.8}$$

unique linear classifiers. That is, using linear classification rules, there are at most $n(n-1) \approx n^2$ unique label assignments for n data points. If we like, we can encode each possibility with $log_2 n(n-1) \approx 2log_2 n$ bits. In d dimensions there are $2\binom{n}{d}$ hyperplane classification rules which can be encoded in roughly $dlog_2 n$ bits. Roughly speaking, the number of bits required for encoding each model is the VC dimension. The remarkable aspect of the VC dimension is that it is often finite even when $\mathcal{F}$ is infinite (as in this example).

If $\mathcal{X}$has d dimensions in total, we might consider linear classifiers based on $1, 2, \cdots, d$ features at a time. Lower dimensional hyperplanes are less complex than higher dimensional ones. Suppose we set

$$\begin{aligned} \mathcal{F}_1 &= \text{linear classifiers using 1 feature} \\ \mathcal{F}_2 &= \text{linear classifiers using 2 features} \\ \cdots &\qquad \text{and so on} \end{aligned} \tag{4.9}$$

These spaces have increasing VC dimensions, and we can try to balance the empirical risk and a cost function depending on the VC dimension. Such procedures are often referred to as **Structural Risk Minimization**. This gives you a glimpse of what the VC dimension is all about. In future lectures we will revisit this topic in greater detail.

4.2.3 Hold-out Methods

The basic idea of "hold-out" methods is to split the n samples $D \equiv \{X_i, Y_i\}_{i=1}^n$ into a training set, D_T, and a test set, D_V.

$$D_T = \{X_i, Y_i\}_{i=1}^m, \qquad D_V = \{X_i, Y_i\}_{i=m+1}^n \ . \tag{4.10}$$

Now, suppose we have a collection of different model spaces $\{\mathcal{F}_\lambda\}$ indexed by $\lambda \in \Lambda$ (e.g., $\mathcal{F}_\lambda$ is the set of polynomials of degree d, with $\lambda = d$), or suppose that we have a collection of complexity penalization criteria $L_\lambda(f)$ indexed by λ (**e.g.,**let $L_\lambda(f) = \hat{R}(f) + \lambda c(f)$, with $\lambda \in \mathbf{R}^+$). We can obtain candidate solutions using the training set as follows. Define

$$\hat{R}_m(f) = \sum_{i=1}^m \ell(f(X_i), Y_i) \tag{4.11}$$

and take

$$\hat{f}_\lambda = \underset{f \in \mathcal{F}_\lambda}{argmin} \hat{R}_m(f) \tag{4.12}$$

or

$$\hat{f}_\lambda = \underset{f \in \mathcal{F}}{argmin} \{\hat{R}_m(f) + \lambda c(f)\} \ . \tag{4.13}$$

This provides us with a set of candidate solutions $\{\hat{f}_\lambda\}$. Then we can define the hold-out error estimate using the test set:

$$\hat{R}_V(f) = \frac{1}{n-m+1} \sum_{i=m+1}^n \ell(f(X_i), Y_i), \tag{4.14}$$

and select the "best" model to be $\hat{f} = \hat{f}_{\hat{\lambda}}$ where

$$\hat{\lambda} = \underset{\lambda}{argmin} \hat{R}_V\left(\hat{f}_\lambda\right) \ . \tag{4.15}$$

This type of procedure has many nice theoretical guarantees, provided both the training and test set grow with n.

4.2.3.1 Leaving-one-out Cross-Validation

A very popular hold-out method is the so call "leaving-one-out cross-validation" studied in depth by Grace Wahba (UW-Madison, Statistics). For each λ we compute

$$\hat{f}_\lambda^{(k)} = \underset{f \in \mathcal{F}}{argmin} \frac{1}{n} \sum_{\substack{i=1 \\ i \neq k}}^n \ell(f(X_i), Y_i) + \lambda C(f) \tag{4.16}$$

or

$$\hat{f}_\lambda^{(k)} = \underset{f \in \mathcal{F}_\lambda}{argmin} \frac{1}{n} \sum_{\substack{i=1 \\ i \neq k}}^n \ell(f(X_i), Y_i) \, . \tag{4.17}$$

Then we have cross-validation function

$$\begin{aligned} V(\lambda) &= \frac{1}{n} \sum_{k=1}^n \ell\left(f_\lambda^{(k)}(X_k), Y_k\right) \\ \lambda^* &= \underset{\lambda}{argmin} V(\lambda) \, . \end{aligned} \tag{4.18}$$

4.3 Summary

To summarize, this lecture gave a brief and incomplete survey of different methods for dealing with the issues of overfitting and model selection. Given a set of training data, $D_n = \{X_i, Y_i\}_{i=1}^n$, our overall goal is to find

$$f^* = \underset{f \in \mathcal{F}}{argmin} R(f) \tag{4.19}$$

from some collection of functions, $\mathcal{F}$. Because we do not know the true distribution P_{XY} underlying the data points D_n, it is difficult to get an exact handle on the risk, $R(f)$. If we only focus on minimizing the empirical risk $\hat{R}(f)$ we end up overfitting to the training data. Two general approaches were presented.

1. In the first approach we consider an indexed collection of spaces $\{\mathcal{F}_\lambda\}_{\lambda \in \Lambda}$ such that the complexity of $\mathcal{F}_\lambda$ increases as λ increases, and
$$\lim_{\lambda \to \infty} \mathcal{F}_\lambda = \mathcal{F}. \tag{4.20}$$
A solution is given by
$$\hat{f}_{\lambda^*} = arg \min_{f \in \mathcal{F}_{\lambda^*}} \hat{R}_n(f) \tag{4.21}$$
where either λ^* is a function which increases with n,
$$\lambda^* = \lambda(n), \tag{4.22}$$
or λ^* is chosen by hold-out validation.
2. The alternative approach is to incorporate a penalty term into the risk minimization problem formulation. Here we consider an indexed collection of penalties $\{C_\lambda\}_{\lambda \in \Lambda}$ satisfying the following properties:
   a. $C_\lambda : \mathcal{F} \to \mathbf{R}^+$;
   b. For each $f \in \mathcal{F}$ and $\lambda_1 < \lambda_2$ we have $C_{\lambda_1}(f) \leq C_{\lambda_2}(f)$;
   c. There exists $\lambda_0 \in \Lambda$ such that $C_{\lambda_0}(f) = 0$ for all $f \in \mathcal{F}$.

   In this formulation we find a solution
$$\hat{f}_{\lambda^*} = \underset{f \in \mathcal{F}}{argmin} \hat{R}_n(f) + C_{\lambda^*}(f), \tag{4.23}$$
where either $\lambda^* = \lambda(n)$, a function growing the number of data samples n, or λ^* is selected by hold-out validation.

4.4 Consistency

If an estimator or classifier $\hat{f}_{\lambda^*}$ satisfies

$$E\left[R\left(\hat{f}_{\lambda^*}\right)\right] \to \inf_{f \in \mathcal{F}} R(f) \qquad \text{as } n \to \infty, \tag{4.24}$$

then we say that $\hat{f}_{\lambda^*}$ is $\mathcal{F}$-consistent with respect to the risk R. When the context is clear, we will simply say that $\hat{f}$ is consistent.

Chapter 5

An Example of the Use of Sieves for Complexity Regularization in Denoising[1]

Consider the following setting. Let

$$Y = f^*(X) + W, \tag{5.1}$$

where X is a random variable (r.v.) on $\mathcal{X} = [0,1]$, W is a r.v. on $\mathcal{Y} = \mathbf{R}$, independent of X and satisfying

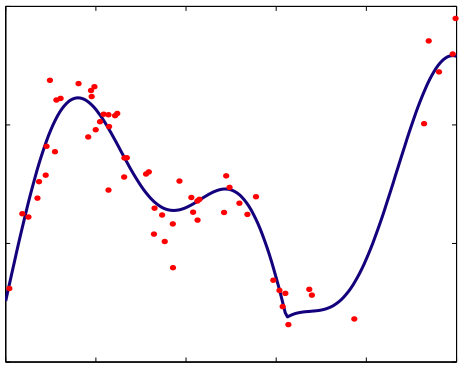
$$E[W] = 0 \quad \text{and} \quad E\left[W^2\right] = \sigma^2 < \infty. \tag{5.2}$$

Finally let $f^* : [0,1] \to \mathbf{R}$ be a function satisfying

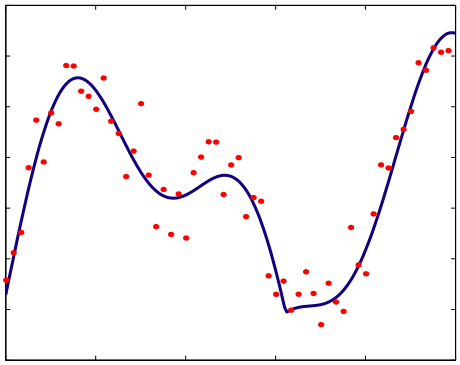
$$|f^*(t) - f^*(s)| \leq L|t-s|, \ \forall t, s \in [0,1], \tag{5.3}$$

where $L > 0$ is a constant. A function satisfying condition (5.3) is said to be Lipschitz on $[0,1]$. Notice that such a function must be continuous, but it is not necessarily differentiable. An example of such a function is depicted in Figure 5.1(a).

[1] This content is available online at <http://cnx.org/content/m16261/1.3/>.



(a)



(b)

Figure 5.1: Example of a Lipschitz function, and our observations setting. (a) random sampling of f^*, the points correspond to (X_i, Y_i), $i = 1, ..., n$; (b) deterministic sampling of f^*, the points correspond to $(i/n, Y_i)$, $i = 1, ..., n$.

Note that

$$\begin{aligned} E[Y|X=x] &= E[f^*(X) + W|X=x] \\ &= E[f^*(x) + W|X=x] \\ &= f^*(x) + E[W] = f^*(x). \end{aligned} \tag{5.4}$$

Consider our usual setup: Estimate f^* using n training examples

$$\begin{aligned} \{X_i, Y_i\}_{i=1}^n &\overset{i.i.d.}{\sim} P_{XY}, \\ Y_i = f^*(X_i) + W_i,\ i &= \{1, ..., n\}, \end{aligned} \tag{5.5}$$

where $\overset{i.i.d.}{\sim}$ means **independently and identically distributed**. Figure 5.1(a) illustrates this setup.

In many applications we can sample $\mathcal{X} = [0, 1]$ as we like, and not necessarily at random. For example

we can take n samples uniformly on [0,1]

$$\begin{aligned} x_i &= \tfrac{i}{n},\ i = 1, ..., n\ , \\ Y_i &= f(x_i) + W_i \\ &= f\left(\tfrac{i}{n}\right) + W_i. \end{aligned} \tag{5.6}$$

We will proceed with this setup (as in Figure 5.1(b)) in the rest of the lecture.

Our goal is to find $\hat{f}_n$ such that $E\left[\| f^* - \hat{f}_n\|^2\right] \to 0$, as $n \to 0$ (here $\| \cdot \|$ is the usual L_2-norm; i.e., $\| f^* - \hat{f}_n\|^2 = \int_0^1 |f^*(t) - \hat{f}_n(t)|^2\, dt$).

Let

$$\mathcal{F} = \{f :\ f \text{ is Lipschitz with constant } L\}. \tag{5.7}$$

The Risk is defined as

$$R(f) = \| f^* - f \|^2 = \int_0^1 |f^*(t) - f(t)|^2 dt. \tag{5.8}$$

The Expected Risk (recall that our estimator $\hat{f}_n$ is based on $\{x_i, Y_i\}$ and hence is a r.v.) is defined as

$$E\left[R\left(\hat{f}_n\right)\right] = E\left[\| f^* - \hat{f}_n\|^2\right]. \tag{5.9}$$

Finally the Empirical Risk is defined as

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\left(f\left(\frac{i}{n}\right) - Y_i\right)^2. \tag{5.10}$$

Let $0 < m_1 \le m_2 \le m_3 \le \cdots$ be a sequence of integers satisfying $m_n \to \infty$ as $n \to \infty$, and $k_n m_n = n$ for some integer $k_n > 0$. That is, for each value of n there is an associated integer value m_n. Define the Sieve $\mathcal{F}_1,\ \mathcal{F}_2,\ \mathcal{F}_3,\ ...,$

$$\mathcal{F}_n = \{f : f(t) = \sum_{j=1}^{m_n} c_j \mathbf{1}_{\{\frac{j-1}{m_n} \le t < \frac{j}{m_n}\}},\ c_j \in \mathbf{R}\}. \tag{5.11}$$

$\mathcal{F}_n$ is the space of functions that are constant on intervals

$$\mathrm{I}_{j,m_n} \equiv \left[\frac{j-1}{m_n}, \frac{j}{m_n}\right),\ j = 1, ..., m_n. \tag{5.12}$$

From here on we will use m and k instead of m_n and k_n (dropping the subscript n) for notational ease. Define

$$f_n(t) = \sum_{j=1}^{m} c_j^* \mathbf{1}_{\{t \in I_{j,m}\}}, \quad \text{where} \quad c_j^* = \frac{1}{k}\sum_{i:\, \frac{i}{n} \in I_{j,m}} f^*\left(\frac{i}{n}\right). \tag{5.13}$$

Note that $f_n \in \mathcal{F}_n$.

Example 5.1: Exercise 1
Upper bound $\| f^* - f_n \|^2$.

$$\begin{aligned}
\| f^* - f \|^2 &= \int_0^1 |f^*(t) - f_n(t)|^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} |f^*(t) - f_n(t)|^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} |f^*(t) - c_j^*|^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} \left| f^*(t) - \tfrac{1}{k} \sum_{i:\frac{i}{n} \in I_{j,m}} f^*\left(\tfrac{i}{n}\right) \right|^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} \left(\tfrac{1}{k} \left| \sum_{i:\frac{i}{n} \in I_{j,m}} \left(f^*(t) - f^*\left(\tfrac{i}{n}\right) \right) \right| \right)^2 dt \qquad (5.14) \\
&\leq \sum_{j=1}^m \int_{I_{j,m}} \left(\tfrac{1}{k} \sum_{i:\frac{i}{n} \in I_{j,m}} \left| f^*(t) - f^*\left(\tfrac{i}{n}\right) \right| \right)^2 dt \\
&\leq \sum_{j=1}^m \int_{I_{j,m}} \left(\tfrac{1}{k} \sum_{i:\frac{i}{n} \in I_{j,m}} \tfrac{L}{m} \right)^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} \left(\tfrac{L}{m}\right)^2 dt \\
&= \sum_{j=1}^m \tfrac{1}{m} \left(\tfrac{L}{m}\right)^2 = \left(\tfrac{L}{m}\right)^2.
\end{aligned}$$

The above implies that $\| f^* - f_n \|^2 \to 0$ as $n \to \infty$, since $m = m_n \to \infty$ as $n \to \infty$. In words, with n sufficiently large we can approximate f^* to arbitrary accuracy using models in $\mathcal{F}_n$ (even if the functions we are using to approximate f^* are not Lipschitz!).

For any $f \in \mathcal{F}_n$,$f = \sum_{j=1}^m c_j \mathbf{1}_{\{t \in I_{j,m}\}}$, we have

$$\hat{R}_n(f) = \frac{1}{n} \sum_{j=1}^m \left(\sum_{i:\frac{i}{n} \in I_{j,m}} (c_j - Y_i)^2 \right). \qquad (5.15)$$

Let $\hat{f}_n = argmin_{f \in \mathcal{F}_n} \hat{R}_n(f)$. Then

$$\hat{f}_n(t) = \sum_{j=1}^m \hat{c}_j \mathbf{1}_{\{t \in I_{j,m}\}}, \quad \text{where} \quad \hat{c}_j = \frac{1}{k} \sum_{i:\frac{i}{n} \in I_{j,m}} Y_i \qquad (5.16)$$

Example 5.2: Exercise 2
Show (5.16).

Note that $E\left[\hat{c}_j\right] = c_j^*$ and therefore $E\left[\hat{f}_n(t)\right] = f_n(t)$. Lets analyze now the expected risk of $\hat{f}_n$:

$$\begin{aligned}
E\left[\| f^* - \hat{f}_n \|^2\right] &= E\left[\| f^* - f_n + f_n - \hat{f}_n \|^2\right] \\
&= \| f^* - f_n \|^2 + E\left[\| f_n - \hat{f}_n \|^2\right] + 2E\left[< f^* - f_n, f_n - \hat{f}_n > \right] \\
&= \| f^* - f_n \|^2 + E\left[\| f_n - \hat{f}_n \|^2\right] + 2 < f^* - f_n, E\left[f_n - \hat{f}_n\right] > \qquad (5.17) \\
&= \| f^* - f_n \|^2 + E\left[\| f_n - \hat{f}_n \|^2\right],
\end{aligned}$$

where the final step follows from the fact that $E\left[\hat{f}_n(t)\right] = f_n(t)$. A couple of important remarks pertaining the right-hand-side of equation (5.17): The first term, $\| f^* - f_n\|^2$, corresponds to the approximation error, and indicates how well can we approximate the function f^* with a function from $\mathcal{F}_n$. Clearly, the larger the class $\mathcal{F}_n$ is, the smallest we can make this term. This term is precisely the squared bias of the estimator $\hat{f}_n$. The second term, $E\left[\| f_n - \hat{f}_n\|^2\right]$, is the estimation error, the variance of our estimator. We will see that the estimation error is small if the class of possible estimators $\mathcal{F}_n$ is also small.

The behavior of the first term in (5.17) was already studied. Consider the other term:

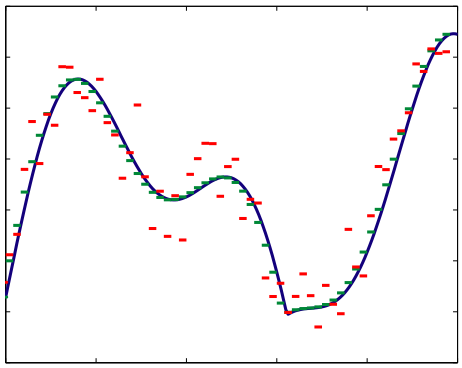
$$\begin{aligned} E\left[\| f_n - \hat{f}_n\|^2\right] &= E\left[\int_0^1 |f_n(t) - \hat{f}_n(t)|^2\, dt\right] \\ &= E\left[\sum_{j=1}^m \int_{I_{j,m}} |c_j^* - \hat{c}_j|^2\, dt\right] \\ &= \sum_{j=1}^m \int_{I_{j,m}} E\left[|c_j^* - \hat{c}_j|^2\right] dt \quad . \\ &= \sum_{j=1}^m \int_{I_{j,m}} \frac{E[W^2]}{k}\, dt \\ &\leq \sum_{j=1}^m \int_{I_{j,m}} \frac{\sigma^2}{k}\, dt \\ &= \sum_{j=1}^m \frac{1}{m}\frac{\sigma^2}{k} = \frac{\sigma^2}{k} = \frac{m}{n}\sigma^2 \end{aligned} \tag{5.18}$$

Combining all the facts derived we have

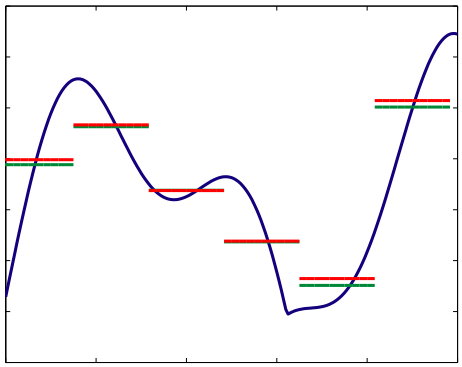
$$E\left[\| f^* - \hat{f}_n\|^2\right] \leq \frac{L^2}{m^2} + \frac{m}{n}\sigma^2 = O\left(max\{\frac{1}{m^2}, \frac{m}{n}\}\right). \tag{5.19}$$

This equation used Big-O notation.

What is the best choice of m? If m is small then the approximation error (**i.e.,** $O\left(1/m^2\right)$) is going to be large, but the estimation error (**i.e.,** $O(m/n)$) is going to be small, and vice-versa. This two conflicting goals provide a tradeoff that directs our choice of m (as a function of n). In Figure 5.2 we depict this tradeoff. In Figure 5.2(a) we considered a large m_n value, and we see that the approximation of f^* by a function in the class $\mathcal{F}_n$ can be very accurate (that is, our estimate will have a small bias), but when we use the measured data our estimate looks very bad (high variance). On the other hand, as illustrated in Figure 5.2(b), using a very small m_n allows our estimator to get very close to the best approximating function in the class $\mathcal{F}_n$, so we have a low variance estimator, but the bias of our estimator (**i.e.,** the difference between f_n and f^*) is quite considerable.



(a)



(b)

Figure 5.2: Approximation and estimation of f^* (in blue) for $n = 60$. The function f_n is depicted in green and the function $\hat{f}_n$ is depicted in red. In (a)we have $m = 60$ and in (b) we have $m = 6$.

We need to balance the two terms in the right-hand-side of (5.19) in order to maximize the rate of decay (with n) of the expected risk. This implies that $\frac{1}{m^2} = \frac{m}{n}$ therefore $m_n = n^{1/3}$ and the Mean Squared Error (MSE) is

$$E\left[\| f_n - \hat{f}_n\|^2\right] = O\left(n^{-2/3}\right). \tag{5.20}$$

So the sieve $\mathcal{F}_1,\ \mathcal{F}_2,\ \cdots$ with

$$\mathcal{F}_n = \{f : f(t) = \sum_{j=1}^{m_n} c_j \mathbf{1}_{\{\frac{j-1}{m_n} \le t < \frac{j}{m_n}\}},\ c_j \in \mathbf{R}\}, \tag{5.21}$$

produces a $\mathcal{F}$-consistent estimator for $f^* = E[Y|X + x] \in \mathcal{F}$.

It is interesting to note that the rate of decay of the MSE we obtain with this strategy cannot be further improved by using more sophisticated estimation techniques (that is, $n^{-2/3}$ is the **minimax** MSE rate for this problem). Also, rather surprisingly, we are considering classes of models $\mathcal{F}_n$ that are actually not Lipschitz, therefore our estimator of f^* is not a Lipschitz function, unlike f^* itself.

Chapter 6

Plug-In Classifier and Histogram Classifier[1]

We return to the topic of classification, and we assume an input (feature) space $\mathcal{X}$ and a binary output (label) space $\mathcal{Y} = \{0, 1\}$. Recall that the Bayes classifier (which minimizes the probability of misclassification) is defined by

$$f^{*}(x) = \begin{cases} 1, & P(Y = 1|X = x) \geq 1/2 \\ 0, & otherwise \end{cases} . \tag{6.1}$$

Throughout this section, we will denote the conditional probability function by

$$\eta(x) \equiv P(Y = 1|X = x) . \tag{6.2}$$

6.1 Plug-in Classifiers

One way to construct a classifier using the training data $\{X_i, Y_i\}^{n}_{i=1}$ is to estimate $\eta(x)$ and then plug-it into the form of the Bayes classifier. That is obtain an estimate,

$$\hat{\eta}_n(x) = \eta(x; \{X_i, Y_i\}^{n}_{i=1}) \tag{6.3}$$

and then form the "plug-in" classification rule

$$\hat{f}(x) = \begin{cases} 1, & \hat{\eta}(x) \geq 1/2 \\ 0, & otherwise \end{cases} . \tag{6.4}$$

REMARK: The function $\eta(x)$ is generally more complicated than the ultimate classification rule (binary-valued), as we can see

$$\begin{array}{rcl} \eta & : & \mathcal{X} \to [0, 1] \\ f & : & \mathcal{X} \to \{0, 1\} \end{array} . \tag{6.5}$$

[1] This content is available online at <http://cnx.org/content/m16280/1.2/>.

Therefore, in this sense plug-in methods are solving a more complicated problem than necessary. However, plug-in methods can perform well, as demonstrated by the next result.

Theorem 6.1: Plug-in Classifier
Let $\tilde{\eta}$ be an approximation to η, and consider the plug-in rule

$$f(x) = \begin{cases} 1, & \tilde{\eta}(x) \geq 1/2 \\ 0, & otherwise \end{cases}. \tag{6.6}$$

Then,

$$R(f) - R^* \leq 2E\left[|\eta(x) - \tilde{\eta}(x)|\right] \tag{6.7}$$

where

$$\begin{aligned} R(f) &= P(f(X) \neq Y) \\ R^* &= R(f^*) = \inf_f R(f) \end{aligned}. \tag{6.8}$$

Proof:
Consider any $x \in \mathbf{R}^d$. In proving the optimality of the Bayes classifier f^* in Lecture 2 (Chapter 3), we showed that

$$P(f(x) \neq Y|X = x) - P(f^*(x) \neq Y|X = x) = (2\eta(x) - 1)\left[\mathbf{1}_{\{f^*(x)=1\}} - \mathbf{1}_{\{f(x)=1\}}\right], \tag{6.9}$$

which is equivalent to

$$P(f(x) \neq Y|X = x) - P(f^*(x) \neq Y|X = x) = |2\eta(x) - 1|\ \mathbf{1}_{\{f^*(x)\neq f(x)\}}, \tag{6.10}$$

since $f^*(x) = 1$ whenever $2\eta(x) - 1 > 0$. Thus,

$$\begin{aligned} P(f(X) \neq Y) - R^* &= \int_{\mathbf{R}^d} 2|\eta(x) - 1/2|\,\mathbf{1}_{\{f^*(x)\neq f(x)\}} p_X(x)\,dx \\ &\quad \text{where } p_X(x) \text{ is the marginal density of } X \\ &\leq \int_{\mathbf{R}^d} 2|\eta(x) - \tilde{\eta}(x)|\,\mathbf{1}_{\{f^*(x)\neq f(x)\}} p_X(x)\,dx \\ &\leq \int_{\mathbf{R}^d} 2|\eta(x) - \tilde{\eta}(x)|\,p_X(x)\,dx \\ &= 2E\left[|\eta(X) - \tilde{\eta}(X)|\right] \end{aligned} \tag{6.11}$$

where the first inequality follows from the fact

$$f(x) \neq f^*(x) \Rightarrow |\eta(x) - \tilde{\eta}(x)| \geq |\eta(x) - 1/2| \tag{6.12}$$

and the second inequality is simply a result of the fact that $\mathbf{1}_{\{f^*(x)\neq f(x)\}}$ is either 0 or 1.

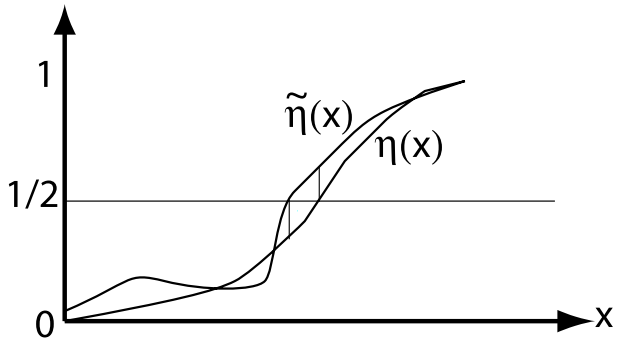


Figure 6.1: Pictorial illustration of $|\eta(x) - \tilde{\eta}(x)| \geq |\eta(x) - 1/2|$ when $f(x) \neq f^*(x)$. Note that the inequality $P(f(X) \neq Y) - R^* \leq \int_{\mathbf{R}^d} 2|\eta(x) - \tilde{\eta}(x)| \mathbf{1}_{\{f^*(x) \neq f(x)\}} p_X(x)\, dx$ shows that the excess risk is at most twice the integral over the set where $f^*(x) \neq f(x)$. The difference $|\eta(x) - \tilde{\eta}(x)|$ may be arbitrarily large away from this set without effecting the error rate of the classifier. This illustrates the fact that estimating η well everywhere (i.e., regression) is unnecessary for the design of a good classifier (we only need to determine where η crosses the 1/2-level). In other words, "classification is easier than regression."

The theorem shows us that a good estimate of η can produce a good plug-in classification rule. By "good" estimate, we mean an estimator $\tilde{\eta}$ that is close to η in expected L_1-norm.

6.2 The Histogram Classifier

Let's assume that the (input) features are randomly distributed over the unit hypercube $\mathcal{X} = [0,1]^d$ (note that by scaling and shifting any set of bounded features we can satisfy this assumption), and assume that the (output) labels are binary, i.e., $\mathcal{Y} = \{0,1\}$. A histogram classifier is based on a partition the hypercube $[0,1]^d$ into M smaller cubes of equal size.

Example 6.1: Partition of hypercube in 2 dimensions
Consider the unit square $[0,1]^2$ and partition it into M subsquares of equal area (assuming M is a squared integer). Let the subsquares be denoted by $\{Q_i\},\ i = 1, ..., M$.

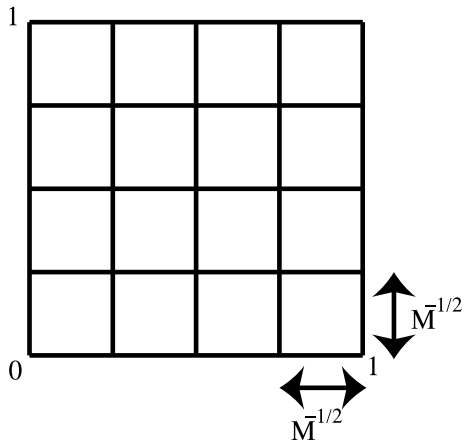


Figure 6.2: Example of hypercube $[0,1]^2$ in M equally sized partition

Define the following piecewise-constant estimator of $\eta(x)$:

$$\hat{\eta}_n(x) = \sum_{j=1}^{M} \hat{P}_j \mathbf{1}_{\{x \in Q_j\}} \tag{6.13}$$

where

$$\hat{P}_j = \frac{\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in Q_j, Y_i = 1\}}}{\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in Q_j\}}}. \tag{6.14}$$

Like our previous denoising examples, we expect that the bias of $\hat{\eta}_n$ will decrease as M increases, but the variance will increase as M increases.

Theorem 6.2: Consistency of Histogram Classifiers
If $M \to \infty$ and $\frac{n}{M} \to \infty$ as $n \to \infty$, then the histogram classifier risk converges to the Bayes risk for every distribution P_{XY} with marginal density $p_X(x) \geq c$, for some constant $c > 0$.[2].

What the theorem tells us is that we need the number of partition cells to tend to infinity (to insure that the bias tends to zero), but they can't grow faster than the number of samples (**i.e.,** we want the number of samples per box tending to infinity to drive the variance to zero).

Proof:

Let $P_j \equiv \frac{\int_{Q_j} \eta(x) p_X(x) dx}{\int_{Q_j} p_X(x) dx}$ (the theoretical analog of $\hat{P}_j$) and define

$$\overline{\eta}(x) = \sum_{j=1}^{M} P_j \mathbf{1}_{\{x \in Q_j\}} \tag{6.15}$$

The function $\overline{\eta}$ is the theoretical analog of $\hat{\eta}$ (i.e., the function obtained by averaging η over the

[2] Actually, the result holds for every distribution P_{XY}. For the more general theorem, refer to Theorem 6.1 in **A probabilistic Theory of Pattern Recognition** by Luc Devroye, László Györfi and Gábor Lugosi.

partition cells). By the triangle inequality,

$$E\left[\left|\hat{\eta}_n(X)-\eta(X)\right|\right] \leq \underbrace{E\left[\left|\hat{\eta}_n(X)-\overline{\eta}(X)\right|\right]}_{EstimationError} + \underbrace{E\left[\left|\overline{\eta}_n(X)-\eta(X)\right|\right]}_{ApproximationError} \tag{6.16}$$

Let's first bound the estimation error. For any $x \in [0,1]^d$, let $Q(x)$ denote the histogram bin in which x falls in. Define the random variable

$$N(x) = \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in Q(x)\}} \tag{6.17}$$

If $Q(x) = Q_j$, then this random variable is simply $n\hat{P}_j$. Note that

$$\hat{\eta}_n(x) = \frac{1}{N(x)} B(x) \tag{6.18}$$

where $B(x) == \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in Q(x),\, Y_i=1\}} = \sum_{i:X_i \in Q(x)} Y_i$. $B(x)$ is simply the number of samples in cell $Q(x)$ labelled 1. Now $\hat{\eta}_n(x)$ is a fairly complicated random variable, but the conditional distribution of $B(x)$ given $N(x)$ is relatively simple. Note that

$$B(x) \mid N(x) = k \;\sim\; \text{Binomial}(k, \overline{\eta}(x)) \tag{6.19}$$

since $\overline{\eta}(x)$ is the probability of a sample in $Q(x)$ having the label 1 and we are conditioning on the event of observing k samples in $Q(x)$.

Now consider the conditional expectation

$$E\left[\left|\hat{\eta}_n(x)-\overline{\eta}(x)\right| \mid N(x)=k\right] \leq \begin{cases} E\left[\left|\frac{B(x)}{N(x)}-\overline{\eta}(x)\right| \mid N(x)=k\right], & k>0 \\ 1, & k=0 \quad (\text{since } 0 \leq \overline{\eta}(x) \leq 1) \end{cases} \tag{6.20}$$

Next note that

$$\begin{aligned} E\left[\left|\frac{B(x)}{N(x)}-\overline{\eta}(x)\right| \mid N(x)=k\right] &= E\left[\left|\frac{B(x)}{k}-\overline{\eta}(x)\right| \mid N(x)=k\right] \\ &= E\left[\frac{1}{k}|B(x) - \underbrace{k\overline{\eta}(x)}_{E[B(x)]}| \mid N(x)=k\right] \\ &\leq \frac{1}{k}\left(\underbrace{E\left[|B(x)-k\overline{\eta}(x)|^2 \mid N(x)=k\right]}_{\text{conditional variance of } B(x)}\right)^{\frac{1}{2}} \end{aligned} \tag{6.21}$$

by the Jensen's inequality, $E[|Z|] \leq \left(E\left[|Z|^2\right]\right)^{\frac{1}{2}}$.

Therefore,

$$\begin{aligned} E\left[\left|\frac{B(x)}{N(x)}-\overline{\eta}(x)\right| \mid N(x)=k\right] &\leq \frac{1}{k}\left(k\overline{\eta}(x)(1-\overline{\eta}(x))\right)^{\frac{1}{2}} \\ &= \sqrt{\frac{\overline{\eta}(x)(1-\overline{\eta}(x))}{k}} \end{aligned} \tag{6.22}$$

and

$$E\left[\left|\hat{\eta}_n(x)-\overline{\eta}(x)\right|\,|\,N(x)=k\right] \;\leq\; \begin{cases} \sqrt{\frac{\overline{\eta}(x)(1-\overline{\eta}(x))}{k}}, & k>0 \\ 1, & k=0 \end{cases} \tag{6.23}$$

or in other words,

$$E\left[\left|\hat{\eta}_n(x)-\overline{\eta}(x)\right|\,|\,N(x)=k\right] \;\leq\; \sqrt{\frac{\overline{\eta}(x)(1-\overline{\eta}(x))}{N(x)}}\mathbf{1}_{\{N(x)>0\}} \;+\; \mathbf{1}_{\{N(x)=0\}} \tag{6.24}$$

Now taking expectation with respect to $N(x)$

$$\begin{array}{lll} E_N\left[E\left[\left|\hat{\eta}_n(x)-\overline{\eta}(x)\right|N(x)=k\right]\right] & \leq & E_N\left[\sqrt{\frac{\overline{\eta}(x)(1-\overline{\eta}(x))}{N(x)}}\mathbf{1}_{\{N(x)>0\}}\right] \;+ \\ P(N(x)=0) \;\leq\; E\left[\frac{1}{2\sqrt{N(x)}}\mathbf{1}_{\{N(x)>0\}}\right] & + & P(N(x)=0) \;\leq\; \frac{1}{2}P(N(x)\leq k) \;+ \\ \frac{1}{2\sqrt{k}}\underbrace{P(N(x)>k)}_{\leq 1} \;+\; P(N(x)=0) & & \end{array} \tag{6.25}$$

Now a key fact is that for any $k>0$, $P(N\leq k)\to 0$ as $n\to\infty$. This follows from the assumption that the marginal density $p_X(x)\geq c$, for some constant $c>0$, and $\frac{n}{M}\to\infty$ as $n\to\infty$. This result is easily verified by contradiction. If $P(N\leq k)\to q>0$ as $n\to\infty$, then $P_X(x)>0$ is contradicted. Thus, for any $\varepsilon>0$ there exists a $k>0$ such that $\frac{1}{2\sqrt{k}}<\varepsilon$ and $P(N\leq k)<\varepsilon$ for n sufficiently large. Therefore, for n sufficiently large and every $x\in[0,1]^d$,

$$E\left[\left|\hat{\eta}_n(x)-\overline{\eta}(x)\right|\right]<3\varepsilon \tag{6.26}$$

where the expectation is with respect to the distribution of the sample $\{X_i,Y_i\}_{i=1}^n$. Thus,

$$E\left[\left|\hat{\eta}_n(X)-\overline{\eta}(X)\right|\right] \;<\; 3\varepsilon \tag{6.27}$$

where the expectation is now with respect to the distribution of the sample and the marginal distribution of X.

Next consider the approximation error $E\left[\left|\overline{\eta}_n(X)-\eta(X)\right|\right]$, where the expectation is over X alone. The function η may not itself be continuous, but there is another function η_ε that is uniformly continuous and such that $E\left[\left|\eta_\varepsilon(X)-\eta(X)\right|\right]<\varepsilon$. Recall that uniformly continuous functions can be well approximated by piecewise constant functions.

By the triangle inequality,

$$E\left[\left|\overline{\eta}-\eta\right|\right]\leq \underbrace{E\left[\left|\overline{\eta}-\overline{\eta}_\varepsilon\right|\right]}_{\leq\varepsilon}+E\left[\left|\overline{\eta}_\varepsilon-\eta_\varepsilon\right|\right]+\underbrace{E\left[\left|\eta_\varepsilon-\eta\right|\right]}_{\leq\varepsilon \text{ by design}} \tag{6.28}$$

where $\overline{\eta}_\varepsilon(x)=\sum_{j=1}^m\left[\int_{Q_j}\eta_\varepsilon(x')\,p_X(x')\,dx'\right]\mathbf{1}_{\{x\in Q_j\}}$.

$$\begin{array}{rcl} E\left[\left|\overline{\eta}(X)-\overline{\eta}_\varepsilon(X)\right|\right] & = & \sum_{j=1}^m\left[\int_{Q_j}\left|\eta(x)-\eta_\varepsilon(x)\right|p_X(x)\,dx\right]\mathbf{1}_{\{x\in Q_j\}} \\ & \leq & \varepsilon \end{array} \tag{6.29}$$

and since η_ε is uniformly continuous,

$$\begin{aligned} E\left[|\overline{\eta}_\varepsilon(X)-\eta_\varepsilon(X)|\right] &= \sum_{j=1}^{M}\int_{Q_j}|\overline{\eta}_\varepsilon(x)-\eta_\varepsilon(x)|\mathbf{1}_{\{x\in Q_j\}}p_X(x)\,dx \\ &\leq \sum_{j=1}^{M}\delta\,P(x\in Q_j), \quad \text{where } \delta \text{ depends on } M \\ &= \delta, \quad \text{since } \sum_{j=1}^{M}P(X\in Q_j)=1 \end{aligned} \tag{6.30}$$

By taking M sufficiently large, δ can be made arbitrarily small. So for large M, $\delta \leq \varepsilon$.

Thus, we have shown

$$E\left[|\overline{\eta}(X)-\eta(X)|\right] < 3\varepsilon \tag{6.31}$$

for sufficiently large M. Since $\varepsilon > 0$ was arbitrary, we have shown that taking

$$\hat{f}_n(x) = \begin{cases} 1, & \hat{\eta}_n(x) \geq 1/2 \\ 0, & otherwise \end{cases} \tag{6.32}$$

satisfies

$$P\left(\hat{f}_n(X)\neq Y\right) - P\left(f^*(X)\neq Y\right) \leq 2E\left[|\hat{\eta}_n(X)-\eta(X)|\right] \to 0 \tag{6.33}$$

if

$$\begin{aligned} M &\to \infty \\ \tfrac{n}{M} &\to \infty \text{ as } n\to\infty \end{aligned} \tag{6.34}$$

NOTE: $P\left(\hat{f}_n(X)\neq Y\right) = E\left[\mathbf{1}_{\{\hat{f}(X)\neq Y\}}\right]$ is the expected risk of $\hat{f}$, with expectation over the distributions of (X,Y) and $\{X_i,Y_i\}_{i=1}^n$.

Chapter 7

Probably Approximately Correct (PAC) Learning[1]

7.1 Introduction

7.1.1 Overview of the Learning Problem

The fundamental problem in learning from data is proper Model Selection. As we have seen in the previous lectures, a model that is too complex could overfit the training data (causing an estimation error) and a model that is too simple could be a bad approximation of the function that we are trying to estimate (causing an approximation error). The estimation error arises because of the fact that we do not know the true joint distribution of data in the input and output space, and therefore we minimize the empirical risk (which, for each candidate model, is a random number depending on the data) and estimate the average risk again from the limited number of training samples we have. The approximation error measures how well the functions in the chosen model space can approximate the underlying relationship between the output space on the input space, and in general improves as the "size" of our model space increases.

7.1.2 Lecture Outline

In the preceding lectures, we looked at some solutions to deal with the overfitting problem. The basic approach followed was the Method of Sieves, in which the complexity of the model space was chosen as a function of the number of training samples. In particular, both the denoising and classification problems we looked at consider estimators based on histogram partitions. The size of the partition was an increasing function of the number of training samples. In this lecture, we will refine our learning methods further introduce model selection procedures that automatically adapt to the distribution of the training data, rather than basing the model class solely on the number of samples. This sort of adaptivity will play a major role in the design of more effective classifiers and denoising methods. The key to designing data-adaptive model selection procedures is obtaining useful upper bounds on the estimation error. To this end, we will introduce the idea of "Probably Approximately Correct" learning methods.

7.2 Recap: Method of Sieves

The method of Sieves underpinned our approaches in the denoising problem and in the histogram classification problem. Recall that the basic idea is to define a sequence of model spaces $\mathcal{F}_1$, $\mathcal{F}_2$, ...of increasing

[1] This content is available online at <http://cnx.org/content/m16282/1.2/>.

complexity, and then given the training data $\{X_i, Y_i\}_{i=1}^n$ select a model according to

$$\hat{f}_n = \underset{f \in \mathcal{F}_n}{argmin} \hat{R}_n(f). \tag{7.1}$$

The choice of the model space $\mathcal{F}_n$ (and hence the model complexity and structure) is determined completely by the sample size n, and does not depend on the (empirical) distribution of training data. This is a major limitation of the sieve method. In a nutshell, the method of sieves tells us to average the data in a certain way (e.g., over a partition of $\mathcal{X}$) based on the sample size, independent on the sample values themselves.

In general, learning basically comprises of two things:

1. Averaging data to reduce variability
2. Deciding **where (or how)** to average

Sieves basically force us to deal with (2) **a priori** (before we analyze the training data). This will lead to suboptimal classifiers and estimators, in general. Indeed deciding where/how to average is the really interesting and fundamental aspect of learning; once this is decided we have effectively solved the learing problem. There are at least two possibilities for breaking the rigidity of the method of sieves, as we shall see in the following section.

7.3 Data Adaptive Model Spaces

7.3.1 Structural Risk Minimization (SRM)

The basic idea is to select $\mathcal{F}_n$ based on the training data themselves. Let $\mathcal{F}_1$, $\mathcal{F}_2$, ...be a sequence of model spaces of increasing sizes/complexities with

$$\lim_{k \to \infty} \inf_{f \in \mathcal{F}_k} R(f) = R^*. \tag{7.2}$$

Let

$$\hat{f}_{n,k} = \underset{f \in \mathcal{F}_k}{argmin} \hat{R}_n(f) \tag{7.3}$$

be a function from $\mathcal{F}_k$ that minimizes the empirical risk. This gives us a sequence of selected models $\hat{f}_{n,1}, \hat{f}_{n,2}, \cdots$ Also associate with each set $\mathcal{F}_k$ a value $C_{n,k} > 0$ that measures the complexity or "size" of the set $\mathcal{F}_k$. Typically, $C_{n,k}$ is monotonically increasing with k (since the sets are of increasing complexity) and decreasing with n (since we become more confident with more training data). More precisely, suppose that the $C_{n,k}$ chosen so that

$$P\left(\sup_{f \in \mathcal{F}_k} |\hat{R}_n(f) - R(f)| > C_{n,k} \right) < \delta \tag{7.4}$$

for some small $\delta > 0$. Then we may conclude that with very high probability (at least $1 - \delta$) the empirical risk $\hat{R}_n$ is within $C_{n,k}$ of R uniformly on the class $\mathcal{F}_k$. This type of bound suffices to bound the estimation error (variance) of the model selection process of the form $R(f) \leq \hat{R}_n(f) + C_{n,k}$, and SRM selects the final model by minimizing this bound over all functions in $\bigcup_{k \geq 1} \mathcal{F}_k$. The selected model is given by $\hat{f}_{\hat{n},k}$, where

$$\hat{k} = \underset{k \geq 1}{argmin} \{ \hat{R}_n\left(\hat{f}_{n,k} \right) + C_{n,k} \}. \tag{7.5}$$

A typical example could be the use of VC dimension to characterize the complexity of the collection of model spaces **i.e.,**$C_{n,k}$ is derived from a bound on the estimation error.

7.3.2 Complexity Regularization

Consider a very large class of candidate models $\mathcal{F}$. To each $f \in \mathcal{F}$ assign a complexity value $C_n(f)$. Assume that the complexity value is chosen so that

$$P\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)| > C_n(f)\right) < \delta. \tag{7.6}$$

This probability bound also implies an upper bound on the estimation error and complexity regularization is based on the criterion

$$\hat{f}_n = \underset{f\in\mathcal{F}}{argmin}\{\hat{R}_n(f) + C_n(f)\}. \tag{7.7}$$

Complexity Regularization and SRM are very similar and equivalent in certain instances. A distinguishing feature of SRM and complexity requlari zation techniques is that the complexity and structure of the model is not fixed prior to examining the data; the data aid in the selection of the best complexity. In fact, the key difference compared to the Method of Sieves is that these techniques can allow the data to play an integral role in deciding where and how to average the data.

7.4 Probably Approximately Correct (PAC) learning

Probability bounds of the forms in (7.4) and (7.6) are the foundation for SRM and complexity regularization techniques. The simplest of these bounds are known as PAC bounds in the machine learning community.

7.4.1 Approximation and Estimation Errors

In order to develop complexity regularization schemes we will need to revisit the estimation error / approximation error trade-off. Let $\hat{f}_n = argmin_{f\in\mathcal{F}} \hat{R}_n(f)$ for some space of models $\mathcal{F}$.

$$R\left(\hat{f}_n\right) - R^* = \underbrace{R\left(\hat{f}_n\right) - inf_{f\in\mathcal{F}} R(f)}_{\text{estimation Error}} + \underbrace{inf_{f\in\mathcal{F}} R(f) - R^*}_{\text{approximation error}} \tag{7.8}$$

The approximation error depends on how close f^* is close to $\mathcal{F}$, and without making assumptions, this is unknown. The estimation error is quantifiable, and depends on the complexity or size of $\mathcal{F}$. The error decomposition is illustrated in Figure 7.1. The estimation error quantifies how much we can "trust" the empirical risk minimization process to select a model close to the best in a given class.

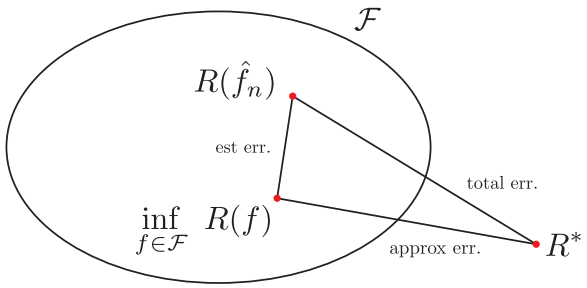


Figure 7.1: Relationship between the errors

Probability bounds of the forms in (7.4) and (7.6) guarantee that the empirical risk is uniformly close to the true risk, and using (7.4) and (7.6) it is possible to show that with high probability the selected model $\hat{f}_n$ satisfies

$$R\left(\hat{f}_n\right) - \inf_{f\in\mathcal{F}_k} R(f) \leq C(n,k) \tag{7.9}$$

or

$$R\left(\hat{f}_n\right) - \inf_{f\in\mathcal{F}_k} R(f) \leq C_n(f). \tag{7.10}$$

7.4.2 The PAC Learning Model

The estimation error will be small if $R\left(\hat{f}_n\right)$ is close to $inf_{f\in\mathcal{F}} R(f)$. PAC learning expresses this as follows. We want $\hat{f}_n$ to be a "probably approximately correct" (PAC) model from $\mathcal{F}$. Formally, we say that $\hat{f}_n$ is ϵ accurate with confidence $1-\delta$, or (ϵ,δ) −PAC for short, if

$$P\left(R\left(\hat{f}_n\right) - \inf_{f\in\mathcal{F}} R(f) > \epsilon\right) < \delta. \tag{7.11}$$

This says that the difference between $R\left(\hat{f}_n\right)$ and $inf_{f\in\mathcal{F}} R(f)$ is greater than ϵ with probability less than δ. Sometimes, especially in the machine learning community, PAC bounds are stated as, "with probability of at least $1-\delta$, $|R\left(\hat{f}_n\right) - inf_{f\in\mathcal{F}} R(f)| \leq \epsilon$"

To introduce PAC bounds, let us consider a simple case. Let $\mathcal{F}$consist of a finite number of models, and let $|\mathcal{F}|$ denote that number. Furthermore, assume that $min_{f\in\mathcal{F}} R(f) = 0$.

Example 7.1
$\mathcal{F}$= set of all histogram classifiers with M bins $\Rightarrow |\mathcal{F}| = 2^M$.

$$\min_{f\in\mathcal{F}} R(f) = 0 \Rightarrow \exists \text{ a classifier in } \mathcal{F} \text{ that has a zero probability of error} \tag{7.12}$$

Theorem 7.1:

Assume $|\mathcal{F}| < \infty$ and $min_{f\in\mathcal{F}} R(f) = 0$, where $R(f) = P(f(X) \neq Y)$. Let $\hat{f}_n = argmin_{f\in\mathcal{F}} \hat{R}_n(f)$, where $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f(X_i)\neq Y_i\}}$. Then for every n and $\epsilon > 0$,

$$P\left(R\left(\hat{f}_n\right) > \epsilon\right) \leq |\mathcal{F}|e^{-n\epsilon} \equiv \delta. \tag{7.13}$$

Proof:
Since $min_{f\in\mathcal{F}} R(f) = 0$, it follows that $\hat{R}_n\left(\hat{f}_n\right) = 0$. In fact, there may be several $f \in \mathcal{F}$ such that $\hat{R}_n(f) = 0$. Let $\mathcal{G} = \{f : \hat{R}_n(f) = 0\}$.

$$\begin{aligned} P\left(R\left(\hat{f}_n\right) > \epsilon\right) &\leq P\left(\bigcup_{f\in\mathcal{G}}\{R(f) > \epsilon\}\right) \\ &= P\left(\bigcup_{f\in\mathcal{F}}\{R(f) > \epsilon,\ \hat{R}_n(f) = 0\}\right) \\ &= P\left(\bigcup_{f\in\mathcal{F}:R(f)>\epsilon}\{\hat{R}_n(f) = 0\}\right) \\ &\leq \sum_{f\in\mathcal{F}:R(f)>\epsilon} P\left(\hat{R}_n(f) = 0\right) \\ &\leq |\mathcal{F}|.(1-\epsilon)^n \end{aligned} \tag{7.14}$$

The last inequality follows from the fact that if $R(f) = P(f(X) \neq Y) > \epsilon$, then the probability that n i.i.d. samples will satisfy $f(X) = Y$ is less than or equal to $(1-\epsilon)^n$. Note that this is simply the probability that $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f(X_i)\neq Y_i\}} = 0$. Finally apply the inequality $1 - x \leq e^{-x}$ to obtain the desired result.

Note that for n sufficiently large, $\delta = |\mathcal{F}|e^{-n\epsilon}$ is arbitrarily small. To achieve a (ϵ, δ)-PAC bound for a desired $\epsilon > 0$ and $\delta > 0$ we require at least $n = \frac{log|\mathcal{F}| - log\delta}{\epsilon}$ training examples.

Corollary 7.1:
Assume that $|\mathcal{F}| < \infty$ and $min_{f\in\mathcal{F}} R(f) = 0$. Then for every n

$$E\left[R\left(\hat{f}_n\right)\right] \leq \frac{1 + log|\mathcal{F}|}{n}. \tag{7.15}$$

Proof:
Recall that for any non-negative random variable Z with finite mean, $E[Z] = \int_0^\infty P(Z > t)\, dt$.

This follows from an application of integration by parts.

$$\begin{aligned} E\left[R\left(\hat{f}_n\right)\right] &= \int_0^\infty P\left(R\left(\hat{f}_n\right) > t\right) dt \\ &= \int_0^u \underbrace{P\left(R\left(\hat{f}_n\right) > t\right)}_{\leq 1} dt + \int_u^\infty P\left(R\left(\hat{f}_n\right) > t\right) dt, \quad \text{for any } u > 0 \\ &\leq u + |\mathcal{F}| \int_u^\infty e^{-nt} dt \\ &= u + \frac{|\mathcal{F}|}{n} e^{-nu} \end{aligned} \tag{7.16}$$

Minimizing with respect to u produces the smallest upper bound with $u = \frac{log|\mathcal{F}|}{n}$

Chapter 8

Chernoff's Bound and Hoeffding's Inequality[1]

8.1 Introduction

8.1.1 Motivation

In the last lecture (Chapter 7) we consider a learning problem in which the optimal function belonged to a finite class of functions. Specifically, for some collection of functions $\mathcal{F}$with finite cardinality $|\mathcal{F}| \leq \infty$, we have

$$\underset{f \in \mathcal{F}}{min} R(f) = 0 \Rightarrow f^* \in \mathcal{F}. \tag{8.1}$$

This is almost always not the situation in the real-world learning problems. Let us suppose we have a finite collection of candidate functions $\mathcal{F}$. Furthermore, we do not assume that the optimal function f^*, which satisfies

$$R(f^*) = \underset{f}{inf} R(f) \tag{8.2}$$

where the inf is taken over all measurable functions, is a member of $\mathcal{F}$. That is, we make few, if any, assumptions about f^*. This situation is sometimes termed as **Agnostic Learning**. The root of the word agnostic literally means **not known**. The term agnostic learning is used to emphasize the fact that often, perhaps usually, we may have no prior knowledge about f^*. The question then arises about how we can reasonably select an $f \in \mathcal{F}$ in this setting.

8.1.2 The Problem

The PAC style bounds discussed in the previous lecture (Chapter 7), offer some help. Since we are selecting a function based on the empirical risk, the question is how close is $\hat{R_n}(f)$ to $R(f) \forall f \in \mathcal{F}$. In other words, we wish that the empirical risk is a good indicator of the true risk for every function in $\mathcal{F}$. If this is case, the selection of f that minimizes the empirical risk

$$\hat{f_n} = arg \underset{f \in \mathcal{F}_n}{min} \hat{R_n}(f) \tag{8.3}$$

[1] This content is available online at <http://cnx.org/content/m16264/1.2/>.

should also yield a small true risk, that is, $R\left(\hat{f}_n\right)$ should be close to $min_{f\in\mathcal{F}}R(f)$. Finally, we can thus state our desired situation as

$$P\left(\underset{f\in\mathcal{F}_n}{max}|\,\hat{R}_n(f)-R(f)\,|>\varepsilon\right) \;<\; \delta, \tag{8.4}$$

for small values of ε and δ. In other words, with probability at least $1-\delta$, $|\,\hat{R}_n(f)-R(f)\,|>\varepsilon$, $\forall f\in\mathcal{F}$. In this lecture, we will start to develop bounds of this form. First we will focus on bounding $P\left(|\,\hat{R}_n(f)-R(f)\,|>\varepsilon\right)$ for one fixed $f\in\mathcal{F}$.

8.2 Developing Initial Bounds

To begin, let us recall the definition of empirical risk for $\{X_i,Y_i\}_{i=1}^n$ be a collection of training data. Then the empirical risk is defined as

$$\hat{R}_n(f)=\frac{1}{n}\sum_{i=1}^{n}\ell\left(f\left(X_i\right),Y_i\right). \tag{8.5}$$

Note that since the training data $\{X_i,Y_i\}_{i=1}^n$ are assumed to be **i.i.d.** pairs, the terms in the sum are **i.i.d** random variables.

Let

$$L_i=\ell\left(f\left(X_i\right),Y_i\right). \tag{8.6}$$

The collection of losses $\{L_i\}_{i=1}^n$ is **i.i.d** according to some unknown distribution (depending on the unknown joint distribution of (X,Y) and the loss function). The expectation of L_i is $E\left[\ell\left(f\left(X_i\right),Y_i\right)\right]=E\left[\ell\left(f\left(X\right),Y\right)\right]=R(f)$, the true risk of f. For now, let's assume that f is fixed.

$$E\left[\hat{R}_n(f)\right]=\frac{1}{n}\sum_{i=1}^{n}E\left[\ell\left(f\left(X_i\right),Y_i\right)\right]=\frac{1}{n}\sum_{i=1}^{n}E\left[L_i\right]=R(f) \tag{8.7}$$

We know from the strong law of large numbers that the average (or empirical mean) $\hat{R}_n(f)$ converges almost surely to the true mean $R(f)$. That is, $\hat{R}_n(f)\to R(f)$ almost surely as $n\to\infty$. The question is how fast.

8.3 Concentration of Measure Inequalities

Concentration inequalities are upper bounds on how fast empirical means converge to their ensemble counterparts, in probability. The area of the shaded tail regions in Figure 1 is $P\left(|\,\hat{R}_n(f)-R(f)\,|>\varepsilon\right)$. We are interested in finding out how fast this probability tends to zero as $n\to\infty$.

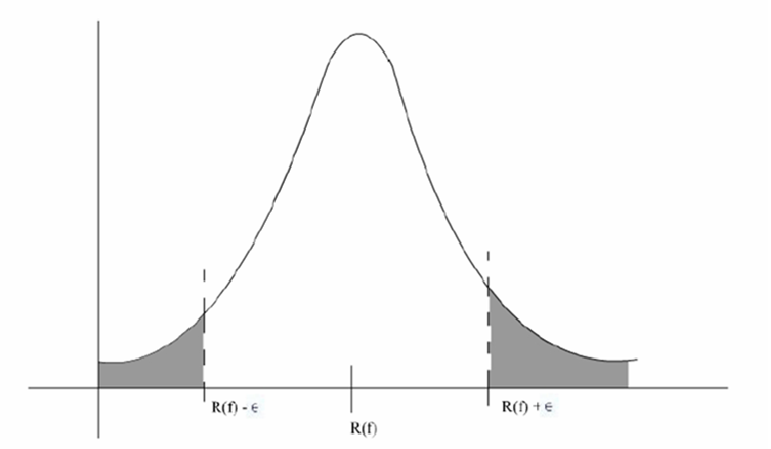


Figure 8.1: Distribution of $\hat{R}_n(f)$

At this stage, we recall **Markov's Inequality**. Let Z be a nonnegative random variable.

$$\begin{array}{rcl} E[Z] & = & \int_0^\infty zp(z)\,dz \\ & = & \int_0^t zp(z)\,dz + \int_u^\infty zp(z)\,dz \\ & \geq & 0 + t\int_t^\infty zp(z)\,dz \\ & = & tP(Z \geq t) \\ \Rightarrow P(Z \geq t) & \leq & \frac{E[Z]}{t} \\ \Rightarrow P\left(Z^2 \geq t^2\right) & \leq & \frac{E\left[Z^2\right]}{t^2} \end{array} \tag{8.8}$$

Take

$$Z = |\,\hat{R}_n(f) - R(f)\,| \quad \text{and} \quad t = \varepsilon \tag{8.9}$$

$$\begin{array}{rcl} P\left(|\,\hat{R}_n(f) - R(f)\,| \geq \varepsilon\right) & \leq & \frac{E\left[|\hat{R}_n(f) - R(f)|^2\right]}{\varepsilon^2} \\ & \leq & \frac{\mathrm{var}\left(\hat{R}_n(f)\right)}{\varepsilon^2} \\ & = & \frac{\sum_{i=1}^n \mathrm{var}\left(\frac{L_i}{n}\right)}{\varepsilon^2} \\ & = & \frac{\mathrm{var}(\ell(X),Y)}{n\varepsilon^2} \\ & = & \frac{\sigma_L^2}{n\varepsilon^2} \end{array} \quad . \tag{8.10}$$

So, the probability goes to zero at a rate of at least n^{-1}. However, it turns out that this is an extremely

loose bound. According to the Central Limit Theorem

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L_i \to N\left(R(f), \frac{\sigma_L^2}{n}\right) \quad \text{as} \quad n \to \infty \tag{8.11}$$

in distribution. This suggests that for large values of n,

$$P\left(|\hat{R}_n(f) - R(f)| \geq \varepsilon\right) \approx O\left(e^{-\frac{n\varepsilon^2}{2\sigma_L^2}}\right). \tag{8.12}$$

That is, the Gaussian tail probability is tending to zero exponentially fast.

8.4 Chernoff's Bound

Note that for any nonnegative random variable Z and $t > 0$,

$$P(Z \geq t) = P\left(e^{sZ} \geq e^{st}\right) \leq \frac{E\left[e^{sZ}\right]}{e^{st}}, \ \forall s > 0 \text{ by Markov's inequality.} \tag{8.13}$$

Chernoff's bound is based on finding the value of s that minimizes the upper bound. If Z is a sum of independent random variables. For example, say

$$Z = \sum_{i=1}^{n} \left(\ell\left(f\left(X_i\right), Y_i\right) - R(f)\right) = n\left(\hat{R}_n(f) - R(f)\right) \tag{8.14}$$

then the bound becomes

$$P\left(\sum_{i=1}^{n}\left(L_i - E\left[L_i\right]\right) \geq t\right) \quad \leq \quad e^{-st}E\left[e^{s\sum_{i=1}^{n}(L_i - E[L_i])}\right] \quad \leq \tag{8.15}$$
$$e^{-st}\prod_{i=1}^{n} E\left[e^{s(L_i - E[L_i])}\right], \text{ from independence.}$$

Thus, the problem of finding a tight bound boils down to finding a good bound for $E\left[s^{s(L_i - E[L_i])}\right]$. Chernoff ('52), first studied this situation for binary random variables. Then, Hoeffding ('63) derived a more general result for arbitrary bounded random variables.

8.5 Hoeffding's Indequality

Theorem 8.1: Hoeffding's Inequality
Let $Z_1, Z_2, ..., Zn$ be independent bounded random variables such that $Z_i \in [a_i, b_i]$ with probability 1. Let $S_n = \sum_{i=1}^{n} Z_i$. Then for any $t > 0$, we have

$$P\left(|S_n - E\left[S_n\right]| \geq t\right) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}. \tag{8.16}$$

Proof:
The key to proving Hoeffding's inequality is the following upper bound: if Z is a random variable with $E[Z] = 0$ and $a \leq Z \leq b$, then

$$E\left[e^{sZ}\right] \leq e^{\frac{s^2(b-a)^2}{8}}. \tag{8.17}$$

This upper bound is derived as follows. By the convexity of the exponential function,

$$e^{sz} \leq \frac{z-a}{b-a}e^{sb} + \frac{b-z}{b-a}e^{sa}, \text{ for } a \leq z \leq b. \tag{8.18}$$

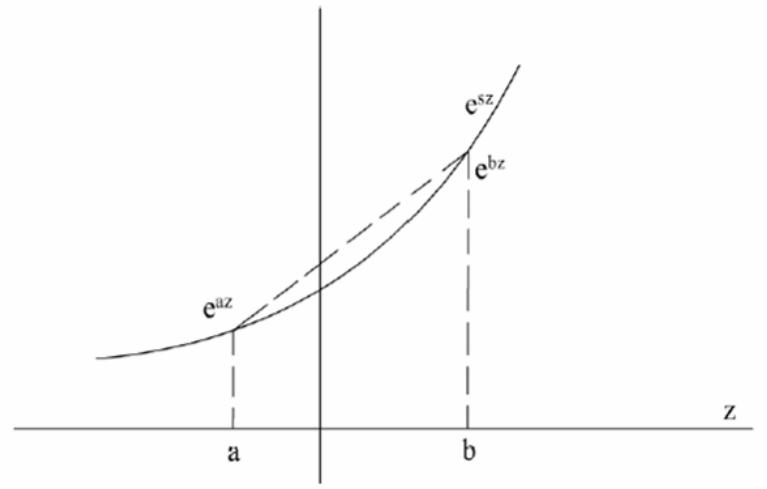


Figure 8.2: Convexity of exponential function.

Thus,

$$\begin{array}{rcl} E\left[e^{sZ}\right] & \leq & E\left[\frac{Z-a}{b-a}\right]e^{sb} + E\left[\frac{b-Z}{b-a}\right]e^{sa} \\ & = & \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}\ , \ \text{ since } E\left[Z\right]=0 \\ & = & \left(1-\theta+\theta e^{s(b-a)}\right)e^{-\theta s(b-a)}\ , \ \text{ where } \theta = \frac{-a}{b-a} \end{array} . \tag{8.19}$$

Now let

$$u = s\left(b-a\right) \quad \text{and define} \quad \phi\left(u\right) \equiv -\theta u + log\left(1-\theta+\theta e^{u}\right). \tag{8.20}$$

Then we have

$$E\left[e^{sZ}\right] \leq \left(1-\theta+\theta e^{s(b-a)}\right)e^{-\theta s(b-a)} = e^{\phi(u)}. \tag{8.21}$$

To minimize the upper bound let's express $\phi\left(u\right)$ in a Taylor's series with remainder :

$$\phi\left(u\right) = \phi\left(0\right) + u\phi^{'}\left(0\right) + \frac{u^2}{2}\phi^{''}\left(v\right) \ \text{ for some } v \in [0,u] \tag{8.22}$$

$$\begin{array}{rcl} \phi^{'}\left(u\right) & = & -\theta + \frac{\theta e^{u}}{1-\theta+\theta e^{u}} \Rightarrow \phi^{'}\left(u\right) = 0 \\ \phi^{''}\left(u\right) & = & \frac{\theta e^{u}}{1-\theta+\theta e^{u}} - \frac{\theta e^{u}}{\left(1-\theta+\theta e^{u}\right)^2} \\ & = & \frac{\theta e^{u}}{1-\theta+\theta e^{u}}\left(1-\frac{\theta e^{u}}{1-\theta+\theta e^{u}}\right) \\ & = & \rho\left(1-\rho\right) \end{array} . \tag{8.23}$$

Now, $\phi^{''}\left(u\right)$ is maximized by

$$\rho = \frac{\theta e^{u}}{1-\theta+\theta e^{u}} = \frac{1}{2} \Rightarrow \phi^{''}\left(u\right) \leq \frac{1}{4}. \tag{8.24}$$

So,

$$\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8} \tag{8.25}$$

$$\Rightarrow E\left[e^{sZ}\right] \leq e^{\frac{s^2(b-a)^2}{8}}. \tag{8.26}$$

Now, we can apply this upper bound to derive Hoeffding's inequality.

$$\begin{aligned} P(S_n - E[S_n] \geq t) &\leq e^{-st}\prod_{i=1}^n E\left[e^{s(L_i - E[L_i])}\right] \\ &\leq e^{-st}\prod_{i=1}^n e^{\frac{s^2(b_i-a_i)^2}{8}} \\ &= e^{-st}e^{s^2\sum_{i=1}^n \frac{(b_i-a_i)^2}{8}} \\ &= e^{\frac{-2t^2}{\sum_{i=1}^n (b_i-a_i)^2}} \\ &\text{by choosing} \quad s = \frac{4t}{\sum_{i=1}^n (b_i-a_i)^2} \end{aligned} \tag{8.27}$$

Similarly, $P(E[S_n] - S_n \geq t) \leq e^{\frac{-2t^2}{\sum_{i=1}^n (b_i-a_i)^2}}$. This completes the proof of the Hoeffding's theorem.

Example

Application

Let $Z_i = 1_{f(X_i)\neq Y_i} - R(f)$, as in the classification problem. Then for a fixed f, it follows from Hoeffding's inequality (i.e., Chernoff's bound in this special case) that

$$\begin{aligned} P\left(|\hat{R}_n(f) - R(f)| \geq \varepsilon\right) &= P\left(\tfrac{1}{n}|S_n - E[S_n]| \geq \varepsilon\right) \\ &= P(|S_n - E[S_n]| \geq n\varepsilon) \quad . \\ &\leq 2e^{-\frac{2(n\varepsilon)^2}{n}} \\ &= 2e^{-2n\varepsilon^2} \end{aligned} \tag{8.28}$$

Now, we want a bound like this to hold uniformly for all $f \in \mathcal{F}$. Assume that $\mathcal{F}$ is a finite collection of models and let $|\mathcal{F}|$ denote its cardinality. We would like to bound the probability that $max_{f\in\mathcal{F}}|\hat{R}_n(f) - R(f)| \geq \varepsilon$. Note that the event

$$\{\max_{f\in\mathcal{F}}|\hat{R}_n(f) - R(f)| \geq \varepsilon\} \equiv \{\bigcup_{f\in\mathcal{F}}|\hat{R}_n(f) - R(f)| \geq \varepsilon\}. \tag{8.29}$$

Therefore

$$P\left(\max_{f\in\mathcal{F}}|\hat{R}_n(f) - R(f)| \geq \varepsilon\right) = P\left(\bigcup_{f\in\mathcal{F}}|\hat{R}_n(f) - R(f)| \geq \varepsilon\right) \leq \tag{8.30}$$

$\sum_{f\in F} P\left(|\hat{R}_n(f) - R(f)| \geq \varepsilon\right)$, the "union of events" bound $\leq$

$2|F|e^{-2n\varepsilon^2}$, by Hoeffding's inequality.

Thus, we have shown that with probability at least $1 - 2|F|e^{-2n\varepsilon^2}$, $\forall f \in \mathcal{F}$

$$|\hat{R}_n(f) - R(f)| < \varepsilon. \tag{8.31}$$

And accordingly, we can be reasonably confident in selecting f from $\mathcal{F}$ based on the empirical risk function $\hat{R}_n$.

Chapter 9

Classification Error Bounds[1]

9.1 Recap: Classifier design

Given a set of training data $\{X_i, Y_i\}_{i=1}^n$ and a finite collection of candidate functions $\mathcal{F}$, select $\hat{f}_n \in \mathcal{F}$ that (hopefully) is a good predictor for future cases. That is

$$\hat{f}_n = \underset{f \in \mathcal{F}}{argmin} \hat{R}_n(f) \tag{9.1}$$

where $\hat{R}_n(f)$ is the empirical risk. For any particular $f \in \mathcal{F}$, the corresponding empirical risk is defined as

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{f(X_i) \neq Y_i\}}. \tag{9.2}$$

9.2 Hoeffding's inequality

Hoeffding's inequality (Chernoff's bound in this case) allows us to gauge how close $\hat{R}_n(f)$ is to the true risk of f, $R(f)$, in probability

$$P\left(|\hat{R}_n(f) - R(f)| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}. \tag{9.3}$$

Since our selection process involves deciding among all $f \in \mathcal{F}$, we would like to gauge how close the empirical risks are to their expected values. We can do this by studying the probability that one or more of the empirical risks deviates significantly from its expected value. This is captured by the probability

$$P\left(\underset{f \in \mathcal{F}}{max} |\hat{R}_n(f) - R(f)| \geq \varepsilon\right). \tag{9.4}$$

Note that the event

$$\underset{f \in \mathcal{F}}{max} |\hat{R}_n(f) - R(f)| \geq \varepsilon \tag{9.5}$$

[1] This content is available online at <http://cnx.org/content/m16265/1.2/>.

is equivalent to union of the events

$$\bigcup_{f\in\mathcal{F}}\{|\hat{R}_n(f)-R(f)|\geq\varepsilon\}. \tag{9.6}$$

Therefore, we can use Bonferonni's bound (aka the "union of events" or "union" bound) to obtain

$$\begin{aligned} P\left(\max_{f\in\mathcal{F}}|\hat{R}_n(f)-R(f)|\geq\varepsilon\right) &= P\left(\bigcup_{f\in\mathcal{F}}|\hat{R}_n(f)-R(f)|\geq\varepsilon\right) \\ &\leq \sum_{f\in\mathcal{F}}P\left(|\hat{R}_n(f)-R(f)|\geq\varepsilon\right) \\ &\leq \sum_{f\in\mathcal{F}}2e^{-2n\varepsilon^2} \\ &= 2|\mathcal{F}|e^{-2n\varepsilon^2} \end{aligned} \tag{9.7}$$

where $|\mathcal{F}|$ is the number of classifiers in $\mathcal{F}$. In the proof of Hoeffding's inequality we also obtained a one-sided inequality that implied

$$P\left(R(f)-\hat{R}_n(f)\geq\varepsilon\right)\leq e^{-2n\varepsilon^2} \tag{9.8}$$

and hence

$$P\left(\max_{f\in\mathcal{F}}R(f)-\hat{R}_n(f)\geq\varepsilon\right)\leq|\mathcal{F}|e^{-2n\varepsilon^2}. \tag{9.9}$$

We can restate the inequality above as follows, For all $f\in\mathcal{F}$ and for all $\delta>0$ with probability at least $1-\delta$

$$R(f)\leq\hat{R}_n(f)+\sqrt{\frac{log|\mathcal{F}|+log(1/\delta)}{2n}}. \tag{9.10}$$

This follows by setting $\delta=|\mathcal{F}|e^{-2n\varepsilon^2}$ and solving for ε. Thus with a high probability $(1-\delta)$, the true risk for all $f\in\mathcal{F}$ is bounded by the empirical risk of f plus a constant that depends on $\delta>0$, the number of training samples n, and the size $\mathcal{F}$. Most importantly the bound does not depend on the unknown distribution P_{XY}. Therefore, we can call this a **distribution-free** bound.

9.3 Error Bounds

We can use the **distribution-free** bound above to obtain a bound on the expected performance of the minimum empirical risk classifier

$$\hat{f}_n=\underset{f\in\mathcal{F}}{argmin}\hat{R}_n(f). \tag{9.11}$$

We are interested in bounding

$$E\left[R\left(\hat{f}_n\right)\right]-\min_{f\in\mathcal{F}}R(f) \tag{9.12}$$

the expected risk of $\hat{f}_n$ minus the minimum risk for all $f\in\mathcal{F}$. Note that this difference is always non-negative since $\hat{f}_n$ is at best as good as

$$f^*=\underset{f\in\mathcal{F}}{argmin}R(f). \tag{9.13}$$

Recall that $\forall f \in \mathcal{F}$ and $\forall \delta > 0$, with probability at least $1 - \delta$

$$R(f) \leq \hat{R}_n(f) + C(\mathcal{F}, n, \delta) \tag{9.14}$$

where

$$C(\mathcal{F}, n, \delta) = \sqrt{\frac{log|\mathcal{F}| + log(1/\delta)}{2n}}. \tag{9.15}$$

In particular, since this holds for all $f \in \mathcal{F}$ including $\hat{f}_n$,

$$R\left(\hat{f}_n\right) \leq \hat{R}_n\left(\hat{f}_n\right) + C(\mathcal{F}, n, \delta) \tag{9.16}$$

and for any other $f \in \mathcal{F}$

$$R\left(\hat{f}_n\right) \leq \hat{R}_n(f) + C(\mathcal{F}, n, \delta) \tag{9.17}$$

since $\hat{R}_n\left(\hat{f}_n\right) \leq \hat{R}_n(f) \forall f \in \mathcal{F}$. In particular,

$$R\left(\hat{f}_n\right) \leq \hat{R}_n(f^*) + C(\mathcal{F}, n, \delta) \tag{9.18}$$

where $f^* = argmin_{f \in \mathcal{F}} R(f)$.

Let Ω denote the set of events on which the above inequality holds. Then by definition

$$P(\Omega) \geq 1 - \delta. \tag{9.19}$$

We can now bound $E\left[R\left(\hat{f}_n\right)\right] - R(f^*)$ as follows

$$\begin{aligned} E\left[R\left(\hat{f}_n\right)\right] - R(f^*) &= E\left[R\left(\hat{f}_n\right) - \hat{R}_n(f^*) + \hat{R}_n(f^*) - R(f^*)\right] \\ &= E\left[R\left(\hat{f}_n\right) - \hat{R}_n(f^*)\right] \end{aligned} \tag{9.20}$$

since $E\left[\hat{R}_n(f^*)\right] = R(f^*)$. The quantity above is bounded as follows.

$$\begin{aligned} E\left[R\left(\hat{f}_n\right) - \hat{R}_n(f^*)\right] &= E\left[R\left(\hat{f}_n\right) - \hat{R}_n(f^*)|\Omega\right] P(\Omega) + \\ E\left[R\left(\hat{f}_n\right) - \hat{R}_n(f^*)|\Omega\right] P\left(\overline{\Omega}\right) &\leq E\left[R\left(\hat{f}_n\right) - \hat{R}_n(f^*)|\Omega\right] + \delta \end{aligned} \tag{9.21}$$

since $P(\Omega) \leq 1$, $1 - P(\Omega) \leq \delta$ and $R\left(\hat{f}_n\right) - \hat{R}_n(f^*) \leq 1$

$$\begin{aligned} E\left[R\left(\hat{f}_n\right) - \hat{R}_n(f^*)\,|\Omega\right] &\leq E\left[R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right)|\Omega\right] \\ &\leq C(\mathcal{F}, n, \delta) \end{aligned} . \tag{9.22}$$

Thus

$$E\left[R\left(\hat{f}_n\right) - \hat{R}_n(f^*)\right] \leq C(\mathcal{F}, n, \delta) + \delta. \tag{9.23}$$

So we have

$$E\left[R\left(\hat{f}_n\right)\right] - \min_{f\in\mathcal{F}} R(f) \leq \sqrt{\frac{log|\mathcal{F}| + log(1/\delta)}{2n}} + \delta,\ \forall\delta > 0. \tag{9.24}$$

In particular, for $\delta = \sqrt{1/n}$, we have

$$\begin{aligned} E\left[R\left(\hat{f}_n\right)\right] - \min_{f\in\mathcal{F}} R(f) &\leq \sqrt{\tfrac{log|\mathcal{F}|+logn}{2n}} + \tfrac{1}{\sqrt{n}} \\ &\leq \sqrt{\tfrac{log|\mathcal{F}|+logn+2}{n}}, \quad \text{since} \quad \sqrt{x}+\sqrt{y} \leq \sqrt{2}\sqrt{x+y},\ \forall\ x, y > 0 \end{aligned} . \tag{9.25}$$

9.4 Application: Histogram Classifier

Let $\mathcal{F}$ be the collection of all classifiers with M equal volume cells. Then $|\mathcal{F}| = 2^M$, and the histogram classification rule

$$\hat{f}_n = \underset{f\in\mathcal{F}}{argmin}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{f(X_i)\neq Y_i\}}\right) \tag{9.26}$$

satisfies

$$E\left[R\left(\hat{f}_n\right)\right] - \min_{f\in\mathcal{F}} R(f) \leq \sqrt{\frac{Mlog2 + 2 + logn}{n}} \tag{9.27}$$

which suggests the choice $M = log_2 n$ (balancing $Mlog2$ with $logn$), resulting in

$$E\left[R\left(\hat{f}_n\right)\right] - \min_{f\in\mathcal{F}} R(f) = O\left(\sqrt{\frac{logn}{n}}\right). \tag{9.28}$$

Chapter 10

Error Bounds in Countably Infinite Spaces[1]

10.1 Introduction

In the last lecture (Chapter 9), we studied bounds of the following form: for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{log|\mathcal{F}| + log\left(\frac{1}{\delta}\right)}{2n}} \quad , \quad \forall f \in \mathcal{F} \tag{10.1}$$

which led to upper bounds on the estimation error of the form

$$E\left[R\left(\hat{f}_n\right)\right] - \min_{f \in F} R(f) \leq \sqrt{\frac{log|\mathcal{F}| + log(n) + 2}{n}}. \tag{10.2}$$

The key assumptions made in deriving the error bounds were:

(i): bounded loss function
(ii): finite collection of candidate functions

The bounds are valid for every P_{XY} and are called distribution-free .

10.2 Deriving Bounds for Countably Infinite Spaces

In this lecture we will generalize the previous results in a powerful way by developing bounds applicable to possibly infinite collections of candidates. To start let us suppose that $\mathcal{F}$ is a countable, possibly infinite, collection of candidate functions. Assign a positive number c(f) to each $f \in \mathcal{F}$, such that

$$\sum_{f \in \mathcal{F}} e^{-c(f)} < \infty. \tag{10.3}$$

The numbers c(f) can be interpreted as

(i): measures of complexity
(ii): -log of prior probabilities
(iii): codelengths

[1] This content is available online at <http://cnx.org/content/m16271/1.2/>.

In particular, if P(f) is the prior probability of f then

$$e^{-(-logp(f))} = p(f) \tag{10.4}$$

so $c(f) \equiv -logp(f)$ produces

$$\sum_{f\in\mathcal{F}} e^{-c(f)} = \sum_{f\in\mathcal{F}} p(f) = 1. \tag{10.5}$$

Now recall Hoeffding's inequality. For each f and every $\varepsilon > 0$

$$P\left(R(f) - \hat{R}_n(f) \geq \varepsilon\right) \leq e^{-2n\varepsilon^2} \tag{10.6}$$

or for every $\delta > 0$

$$P\left(R(f) - \hat{R}_n(f) \geq \sqrt{\frac{log\left(\frac{1}{\delta}\right)}{2n}}\right) \leq \delta. \tag{10.7}$$

Suppose $\delta > 0$ is specified. Using the values c(f) for $f \in \mathcal{F}$, define

$$\delta(f) = e^{-c(f)}\delta. \tag{10.8}$$

Then we have

$$P\left(R(f) - \hat{R}_n(f) \geq \sqrt{\frac{log\left(\frac{1}{\delta(f)}\right)}{2n}}\right) \leq \delta(f). \tag{10.9}$$

Furthermore we can apply the union bound as follows

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}}\{R(f) - \hat{R}_n(f) - \sqrt{\tfrac{log(1/\delta(f))}{2n}}\} \geq 0\right) &\leq P\left(\textstyle\bigcup_{f\in\mathcal{F}} R(f) - \hat{R}_n(f) \geq \sqrt{\frac{log\left(\frac{1}{\delta(f)}\right)}{2n}}\right) \\ &\leq \textstyle\sum_{f\in\mathcal{F}} P\left(R(f) - \hat{R}_n(f) \geq \sqrt{\frac{log\left(\frac{1}{\delta(f)}\right)}{2n}}\right) \\ &\leq \textstyle\sum_{f\in\mathcal{F}} \delta(f) = \sum_{f\in\mathcal{F}} e^{-c(f)}\delta = \delta \end{aligned}. \tag{10.10}$$

So for any $\delta > 0$ with probability at least $1 - \delta$, we have that $\forall f \in \mathcal{F}$

$$\begin{aligned} R(f) &\leq \hat{R}_n(f) + \sqrt{\frac{log\left(\frac{1}{\delta(f)}\right)}{2n}} \\ &= \hat{R}_n(f) + \sqrt{\frac{c(f)+log\left(\frac{1}{\delta}\right)}{2n}} \end{aligned}. \tag{10.11}$$

Special Case
Suppose $\mathcal{F}$ is finite and $c(f) = log|\mathcal{F}| \quad \forall f \in \mathcal{F}$. Then

$$\sum_{f\in\mathcal{F}} e^{-c(f)} = \sum_{f\in\mathcal{F}} e^{-log|\mathcal{F}|} = \sum_{f\in\mathcal{F}} \frac{1}{|\mathcal{F}|} = 1 \tag{10.12}$$

and

$$\delta(f) = \frac{\delta}{|\mathcal{F}|} \tag{10.13}$$

which implies that for any $\delta > 0$ with probability at least $1 - \delta$, we have

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{log|\mathcal{F}| + log\left(\frac{1}{\delta(f)}\right)}{2n}} \quad , \qquad \forall f \in \mathcal{F}. \tag{10.14}$$

Note that this is precisely the bound we derived in the last lecture (Chapter 9).

Choosing $c(f)$

The generalized bounds allow us to handle countably infinite collections of candidate functions, but we require that

$$\sum_{f \in \mathcal{F}} e^{-c(f)} < \infty. \tag{10.15}$$

Of course, if $c(f) = -logp(f)$ where $p(f)$ is a proper prior probability distribution then we have

$$\sum_{f \in \mathcal{F}} e^{-c(f)} = 1. \tag{10.16}$$

However, it may be difficult to design a probability distribution over an infinite class of candidates. The coding perspective provides a very practical means to this end.

Assume that we have assigned a uniquely decodable binary code to each $f \in \mathcal{F}$, and let $c(f)$ denote the codelength for f. That is, the code for f is $c(f)$ bits long. A very useful class of uniquely decodable codes are called prefix codes .

Definition 10.1: Prefix Code

A code is called a prefix code if no codeword is a prefix of any other codeword.

Example: From Cover & Thomas '91

Consider an alphabet of symbols, say A, B, C, and D and the codebooks below

Symbol	Singular Codebook	Nonsingular But Not Uniquely Decodable	Uniquely Decodable But Not a Prefix Code	Prefix Code
A	0	0	10	0
B	0	010	00	10
C	0	01	11	110
D	0	10	110	1110

Figure 10.1

In the singular codebook we assign the same codeword to each symbol - a system that is obviously flawed! In the second case, the codes are not singular but the codeword 010 could represent B or CA or AD. Hence it is not a uniquely decodable codebook.

The third and fourth cases are both examples of uniquely decodable codebooks, but the fourth has the added feature that no codeword is a prefix of another. Prefix codes can be decoded from left to right since each codeword is "self-punctuating" - in this case with a zero to indicate the end of each word.

To design a uniquely decodable codebook in general is as challenging as the problem of selecting $c(f)$ to satisfy

$$\sum_{f \in \mathcal{F}} e^{-c(f)} < \infty. \tag{10.17}$$

However, prefix codes can often be easily designed or specified and they are inherently decodable. Moreover, prefix codes satisfy an important inequality called the Kraft Inequality .

10.3 The Kraft Inequality

For any binary prefix code, the codeword lengths c_1, c_2, ... satisfy

$$\sum_{i=1}^{\infty} 2^{-c_i} \leq 1. \tag{10.18}$$

Conversely, given any c_1, c_2, ... satisfying the inequality above we can construct a prefix code with these codeword lengths. We will prove this result a bit later, but now let's see how this is useful in our learning problem.

Assume that we have assigned a binary prefix codeword to each $f \in \mathcal{F}$, and let c(f) denote the bit-length of the codeword for f. Set $\delta(f) = 2^{-c(f)}\delta$. Then

$$\begin{array}{rcl} P\left(\bigcup_{f\in\mathcal{F}} R(f) - \hat{R}_n(f) \geq \sqrt{\frac{log\left(\frac{1}{\delta(f)}\right)}{2n}}\right) & \leq & \sum_{f\in\mathcal{F}} P\left(R(f) - \hat{R}_n(f) \geq \sqrt{\frac{log\left(\frac{1}{\delta(f)}\right)}{2n}}\right) \\ & \leq & \sum_{f\in\mathcal{F}} \delta(f) = \sum_{f\in\mathcal{F}} 2^{-c(f)}\delta = \delta \end{array} . \tag{10.19}$$

This implies that for any $\delta > 0$ with probability at least $1-\delta$ we have $\forall f \in \mathcal{F}$

$$\begin{array}{rcl} R(f) & \leq & \hat{R}_n(f) + \sqrt{\frac{log\left(\frac{1}{\delta(f)}\right)}{2n}} \\ & = & \hat{R}_n(f) + \sqrt{\frac{c(f)log2 + log\left(\frac{1}{\delta}\right)}{2n}} \end{array} . \tag{10.20}$$

Application

Let $\mathcal{F}_1$, $\mathcal{F}_1$, ... be a sequence of finite sets of candidate functions with $|\mathcal{F}_1| < |\mathcal{F}_1| < ...$ We can design prefix codes as follows. Use the codes 0, 10, 110, 1110, ... to encode the subscript i in $|\mathcal{F}_i|$. For each class $|\mathcal{F}_i|$, construct a set of binary codewords of length $\lceil log_2|\mathcal{F}|\rceil$ to uniquely encode each function in $\mathcal{F}_i$. Then, encode any given function f by first using the code for i corresponding to the smallest $\mathcal{F}_i$ that f belongs to, followed by the length $\lceil log_2|\mathcal{F}|\rceil$ codeword for $f \in \mathcal{F}_i$. This is a prefix code.

Example 10.1: Histogram Classifiers

X=[0,1]^d, Y={0,1}. Let $\mathcal{F}_k$, k=1, 2, ... denote the collection of histogram classification rules with k equal volume bins. We can use the following codebook for the index k.

k	Prefix Code
1	0
2	10
3	110
4	1110
.	.
.	.
.	.

Figure 10.2

And follow this codeword with $k = log_2|\mathcal{F}_k|$ bits to indicate which of the 2^k possible histogram rules is under consideration. Thus for any $f \in \mathcal{F}_k$ for some k ≥ 1 there is a prefix code of length

$$c(f) = k + k = 2k \quad bits. \tag{10.21}$$

It follows that for any $\delta > 0$ with probability at least $1 - \delta$ we have $\forall f \in \bigcup_{k \geq 1} \mathcal{F}_k$

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{2k_f log2 + log\left(\frac{1}{\delta}\right)}{2n}} \tag{10.22}$$

where k_f is the number of bins in histogram corresponding to f. Contrast with the bound we had for the class of m bin histograms alone: with probability $\geq 1 - \delta$, $\forall f \in \mathcal{F}_m$

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{mlog2 + log\left(\frac{1}{\delta(f)}\right)}{2n}}. \tag{10.23}$$

Notice the bound for all histograms rules is almost as good as the bound for only the m-bin rules. That is, when $k_f = m$ the bounds are within a factor of $\sqrt{2}$. On the other hand, the new bound is a big improvement, since it also gives us a guide for selecting the number of bins.

Proof 10.1: Proof of the Kraft Inequality
We will prove that for any binary prefix code, the codeword lengths $c_1, c_2, \ldots$ satisfy $\sum_{k \geq 1} 2^{-c_k} \leq 1$. The converse is easy to prove also, but it not central to our purposes here (for a proof, see Cover & Thomas '91). Consider a binary tree like the one shown below.

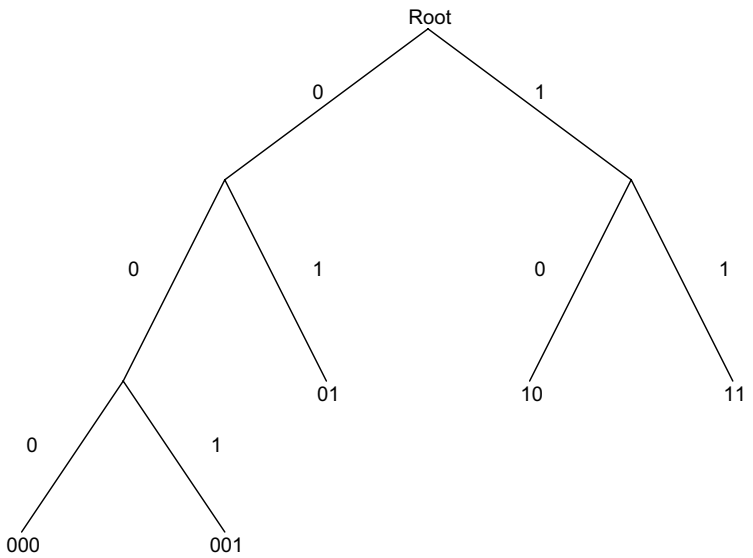


Figure 10.3

The sequence of bit values leading from the root to a leaf of the tree represents a codeword. The prefix condition implies that no codeword is a descendant of any other codeword in the tree. Let c_{max} be the length of the longest codeword (also the number of branches to the deepest leaf) in the tree.

Consider a leaf i in the tree at level c_i. This leaf would have $2^{c_{max}-c_i}$ descendants at level c_{max}. Furthermore, for each leaf the set of possible descendants at level c_{max} is disjoint (since no codeword can be a prefix of another). Therefore, since the total number of possible leafs at level c_{max} is $2^{c_{max}}$, we have

$$\sum_{i \in \text{leafs}} 2^{c_{max}-c_i} \leq 2^{c_{max}} \quad \Rightarrow \quad \sum_{i \in \text{leafs}} 2^{-c_i} \leq 1 \tag{10.24}$$

which proves the case when the number of codewords is finite.

Suppose now that we have a countably infinite number of codewords. Let $b_1\ b_2\ ...\ b_{c_i}$ be the ith codeword and let

$$r_i = \sum_{j=i}^{c_i} b_j 2^{-j} \tag{10.25}$$

be the real number corresponding to the binary expansion of the codeword. We can associate the interval $[r_i, r_i + 2^{-c_i})$ with the ith codeword. This is the set of all real numbers whose binary expansion begins with $b_1\ b_2\ ...\ b_{c_i}$. Since this is a subinterval of $[0, 1]$, and all such subintervals corresponding to prefix codewords are disjoint, the sum of their lengths must be less than or equal to 1. This proves the case where the number of codewords is infinite.

Chapter 11

Complexity Regularization[1]

11.1 Review: PAC Bounds

Consider a finite collection of models $\mathcal{F}$, and recall the basic PAC bound: for any $\delta > 0$, with probability at least $1 - \delta$

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{log|\mathcal{F}| + log(1/\delta)}{2n}} \quad , \quad \forall f \in \mathcal{F} \tag{11.1}$$

where

$$\begin{aligned} \hat{R}_n(f) &= \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) \\ R(f) &= E[\ell(f(X), Y)] \end{aligned} \tag{11.2}$$

and the loss ℓ is assumed to be bounded between 0 and 1. Note that we can write the inequality above as:

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{log\left(\frac{|\mathcal{F}|}{\delta}\right)}{2n}} \tag{11.3}$$

Letting $\delta_f = \frac{\delta}{|\mathcal{F}|}$, we have:

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{log(1/\delta_f)}{2n}} \tag{11.4}$$

This is precisely the form of Hoeffding's inequality, with δ_f in place of the usual δ. In effect, in order to have Hoeffding's inequality hold with probability $1 - \delta$ for all $f \in \mathcal{F}$, we must distribute the "δ-budget" or "confidence-budget" over all $f \in \mathcal{F}$ (in this case, evenly distributed):

$$\begin{aligned} \sum_{f\in\mathcal{F}} \delta_f &= \sum_{f\in\mathcal{F}} \frac{\delta}{|\mathcal{F}|} \\ &= \delta \end{aligned} \tag{11.5}$$

However, to apply the union bound, we do not need to distribute δ evenly among the candidate models. We only require:

$$\sum_{f\in\mathcal{F}} \delta_f = \delta \tag{11.6}$$

So, if $p(f)$ are positive numbers satisfying $\sum_{f\in\mathcal{F}} p(f) = 1$, then we can take $\delta_f = p(f)\,\delta$. This provides two advantages:

[1] This content is available online at <http://cnx.org/content/m16266/1.2/>.

1. By choosing $p(f)$ larger for certain f, we can preferentially treat those candidates
2. We do not need $\mathcal{F}$ to be finite and we only require $\sum_{f\in\mathcal{F}} p(f) = 1$

Prefix codes are one way to achieve this. If we assign a binary prefix code of length $c(f)$ to each $f \in \mathcal{F}$, then the values $p(f) = 2^{-c(f)}$ satisfy $\sum_{f\in\mathcal{F}} p(f) \leq 1$ according to the Kraft inequality.

The main point of this lecture is to examine how PAC bounds of the form w.p. $\geq 1-\delta$

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{c(f)log2+log(1/\delta)}{2n}} \quad , \qquad \forall f \in \mathcal{F} \tag{11.7}$$

can be used to select a model that comes close to achieving the best possible performace

$$\inf_{f\in\mathcal{F}} R(f) \tag{11.8}$$

Let $\hat{f}_n$ be the model selected from $\mathcal{F}$ using the training data $\{X_i, Y_i\}_{i=1}^n$. We will specify this model in a moment, but keep in mind that it is not necessarily the model with minimum empirical risk as before. We would like to have

$$E\left[R\left(\hat{f}_n\right)\right] - \inf_{f\in\mathcal{F}} R(f) \tag{11.9}$$

as small as possible. First, for any $\delta > 0$, define

$$\hat{f}_n^{\delta} \quad = \quad \underset{f\in\mathcal{F}}{argmin}\{\hat{R}_n(f) + C(f,n,\delta)\} \tag{11.10}$$

where

$$C(f,n,\delta) \equiv \sqrt{\frac{c(f)log2+log(1/\delta)}{2n}} \tag{11.11}$$

Then w.p. $\geq 1-\delta$

$$R(f) \leq \hat{R}_n(f) + C(f,n,\delta) \quad , \qquad \forall f \in \mathcal{F} \tag{11.12}$$

and in particular,

$$R\left(\hat{f}_n^{\delta}\right) \leq \hat{R}_n\left(\hat{f}_n^{\delta}\right) + C\left(\hat{f}_n^{\delta}, n, \delta\right) \quad , \tag{11.13}$$

so, by the definition of $\hat{f}_n^{\delta}$, $\forall f \in \mathcal{F}$

$$R\left(\hat{f}_n^{\delta}\right) \leq \hat{R}_n(f) + C(f,n,\delta) \quad . \tag{11.14}$$

We will make use of the inequality above in a moment. First note that $\forall f \in \mathcal{F}$

$$E\left[R\left(\hat{f}_n^{\delta}\right)\right] - R(f) \quad = \quad E\left[R\left(\hat{f}_n^{\delta}\right) - \hat{R}_n(f)\right] + E\left[\hat{R}_n(f) - R(f)\right] \tag{11.15}$$

The second term is exactly 0, since $E\left[\hat{R}_n(f)\right] = R(f)$.

Now consider the first term $E\left[R\left(\hat{f}_n^{\delta}\right)-\hat{R}_n(f)\right]$. Let Ω be the set of events on which

$$R\left(\hat{f}_n^{\delta}\right)\leq \hat{R}_n(f)-C(f,n,\delta),\ \forall\ f\in\mathcal{F} \tag{11.16}$$

From the bound above, we know that $P(\Omega)\geq 1-\delta$. Thus,

$$E\left[R\left(\hat{f}_n^{\delta}\right)-\hat{R}_n(f)\right] = E\left[R\left(\hat{f}_n^{\delta}\right)-\hat{R}_n(f)\,|\Omega\right]P(\Omega) + \tag{11.17}$$

$$E\left[R\left(\hat{f}_n^{\delta}\right)-\hat{R}_n(f)\,|\Omega^c\right](1-P(\Omega)) \leq C(f,n,\delta) +$$

$$\delta \quad \left(\text{since } 0\leq R,\ \hat{R}\leq 1,\ P(\Omega)\leq 1 \text{ and } 1-P(\Omega)\leq\delta\right) = \sqrt{\frac{c(f)log2+log(1/\delta)}{2n}}+\delta =$$

$$\sqrt{\frac{c(f)log2+\frac{1}{2}logn}{2n}}+\frac{1}{\sqrt{n}} \quad \left(\text{by setting } \delta=\frac{1}{\sqrt{n}}\right)$$

We can summarize our analysis with the following theorem.

Theorem 11.1: Complexity Regularized Model Selection
Let $\mathcal{F}$ be a countable collection of models, and assign a positive number $c(f)$ to each $f\in\mathcal{F}$ such that $\sum_{f\in\mathcal{F}}2^{-c(f)}\leq 1$. Define the minimum complexity regularized risk model

$$\hat{f}_n = \underset{f\in\mathcal{F}}{argmin}\{\hat{R}_n(f)+\sqrt{\frac{c(f)log2+\frac{1}{2}logn}{2n}}\} \tag{11.18}$$

Then,

$$E\left[R\left(\hat{f}_n\right)\right] \leq \underset{f\in\mathcal{F}}{inf}\{R(f)+\sqrt{\frac{c(f)log2+\frac{1}{2}logn}{2n}}+\frac{1}{\sqrt{n}}\} \tag{11.19}$$

This shows that

$$\hat{R}_n(f)+\sqrt{\frac{c(f)log2+\frac{1}{2}logn}{2n}} \tag{11.20}$$

is a reasonable surrogate for

$$R(f)+\sqrt{\frac{c(f)log2+\frac{1}{2}logn}{2n}} \tag{11.21}$$

Example: Histogram Classifiers
Let $\mathcal{X}=[0,1]^d$ be the input space and $\mathcal{Y}=\{0,1\}$ be the output space. Let $\mathcal{F}_k$, k = 1, 2, ... denotes the collection of histogram classification rules with k equal volume bins. One choice of prefix code for this example is: $k=1\Rightarrow$ code $=0, k=3\Rightarrow$ code $=10, k=3\Rightarrow$ code $=110$ and so on .... Then, if first code is corresponding to k $\Rightarrow f\in\mathcal{F}_k$, followed by $k=log_2|\mathcal{F}_k|$ bits to indicate which of the 2^k histogram rules in $\mathcal{F}_k$ is under consideration, we have

$$f\in\mathcal{F}_k\Rightarrow c(f)=2k\ bits \tag{11.22}$$

Let $\hat{f}_n$ be the model that solves the minimization **i.e.,**

$$\min_{k\geq 1}\{\min_{f\in\mathcal{F}_k}\hat{R}_n(f)+\sqrt{\frac{2klog2+\frac{1}{2}logn}{2n}}\} \tag{11.23}$$

That is, for each k, let

$$\hat{f}_n^{(k)} = \arg\min_{f\in\mathcal{F}_k}\hat{R}_n(f) \tag{11.24}$$

Then select the best k according to

$$\hat{k} = \arg\min_{k\geq 1}\{\hat{R}_n\left(\hat{f}_n^{(k)}\right)+\sqrt{\frac{2klog2+\frac{1}{2}logn}{2n}}\} \tag{11.25}$$

and set

$$\hat{f}_n = \hat{f}_n^{\left(\hat{k}\right)} \tag{11.26}$$

Then,

$$E\left[R\left(\hat{f}_n\right)\right] \leq \inf_{k\geq 1}\{\min_{f\in\mathcal{F}_k}R(f)+\sqrt{\frac{2klog2+\frac{1}{2}logn}{2n}}+\frac{1}{\sqrt{n}}\} \tag{11.27}$$

It is a simple exercise to show that if $d=2$ and the Bayes decision boundary is a 1-d curve, then by setting $k=\sqrt{n}$ and selecting the best f from $\mathcal{F}_{\sqrt{n}}$ we have

$$E\left[R\left(\hat{f}_n\right)\right] = O\left(n^{-1/4}\right) \tag{11.28}$$

NOTE: The complexity regularized classifier $\hat{f}_n$ adaptively achieves this rate, without user intervention.

Chapter 12

Decision Trees[1]

12.1 Minimum Complexity Penalized Function

Recall the basic results of the last lectures: let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces respectively. Let $X \in \mathcal{X}$ and $Y \in \mathcal{X}$ be random variables with unknown joint probability distribution P_{XY}. We would like to use X to "predict" Y. Consider a loss function $0 \leq \ell(y_1, y_2) \leq 1,\ \forall y_1, y_2 \in \mathcal{Y}$. This function is used to measure the accuracy of our prediction. Let $\mathcal{F}$ be a collection of candidate functions (models), $f : \mathcal{X} \to \mathcal{Y}$. The expected risk we incur is given by $R(f) \equiv E_{XY}[\ell(f(X), Y)]$. We have access only to a number of i.i.d. samples, $\{X_i, Y_i\}_{i=1}^n$. These allow us to compute the empirical risk $\hat{R}_n(f) \equiv \frac{1}{n}\sum_{i=1}^n \ell(f(X_i), Y_i)$.

Assume in the following that $\mathcal{F}$ is countable. Assign a positive number $c(f)$ to each $f \in \mathcal{F}$ such that $\sum_{f\in\mathcal{F}} 2^{-c(f)} \leq 1$. If we use a prefix code to describe each element of $\mathcal{F}$ and define $c(f)$ to be the codeword length (in bits) for each $f \in \mathcal{F}$, the last inequality is automatically satisfied.

We define the **minimum complexity penalized estimator** as

$$\hat{f}_n \equiv \underset{f\in\mathcal{F}}{argmin}\{\hat{R}_n(f) + \sqrt{\frac{c(f)\,log2 + \frac{1}{2}logn}{2n}}\}. \tag{12.1}$$

As we showed previously we have the bound

$$E\left[R\left(\hat{f}_n\right)\right] \leq \underset{f\in\mathcal{F}}{min}\{R(f) + \sqrt{\frac{c(f)\,log2 + \frac{1}{2}logn}{2n}} + \frac{1}{\sqrt{n}}\}. \tag{12.2}$$

The performance (risk) of $\hat{f}_n$ is on average better than

$$R(f_n^*) + \sqrt{\frac{c(f_n^*)\,log2 + \frac{1}{2}logn}{2n}} + \frac{1}{\sqrt{n}}, \tag{12.3}$$

where

$$f_n^* = \underset{f\in\mathcal{F}}{argmin}\{R(f) + \sqrt{\frac{c(f)\,log2 + \frac{1}{2}logn}{2n}}\}. \tag{12.4}$$

If it happens that the optimal function, that is

$$f^* = arg \underset{f \text{ measurable}}{min} R(f), \tag{12.5}$$

[1] This content is available online at <http://cnx.org/content/m16287/1.2/>.

is close to an $f \in \mathcal{F}$ with a small $c(f)$, then $\hat{f}_n$ will perform almost as well as the optimal function.

Example 12.1
Suppose $f^* \in \mathcal{F}$, then

$$E\left[R\left(\hat{f}_n\right)\right] \leq R(f^*) + \sqrt{\frac{c(f^*)\,log2 + \frac{1}{2}logn}{2n}} + \frac{1}{\sqrt{n}}. \tag{12.6}$$

Furthermore if $c(f^*) = O(logn)$ then

$$E\left[R\left(\hat{f}_n\right)\right] \leq R(f^*) + O\left(\sqrt{\frac{logn}{n}}\right), \tag{12.7}$$

that is, only within a small $O\left(\sqrt{\frac{logn}{n}}\right)$ offset of the optimal risk.

In general, we can also bound the excess risk $E\left[R\left(\hat{f}_n\right)\right] - R^*$, where R^* is the Bayes risk,

$$R^* = \inf_{f \text{ measurable}} R(f). \tag{12.8}$$

By subtracting R^* (a constant) from both sides of the inequality

$$E\left[R\left(\hat{f}_n\right)\right] \leq \min_{f\in\mathcal{F}}\{R(f) + \sqrt{\frac{c(f)\,log2 + \frac{1}{2}logn}{2n}} + \frac{1}{\sqrt{n}}\} \tag{12.9}$$

we obtain

$$E\left[R\left(\hat{f}_n\right)\right] - R^* \leq \min_{f\in\mathcal{F}}\{R(f) - R^* + \sqrt{\frac{c(f)\,log2 + \frac{1}{2}logn}{2n}} + \frac{1}{\sqrt{n}}\}. \tag{12.10}$$

Note that two terms in this upper bound: $R(f) - R^*$ is a bound on the approximation error of a model f, and remainder is a bound on the estimation error associated with f. Thus, we see that complexity regularization automatically optimizes a balance between approximation and estimation errors. In other words, complexity regularization is **adaptive** to the unknown tradeoff between approximation and estimation.

12.2 Classification

Consider the particularization of the above to a classification scenario. Let $\mathcal{X} = [0,1]^d$, $\mathcal{Y} = \{0,1\}$ and $\ell\left(\hat{y}, y\right) \equiv \mathbf{1}_{\{\hat{y}\neq y\}}$. Then $R(f) = E_{XY}\left[\mathbf{1}_{\{f(X)\neq Y\}}\right] = P(f(X) \neq Y)$. The Bayes risk is given by

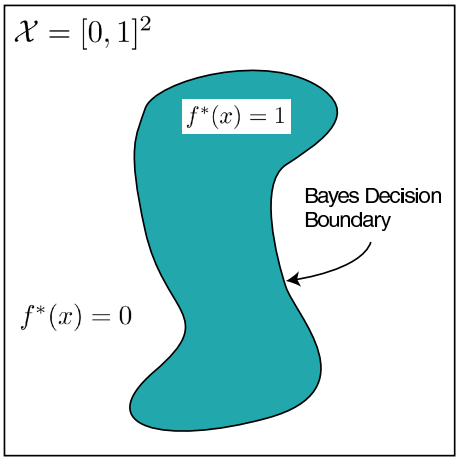
$$R^* = \inf_{f \text{ measurable}} R(f). \tag{12.11}$$

As it was observed before, the Bayes classifier (**i.e.,** a classifier that achieves the Bayes risk) is given by

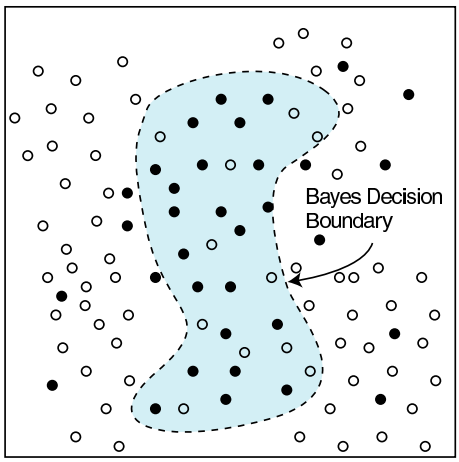
$$f^*(x) = \begin{cases} 1, & P(Y=1|X=x) \geq \frac{1}{2} \\ 0, & P(Y=1|X=x) < \frac{1}{2} \end{cases}. \tag{12.12}$$

This classifier can be expressed in a different way. Consider the set $G^* = \{x : P(Y=1|X=x) \geq 1/2\}$. The Bayes classifier can written as $f^*(x) = \mathbf{1}_{\{x\in G^*\}}$. Therefore the classifier is characterized entirely by the

set G^*, if $X \in G^*$ then the "best" guess is that Y is one, and vice-versa. The boundary of this set corresponds to the points where the decision is harder. The boundary of G^* is called the **Bayes Decision Boundary**. In Figure 12.1(a) this concept is illustrated. If $\eta(x) = P(Y = 1|X = x)$ is a continuous function then the Bayes decision boundary is simply given by $\{x : \ P(Y = 1|X = x) = 1/2\}$. Clearly the structure of the decision boundary provides important information on the difficulty of the problem.



(a)



(b)

Figure 12.1: (a) The Bayes classifier and the Bayes decision boundary ; (b) Example of the i.i.d. training pairs.

12.2.1 Empirical Classifier Design

Given n i.i.d. training pairs, $\{X_i, Y_i\}_{i=1}^n$, we want to construct a classifier $\hat{f}_n$ that performs well on average, **i.e.,** we want $E\left[R\left(\hat{f}_n\right)\right]$ as close to R^* as possible. In Figure 12.1(b) an example of the i.i.d. training pairs is depicted.

The construction of a classifier boils down to the estimation of the Bayes decision boundary. The histogram rule, discussed in a previous lecture, approaches the problem by subdividing the feature space into small boxes and taking a majority vote of the training data in each box. A typical result is depicted in Figure 12.2(a).

The main problem with the histogram rule is that it is solving a more complicated problem than it is actually necessary. We do not need to determine the correct label for each individual box directly (the histogram rule is essentially estimating $\eta(x)$). In principle we only need to locate the decision boundary and assign the correct label on either side (notice that the accuracy of a majority vote over a region increases with the size of the region). The next example illustrates this.

Example 12.2: Three Different Classifiers

The pictures below correspond to the approximation of the Bayes classifier by three different classifiers:

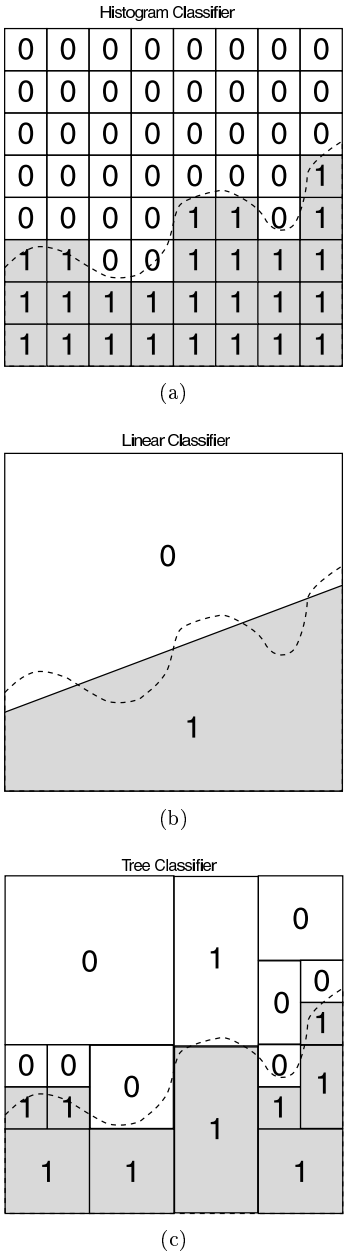


Figure 12.2: (a) Histogram classifier ; (b) Linear classifier; (c)Tree classifier.

The linear classifier and the tree classifier (to be defined formally later) both attack the problem of finding the boundary more directly than the histogram classifier, and therefore they tend to produce much better results in theory and practice. In the following we will demonstrate this for classification trees.

12.3 Binary Classification Trees

Binary classification trees are constructed by a two-step process:

1. Tree growing
2. Tree pruning

The basic idea is to first grow a very large, complicated tree classifier, that explains the the training data very accurately, but has poor generalization characteristics, and then prune this tree, to avoid overfitting.

12.3.1 Growing Trees

The growing process is based on recursively subudividing the feature space. Usually the subdivisions are splits of existing regions into two smaller regions (**i.e.,** binary splits) and usually the splits are perpendicular to one of the feature axes. An example of such construction is depicted in Figure 12.3.

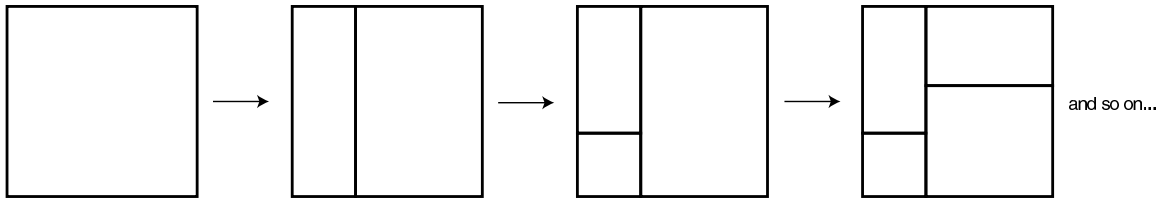


Figure 12.3: Growing a recursive binary tree ($\mathcal{X} = [0,1]^2$).

Often the splitting process is based on the training data, and is designed to separate data with different labels as much as possible. It such constructions, the "splits," and hence the tree-structure itself, are data dependent. Alternatively, the splitting and subdivision could be independent from the training data. The latter approach is the one we are going to investigate in detail, and we will consider Dyadic Decision Trees and Recursive Dyadic Partitions (depicted in Figure 12.4) in particular.

Until now we have been referring to trees, but did not make clear how do trees relate to partitions. It turns out that any decision tree can be associated with a partition of the input space $\mathcal{X}$ and vice-versa. In particular, a Recursive Dyadic Partition (RDP) can be associated with a (binary) tree. In fact, this is the most efficient way of describing a RDP. In Figure 12.4 we illustrate the procedure. Each leaf of the tree corresponds to a cell of the partition. The nodes in the tree correspond to the various partition cells that are generated through in the construction of the tree. The orientation of the dyadic split alternates between the levels of the tree (for the example of Figure 12.4, at the root level the split is done in the horizontal axis, at the level below that (the level of nodes 2 and 3) the split is done in the vertical axis, and so on...). The tree is called dyadic because the splits of cells are always at the midpoint along one coordinate axis, and consequently the sidelengths of all cells are dyadic (i.e., powers of 2).

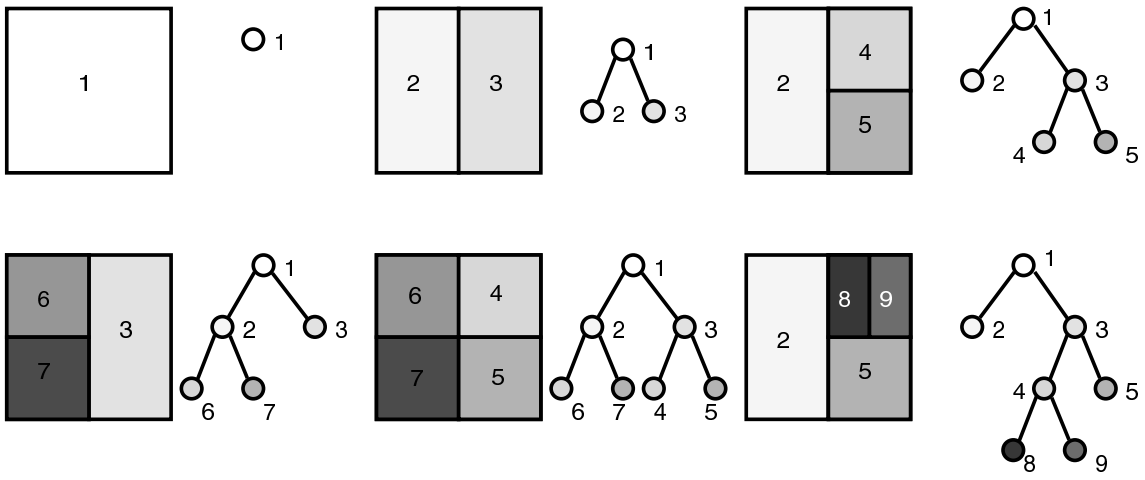


Figure 12.4: Example of Recursive Dyadic Partition (RDP) growing ($\mathcal{X} = [0,1]^2$).

In the following we are going to consider the 2-dimensional case, but all the results can be easily generalized for the d-dimensional case ($d \geq 2$), provided the dyadic tree construction is defined properly. Consider a recursive dyadic partition of the feature space into k boxes of equal size. Associated with this partition is a tree T. Minimizing the empirical risk with respect to this partition produces the histogram classifier with k equal-sized bins. Consider also all the possible partitions corresponding to pruned versions of the tree T. Minimizing the empirical risk with respect to those other partitions results in other classifiers (dyadic decision trees) that are fundamentally different than the histogram rule we analyzed earlier.

12.3.2 Pruning

Let $\mathcal{F}$ be the collection of all possible dyadic decision trees corresponding to recursive dyadic partitions of the feature space. Each such tree can be prefix encoded with a bit-string proportional to the number of leafs in the tree as follows; encode the structure of the tree in a top-down fashion: (i) assign a zero at each branch node and a one at each leaf node (terminal node) (ii) read the code in a breadth-first fashion, top-down, left-right. Figure 12.5 exemplifies this coding strategy. Notice that, since we are considering binary trees, the total number of nodes is twice the number of leafs minus one, that is, if the number of leafs in the tree is k then the number of nodes is $2k-1$. Therefore to encode a tree with k leafs we need $2k-1$ bits.

Since we want to use the partition associated with this tree for classification we need to assign a decision label (either zero or one) to each leaf. Hence, to encode a decision tree in this fashion we need $3k-1$ bits, where k is the number of leafs. For a tree with k leafs the first $2k-1$ bits of the codeword encode the tree structure, and the remaining k bits encode the classification labels. This is easily shown to be a prefix code, therefore we can use this under our classification scenario.

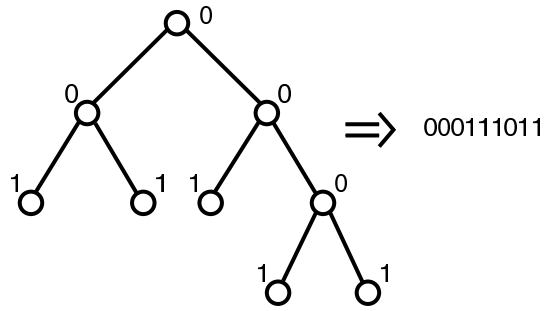


Figure 12.5: Illustration of the tree coding technique: example of a tree and corresponding prefix code.

Let

$$\hat{f}_n^* = \underset{f \in \mathcal{F}}{argmin}\{\hat{R}_n(f) + \sqrt{\frac{(3k-1)\,log2 + \frac{1}{2}logn}{2n}}\}. \tag{12.13}$$

This optimization can be solved through a bottom-up pruning process (starting from a very large initial tree T_0) in $O\left(|T_0|^2\right)$ operations, where $|T_0|$ is the number of leafs in the initial tree. The complexity regularization theorem tells us that

$$E\left[R\left(\hat{f}_n\right)\right] \leq \underset{f \in \mathcal{F}}{min}\{R(f) + \sqrt{\frac{(3k-1)\,log2 + \frac{1}{2}logn}{2n}}\} + \frac{1}{\sqrt{n}}. \tag{12.14}$$

12.4 Comparison between Histogram Classifiers and Classification Trees

In the following we will illustrate the idea behind complexity regularization by applying the basic theorem to histogram classifiers and classification trees (using our setup above).

Consider the classification setup described in "Classification" (Section 12.2: Classification), with $\mathcal{X} = [0,1]^2$.

12.4.1 Histogram Risk Bound

Recall the setup and results of a previous lecture[2]. Let

$$\mathcal{F}_k^H = \{\text{histogram rules with } k^2 \text{ bins}\}. \tag{12.15}$$

Then $|\mathcal{F}_k^H| = 2^{k^2}$. Let $\mathcal{F}^H = \bigcup_{k \geq 1} \mathcal{F}_k^H$. We can encode each element f of $\mathcal{F}^H$ with $c_H(f) = k + k^2$ bits, where the first k bits indicate the smallest k such that $f \in \mathcal{F}_k^H$ and the following k^2 bits encode the labels of each bin. This is a prefix encoding of all the elements in $\mathcal{F}^H$.

[2]The description here is slightly different than the one in the previous lecture.

We define our estimator as

$$\hat{f}_n^H = \hat{f}_n^{\hat{k}}, \tag{12.16}$$

where

$$\hat{f}_n^{(k)} = arg\min_{f \in \mathcal{F}_k^H} \hat{R}_n(f), \tag{12.17}$$

and

$$\hat{k} = arg\min_{k \geq 1} \{\hat{R}_n\left(\hat{f}_n^{(k)} + \sqrt{\frac{(k+k^2)\,log2 + \frac{1}{2}logn}{2n}}\right\}. \tag{12.18}$$

Therefore $\hat{f}_n^H$ minimizes

$$\hat{R}_n(f) + \sqrt{\frac{c_H(f)\,log2 + \frac{1}{2}logn}{2n}}, \tag{12.19}$$

over all $f \in \mathcal{F}^H$. We showed before that

$$E\left[R\left(\hat{f}_n^H\right)\right] - R^* \leq \min_{f \in \mathcal{F}^H} \{R(f) - R^* + \sqrt{\frac{c_H(f)\,log2 + \frac{1}{2}logn}{2n}}\} + \frac{1}{\sqrt{n}}. \tag{12.20}$$

To proceed with our analysis we need to make some assumptions on the intrinsic difficulty of the problem. We will assume that the Bayes decision boundary is a "well-behaved" 1-dimensional set, in the sense that it has box-counting dimension one (see Appendix "Box Counting Dimension" (Section 12.6: Box Counting Dimension)). This implies that, for an histogram with k^2 bins, the Bayes decision boundary intersects less than Ck bins, where C is a constant that does not depend on k. Furthermore we assume that the marginal distribution of X satisfies $P_X(A) \leq K|A|$, for any measurable subset $A \subseteq [0,1]^2$. This means that the samples collected do not accumulate anywhere in the unit square.

Under the above assumptions we can conclude that

$$\min_{f \in \mathcal{F}_k^H} R(f) - R^* \;\leq\; \frac{K}{k^2}Ck = \frac{CK}{k}. \tag{12.21}$$

Therefore

$$E\left[R\left(\hat{f}_n^H\right)\right] - R^* \;\leq\; CK/k + \sqrt{\frac{(k+k^2)\,log2 + \frac{1}{2}logn}{2n}} + \frac{1}{\sqrt{n}}. \tag{12.22}$$

We can balance the terms in the right side of the above expression using $k = n^{1/4}$ (for n large) therefore

$$E\left[R\left(\hat{f}_n^H\right)\right] - R^* \;=\; O\left(n^{-1/4}\right), \quad \text{as } n \to \infty. \tag{12.23}$$

12.4.2 Dyadic Decision Trees

Now let's consider the dyadic decision trees, under the assumptions above, and contrast these with the histogram classifier. Let

$$\mathcal{F}_k^T = \{\text{tree classifiers with } k \text{ leafs}\}. \tag{12.24}$$

Let $\mathcal{F}^T = \bigcup_{k\geq 1}\mathcal{F}_k^T$. We can prefix encode each element f of $\mathcal{F}^T$ with $c_T(f) = 3k-1$ bits, as described before.

Let

$$\hat{f}_n^T = \hat{f}_n^{\left(\hat{k}\right)}, \tag{12.25}$$

where

$$\hat{f}_n^{(k)} = arg\min_{f\in\mathcal{F}_k^T} \hat{R}_n(f), \tag{12.26}$$

and

$$\hat{k} = argmin_{k\geq 1}\{\hat{R}_n\left(\hat{f}_n^{(k)} + \sqrt{\frac{(3k-1)\,log2 + \frac{1}{2}logn}{2n}}\right\}. \tag{12.27}$$

Hence $\hat{f}_n^T$ minimizes

$$\hat{R}_n(f) + \sqrt{\frac{c_T(f)\,log2 + \frac{1}{2}logn}{2n}}, \tag{12.28}$$

over all $f \in \mathcal{F}^T$. Moreover

$$E\left[R\left(\hat{f}_n^T\right)\right] - R^* \;\leq\; \min_{f\in\mathcal{F}^T}\{R(f) - R^* + \sqrt{\frac{c_T(f)\,log2 + \frac{1}{2}logn}{2n}}\} + \frac{1}{\sqrt{n}}. \tag{12.29}$$

If the Bayes decision boundary is a 1-dimensional set, as in "Histogram Risk Bound" (Section 12.4.1: Histogram Risk Bound), there exists a tree with at most $8Ck$ leafs such that the boundary is contained in at most Ck squares, each of volume $1/k^2$. To see this, start with a tree yielding the histogram partition with k^2 boxes (**i.e.,** the tree partitioning the unit square into k^2 equal sized squares). Now prune all the nodes that do not intersect the boundary. In Figure 12.6 we illustrate the procedure. If you carefully bound the number of leafs you need at each level you can show that you will have in total less than $8Ck$ leafs. We conclude then that there exists a tree with at most $8Ck$ leafs that has the same risk as a histogram with $O\left(k^2\right)$ bins. Therefore, using (12.14) we have

$$E\left[R\left(\hat{f}_n^T\right)\right] - R^* \;\leq\; CK/k + \sqrt{\frac{(3\,(8Ck)-1)\,log2 + \frac{1}{2}logn}{2n}} + \frac{1}{\sqrt{n}}. \tag{12.30}$$

We can balance the terms in the right side of the above expression using $k = n^{1/3}$ (for n large) therefore

$$E\left[R\left(\hat{f}_n^T\right)\right] - R^* \;=\; O\left(n^{-1/3}\right), \quad \text{as } n\to\infty. \tag{12.31}$$

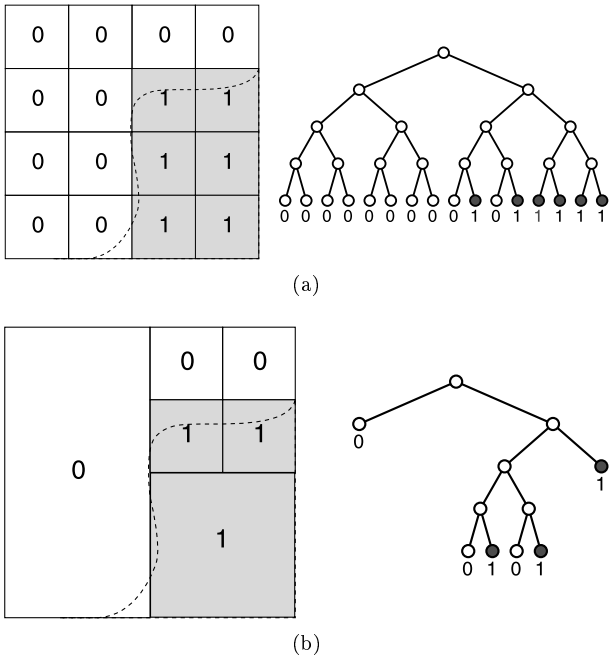


Figure 12.6: Illustration of the tree pruning procedure: (a) Histogram classification rule, for a partition with 16 bins, and corresponding binary tree representation (with 16 leafs). (b) Pruned version of the histogram tree, yielding exactly the same classification rule, but now requiring only 6 leafs. (**Note:** The trees where constructed using the procedure of Figure)

12.5 Final Comments

Trees generally work much better than histogram classifiers. This is essentially because they provide much more efficient ways of approximating the Bayes decision boundary (as we saw in our example, under reasonable assumptions on the Bayes boundary, a tree encoded with $O(k)$ bits can describe the same classifier as an histogram that requires $O(k^2)$ bits).

The dyadic decision trees studied here are different than classical tree rules, such as CART or C4.5. Those techniques select a tree according to

$$\hat{k} = \underset{k \geq 1}{argmin} \{ \hat{R}_n \left(\hat{f}_n^{(k)} \right) + \alpha k \}, \tag{12.32}$$

for some $\alpha > 0$ whereas ours was roughly

$$\hat{k} = \underset{k \geq 1}{argmin} \{ \hat{R}_n \left(\hat{f}_n^{(k)} \right) + \alpha \sqrt{k} \}, \tag{12.33}$$

for $\alpha \approx \sqrt{\frac{3log2}{2n}}$. The square root penalty is essential for the risk bound. No such bound exists for CART or C4.5 . Moreover, recent experimental work has shown that the square root penalty often performs better

in practice. Finally, recent results show that a slightly tighter bounding procedure for the estimation error can be used to show that dyadic decision trees (with a slightly different pruning procedure) achieve a rate of

$$E\left[R\left(\hat{f}_n^T\right)\right] - R^* = O\left(n^{-1/2}\right), \quad \text{as } n \to \infty, \tag{12.34}$$

which turns out to be the minimax optimal rate (i.e., under the boundary assumptions above, no method can achieve a faster rate of convergence to the Bayes error).

12.6 Box Counting Dimension

The notion of dimension of a sets arises in many aspects of mathematics, and it is particularly relevant to the study of fractals (that besides some important applications make really cool t-shirts). The dimension somehow indicates how we should measure the contents of a set (length, area, volume, etc...). The box-counting dimension is a simple definition of the dimension of a set. The main idea is to cover the set with boxes with sidelength r. Let $N(r)$ denote the smallest number of such boxes, then the box counting dimension is defined as

$$\lim_{r \to 0} \frac{logN(r)}{-logr}. \tag{12.35}$$

Although the boxes considered above do not need to be aligned on a rectangular grid (and can in fact overlap) we can usually consider them over a grid and obtain an upper bound on the box-counting dimension. To illustrate the main ideas let's consider a simple example, and connect it to the classification scenario considered before.

Let $f : [0,1] \to [0,1]$ be a Lipschitz function, with Lipschitz constant L (**i.e.,** $|f(a) - f(b)| \leq L|a - b|, \ \forall a, b \in [0,1]$). Define the set

$$A = \{x = (x_1, x_2) : x_2 = f(x_1)\}, \tag{12.36}$$

that is, the set A is the graphic of function f.

Consider a partition with k^2 squared boxes (just like the ones we used in the histograms), the points in set A intersect at most $C^{'}k$ boxes, with $C^{'} = (1 + \lceil L \rceil)$ (and also the number of intersected boxes is greater than k). The sidelength of the boxes is $1/k$ therefore the box-counting dimension of A satisfies

$$\begin{aligned} dim_B(A) &\leq \lim_{1/k \to 0} \frac{logC^{'}k}{-log(1/k)} \\ &= \lim_{k \to \infty} \frac{logC^{'} + log(k)}{log(k)} \\ &= 1. \end{aligned} \tag{12.37}$$

The result above will hold for any "normal" set $A \subseteq [0,1]^2$ that does not occupy any area. For most sets the box-counting dimension is always going to be an integer, but for some "weird" sets (called fractal sets) it is not an integer. For example, the Koch curve has box-counting dimension $log(4)/log(3) = 1.26186....$ This means that it is not quite as small as a 1-dimensional curve, but not as big as a 2-dimensional set (hence occupies no area).

To connect these concepts to our classification scenario consider a simple example. Let $\eta(x) = P(Y = 1|X = x)$ and assume $\eta(x)$ has the form

$$\eta(x) = \frac{1}{2} + x_2 - f(x_1), \quad \forall x \equiv (x_1, x_2) \in \mathcal{X}, \tag{12.38}$$

where $f : [0,1] \to [0,1]$ is Lipschitz with Lipschitz constant L. The Bayes classifier is then given by

$$f^*(x) = \mathbf{1}_{\{\eta(x) \geq 1/2\}} \equiv \mathbf{1}_{\{x_2 \geq f(x_1)\}}. \tag{12.39}$$

This is depicted in Figure 12.7. Note that this is a special, restricted class of problems. That is, we are considering the subset of all classification problems such that the joint distribution P_{XY} satisfies $P(Y = 1|X = x) = 1/2 + x_2 - f(x_1)$ for some function f that is Lipschitz. The Bayes decision boundary is therefore given by

$$A = \{x = (x_1, x_2) : x_2 = f(x_1)\}. \tag{12.40}$$

Has we observed before this set has box-counting dimension 1.

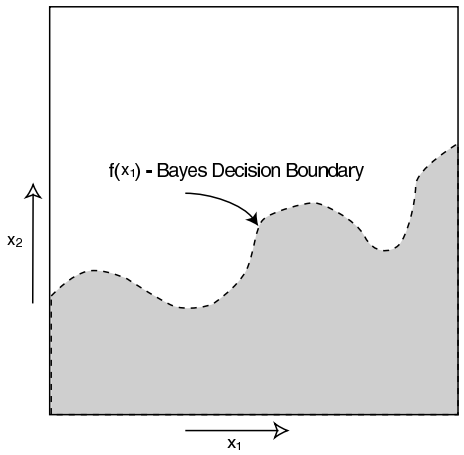


Figure 12.7: Bayes decision boundary for the setup described in Appendix .

Chapter 13

Complexity Regularization for Squared Error Loss[1]

13.1 Complexity Regularization in Regression

Recall the classification problem. In Lecture 6 (Chapter 7), where we assumed that $min_{f\in\mathcal{F}}R(f)=0$, we obtained the PAC bound $\forall f\in\mathcal{F}$

$$\mathcal{P}\{R\left(\hat{f}_n\right)>\varepsilon\}\leq|\mathcal{F}|e^{-n\varepsilon}. \tag{13.1}$$

From Corrolary 1 in Lecture 6 (Corollary 7.1, p. 47),

$$E\left[R\left(\hat{f}_n\right)\right]\leq\frac{1+log|\mathcal{F}|}{n}. \tag{13.2}$$

In Lectures 7 (Chapter 8) and 8 (Chapter 9), we dropped the assumption that $min_{f\in\mathcal{F}}R(f)=0$ and obtained, $\forall f\in\mathcal{F}$

$$\mathcal{P}\{R\left(\hat{f}_n\right)>\varepsilon\}\leq|\mathcal{F}|e^{-2n\varepsilon^2}. \tag{13.3}$$

This led to

$$E\left[R\left(\hat{f}_n\right)-\min_{f\in\mathcal{F}}R(f)\right]\leq\sqrt{\frac{log|\mathcal{F}|+logn+2}{n}}. \tag{13.4}$$

Hoeffding's inequality was central to our analysis of learning under bounded loss functions. In many regression and signal estimation problems it is natural to consider squared error loss functions (rather than 0/1 or absolute error). In such cases, we will need to derive bounds using different techniques.

Example 13.1
To illustrate the distinction between classification and regression, consider a simple, scalar signal plus noise problem. Consider $Y_i=\theta+W_i$, $i=1,\cdots,n$, where θ is a fixed unknown scalar parameter and the W_i are independent, zero-mean, unit variance random variables. Let $\overline{Y}=1/n\sum_{i=1}^{n}Y_i$.

[1] This content is available online at <http://cnx.org/content/m16267/1.2/>.

Then, according to the Central Limit Theorem, $\overline{Y}$ is distributed approximately $N(\theta, 1/n)$. A simple tail-bound on the Gaussian distribution gives us

$$P\left(\overline{Y} - \theta > \varepsilon\right) = P\left(W > \varepsilon\right) \leq \frac{1}{2} e^{-n\varepsilon^2/2}, \tag{13.5}$$

which implies that

$$P\left(\left|\overline{Y} - \theta\right|^2 > \varepsilon\right) \leq e^{-n\varepsilon^2/2}. \tag{13.6}$$

This is a bound on the deviations of the squared error $\mathrm{err}^2 = \left|\overline{Y} - \theta\right|^2$. Notice that the exponential decay rate is a function of ε rather than ε^2, as in Hoeffding's inequality. The squared error concentration inequality implies that $E\left[\left|\overline{Y} - \theta\right|^2\right] = O\left(\frac{1}{n}\right)$ (just write $E\left[\mathrm{err}^2\right] = \int_0^\infty P\left(\mathrm{err}^2 > t\right) dt$). Therefore, in regression with a squared error loss, we can hope to get a rate of convergence as fast as n^{-1} instead of $n^{-1/2}$. The reason is simply because we are using an squared error loss instead of the 0/1 or absolute error loss.

To begin our investigation into regression and function estimation, let us consider the following. Let $\mathcal{X} = \mathbf{R}^d$ and $\mathcal{Y} = \mathbf{R}$. Take $\mathcal{F}$such that $f \in \mathcal{F}$ is a map $f : \mathbf{R}^d \mapsto \mathbf{R}$. We have training data $\{X_i, Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} P_{XY}$. As our loss function, we take the squared error, **i.e.,**

$$l\left(f\left(X_i\right), Y_i\right) = \left(f\left(X_i\right) - Y_i\right)^2. \tag{13.7}$$

The risk is then the MSE:

$$R\left(f\right) = E\left[\left(f\left(X\right) - Y\right)^2\right]. \tag{13.8}$$

We know that the function f^* that minimizes the MSE is just the conditional expectation of Y given X:

$$f^* = E\left[Y | X = x\right]. \tag{13.9}$$

Now let $R^* = R\left(f^*\right)$. We would like to select an $\hat{f}_n \in \mathcal{F}$ using the training data $\{X_i, Y_i\}_{i=1}^n$ such that the **excess risk**

$$E\left[R\left(\hat{f}_n\right)\right] - R^* \geq 0 \tag{13.10}$$

is small. Let's consider the difference between the empirical risks:

$$\hat{R}\left(f\right) - \hat{R}\left(f^*\right) = \frac{1}{n}\sum_{i=1}^n \left(f\left(X_i\right) - Y_i\right)^2 - \frac{1}{n}\sum_{i=1}^n \left(f^*\left(X_i\right) - Y_i\right)^2. \tag{13.11}$$

Note that $E\left[\hat{R}\left(f\right) - \hat{R}\left(f^*\right)\right] = R\left(f\right) - R\left(f^*\right)$. Hence, by the Strong Law of Large Numbers (SLLN), we know that

$$\hat{R}\left(f\right) - \hat{R}\left(f^*\right) \to R\left(f\right) - R\left(f^*\right) \tag{13.12}$$

as $n \to \infty$. But how fast is this convergence?

We will derive a PAC style bound for the difference $\hat{R}(f) - \hat{R}(f^*) - (R(f) - R(f^*))$. The following derivation is from Barron 1991. The excess risk and it empirical counterpart will be denoted by

$$\begin{array}{l} r(f, f^*) = R(f) - R(f^*) \\ \hat{r}(f, f^*) = \hat{R}(f) - \hat{R}(f^*) \end{array}. \tag{13.13}$$

Note that $\hat{r}(f, f^*)$ is the sum of independent random variables:

$$\hat{r}(f, f^*) = -\frac{1}{n}\sum_{i=1}^{n} U_i, \tag{13.14}$$

where $U_i = -(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2$. Therefore, $r(f, f^*) - \hat{r}(f, f^*) = \frac{1}{n}\sum_{i=1}^{n}(U_i - E[U_i])$.

We are looking for a PAC bound of the form

$$\mathcal{P}\left(r(f, f^*) - \hat{r}(f, f^*) > \varepsilon\right) < \delta. \tag{13.15}$$

If the variables U_i are bounded, then we can apply Hoeffding's inequality. However, a more useful bound for our regression problem can be derived if the the variables U_i satisfy the following moment condition:

$$E\left[|U_i - E[U_i]|^k\right] \leq \frac{var(U_i)}{2}\ k!\ h^{k-2} \tag{13.16}$$

for some $h > 0$.

The moment condition can be difficult to verify in general, but it does hold, for example, for bounded random variables. If (13.16) holds, then the Craig-Bernstein (CB) inequality states:

$$\mathcal{P}\left(\frac{1}{n}\sum_{i=1}^{n}(U_i - E[U_i]) \geq \frac{t}{n\varepsilon} + \frac{n\varepsilon\ var\left(\frac{1}{n}\sum U_i\right)}{2(1-c)}\right) \leq e^{-t}, \tag{13.17}$$

for $0 < \varepsilon h \leq c < 1$ and $t > 0$. This shows that the tail decays exponentially in t, rather than exponentially in t^2. Recall Hoeffding's inequality:

$$\mathcal{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i]) \geq \frac{t}{n}\right) \leq e^{\frac{-2t^2}{n}}. \tag{13.18}$$

If $\frac{t}{n} \ll 1$, then $\frac{t^2}{n} \ll t$, which implies $e^{\frac{-2t^2}{n}} \gg e^{-t}$. This indicates that the CB inequality may be much tighter than Hoeffding's. To use the CB inequality, we need to bound the variance of $\frac{1}{n}\sum_{i=1}^{n} U_i$. Note that

$$var(U_i) = var\left(-(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2\right). \tag{13.19}$$

Assumption 1

The support of Y and the range f(X) is in a known interval of length b.

Proposition 1

With the above assumption, (13.16) holds with $h = \frac{2b^2}{3}$.

Proposition 2
Again, with the above assumption, it may be shown that

$$var\left(U_i\right) \leq 5b^2 r\left(f, f^*\right). \tag{13.20}$$

You can write U_i as

$$\begin{array}{l} U_i = 2Y_i f\left(X_i\right) - 2Y_i f^*\left(X_i\right) + f^*(X_i)^2 - f(X_i)^2 = 2Y_i f\left(X_i\right) - 2Y_i f^*\left(X_i\right) + \\ 2f^*(X_i)^2 - f^*(X_i)^2 - f(X_i)^2 + 2f\left(X_i\right) f^*\left(X_i\right) - 2f\left(X_i\right) f^*\left(X_i\right) = \\ 2\left(Y_i - f^*\left(X_i\right)\right)\left(f\left(X_i\right) - f^*\left(X_i\right)\right) - \left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2. \end{array} \tag{13.21}$$

Note that the variance of U_i is upper-bounded by its second moment. Also note that the covariance of the two terms above is zero:

$$\begin{array}{rcl} E\left[2\left(Y_i - f^*\left(X_i\right)\right)\left(f\left(X_i\right) - f^*\left(X_i\right)\right)\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2\right] & = & E\left[T_1 T_2\right] \\ & = & E_X\left[E_{Y|X}\left[T_1 T_2\right]\right] \\ & = & E_X\left[T_2 E_{Y|X}\left[T_1\right]\right] \\ & = & E_X\left[T_2 * 0\right] \\ & = & 0 \end{array} \tag{13.22}$$

This is evident when you recall that $f^*\left(X_i\right) = E\left[Y|X = X_i\right]$. Now we can bound the second moments of T_1 and T_2 :

$$\begin{array}{rcl} E\left[T_1\right] & = & 4E\left[\left(\left(Y_i - f^*\left(X_i\right)\right)\left(f\left(X_i\right) - f^*\left(X_i\right)\right)\right)^2\right] \\ & = & 4E\left[\left(Y_i - f^*\left(X_i\right)\right)^2\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2\right] \\ & \leq & 4E\left[b^2\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2\right] \\ E\left[T_2\right] & = & E\left[\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^4\right] \\ & = & E\left[\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2\right] \\ & \leq & E\left[b^2\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2\right] \end{array}. \tag{13.23}$$

So $var\left(U_i\right) \leq 5b^2 E\left[\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2\right]$. The final step is to see that

$$r\left(f, f^*\right) = E\left[U_i\right] = E_X\left[E_{Y|X}\left[U_i\right]\right] = E\left[\left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2\right]. \tag{13.24}$$

Thus, $n\ var\left(\frac{1}{n}\sum_{i=1}^n U_i\right) \leq 5b^2 r\left(f, f^*\right)$. And therefore, we can say that, with probability at least $1 - e^{-t}$,

$$r\left(f, f^*\right) - \hat{r}\left(f, f^*\right) \leq \frac{t}{n\ \varepsilon} + \frac{5\varepsilon\ b^2\ r\left(f, f^*\right)}{2\left(1 - c\right)}. \tag{13.25}$$

In other words, with probability at least $1 - \delta$ (where $\delta = e^{-t}$),

$$r\left(f, f^*\right) - \hat{r}\left(f, f^*\right) \leq \frac{log\frac{1}{\delta}}{n\ \varepsilon} + \frac{5\varepsilon\ b^2\ r\left(f, f^*\right)}{2\left(1 - c\right)}. \tag{13.26}$$

Now, suppose we have assigned positive numbers $c\left(f\right)$ to each $f \in \mathcal{F}$ satisfying the Kraft inequality:

$$\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1. \tag{13.27}$$

Note that (13.26) holds $\forall\delta > 0$. In particular, we let δ be a function of f:

$$\delta(f) = 2^{-c(f)}\delta. \tag{13.28}$$

So we can use this δ along with the procedure introduced in Lecture 9 (Chapter 10) (**i.e.,** Union of events bound followed by the Kraft inequality) to obtain the following. For all $f \in \mathcal{F}, \forall\delta > 0$,

$$r(f, f^*) - \hat{r}(f, f^*) \leq \frac{c(f)\,log2 + log\frac{1}{\delta}}{n\ \varepsilon} + \frac{5\varepsilon\ b^2\ r(f, f^*)}{2(1-c)} \tag{13.29}$$

with probability at least $1-\delta$. Now set $c = \varepsilon\ h = \frac{2b^2\ \varepsilon}{3}$ and assume $\varepsilon < \frac{6}{19b^2}$. Then define

$$\alpha = \frac{5\varepsilon\ b^2}{2(1-c)} < 1. \tag{13.30}$$

Now, after using α and rearranging terms, we have:

$$(1-\alpha)\, r(f, f^*) \leq \hat{r}(f, f^*) + \frac{c(f)\,log2 + log\frac{1}{\delta}}{\varepsilon\ n}. \tag{13.31}$$

We want to choose f to minmize this upper bound. So take

$$\hat{f}_n = \underset{f\in\mathcal{F}}{argmin}\{\hat{R}_n(f) + \frac{c(f)\,log2}{n\varepsilon}\}. \tag{13.32}$$

So, with probability at least $1-\delta$,

$$\begin{aligned}(1-\alpha)\, r\left(\hat{f}_n, f^*\right) &\leq \hat{r}\left(\hat{f}_n, f^*\right) + \frac{c\left(\hat{f}_n\right) log2 + log\frac{1}{\delta}}{\varepsilon\ n} \\ &\leq \hat{r}(f_n^*, f^*) + \frac{c(f_n^*)log2 + log\frac{1}{\delta}}{\varepsilon\ n}\end{aligned} \tag{13.33}$$

where $f_n^* = argmin_{f\in\mathcal{F}}\{R(f) + \frac{c(f)log2}{n\varepsilon}\}$.

Now we use the Craig-Bernstein inequality to bound the difference between $\hat{r}(f_n^*, f^*)$ and $r(f_n^*, f^*)$: With probability at least $1-\delta$,

$$\hat{r}(f_n^*, f^*) \leq r(f_n^*, f^*) + \alpha\ r(f_n^*, f^*) + \frac{log\left(\frac{1}{\delta}\right)}{n\varepsilon}. \tag{13.34}$$

Now we can again use the union bound to combine (13.33) and (13.34): With probability at least $1-2\delta, \forall\delta > 0$,

$$r\left(\hat{f}_n, f^*\right) \leq \frac{1+\alpha}{1-\alpha} r(f_n^*, f^*) + \frac{c(f_n^*)\,log2 + 2log1/\delta}{n\varepsilon}. \tag{13.35}$$

Now set $\delta = e^{\frac{-n\varepsilon t}{2}}$, then we have

$$\mathcal{P}\left(r\left(\hat{f}_n, f^*\right) - \frac{1+\alpha}{1-\alpha} r(f_n^*, f^*) + \frac{c(f_n^*)\,log2}{n\varepsilon} \geq t\right) \leq 2e^{\frac{-n\varepsilon t}{2}}. \tag{13.36}$$

Integrating, we get

$$\begin{aligned}E\left[r\left(\hat{f}_n, f^*\right) - \tfrac{1+\alpha}{1-\alpha} r(f_n^*, f^*) + \tfrac{c(f_n^*)log2}{n\varepsilon}\right] &\leq \int_0^\infty \mathcal{P}(\ " \geq t)\ dt \\ &\leq \int_0^\infty 2e^{\frac{-n\varepsilon t}{2}} \\ &= \tfrac{4}{n\varepsilon}\end{aligned}. \tag{13.37}$$

To sum up, we have shown that for $\varepsilon < \frac{6}{19b^2}$,

$$E\left[r\left(\hat{f}_n, f^*\right)\right] \leq \left(\frac{1+\alpha}{1-\alpha}\right) r\left(f_n^*, f^*\right) + \frac{c\left(f_n^*\right) log2 + 4}{n\varepsilon}, \tag{13.38}$$

or,

$$E\left[r\left(\hat{f}_n, f^*\right)\right] \leq \left(\frac{1+\alpha}{1-\alpha}\right) \min_{f \in \mathcal{F}} \{r\left(f, f^*\right) + \frac{c\left(f\right) log2}{n\varepsilon}\} + \frac{4}{n\varepsilon}, \tag{13.39}$$

since $\alpha < 1$. Or, in expanded form:

$$E\left[R\left(\hat{f}_n\right)\right] - R\left(f^*\right) \quad \leq \quad \left(\frac{1+\alpha}{1-\alpha}\right) \min_{f \in \mathcal{F}} \{R\left(f\right) - R\left(f^*\right) + \frac{c(f) log2}{n\varepsilon}\} + \frac{4}{n\varepsilon} \ . \tag{13.40}$$

Notice that if $f^* \in \mathcal{F}$ and if $c\left(f^*\right)$ is not too large (e.g., $c\left(f^*\right) \approx logn$), then we have $E\left[R\left(\hat{f}_n\right)\right] - R\left(f^*\right) = O\left(n^{-1} logn\right)$, within a logarithmic factor of the parametric rate of convergence!

Chapter 14

Maximum Likelihood Estimation[1]

In the last lecture (Chapter 13) we derived a risk (MSE) bound for regression problems; i.e., select an $f \in \mathcal{F}$ so that $E\left[(f(X)-Y)^2\right] - E\left[(f^*(X)-Y)^2\right]$ is small, where $f^*(x) = E[Y|X=x]$. The result is summarized below.

Theorem 14.1: Complexity Regularization with Squared Error Loss
Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = [-b/2, b/2]$, $\{X_i, Y_i\}_{i=1}^n$ iid, P_{XY} unknown, $\mathcal{F}$ = {collection of candidate functions},

$$f : \mathbb{R}^d \to \mathcal{Y}, \quad R(f) = E\left[(f(X)-Y)^2\right]. \tag{14.1}$$

Let $c(f)$, $f \in \mathcal{F}$, be positive numbers satisfying $\sum_{f\in\mathcal{F}} 2^{-c(f)} \leq 1$, and select a function from $\mathcal{F}$ according to

$$\hat{f}_n = argmin\{\hat{R}_n(f) + \frac{1}{\varepsilon}\frac{c(f)\,log2}{n}\}, \tag{14.2}$$

with $\varepsilon \leq \frac{3}{5b^2}$ and $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n (f(X_i)-Y_i)^2$. Then,

$$E\left[R\left(\hat{f}_n\right)\right] - R(f^*) \leq \left(\frac{1+\alpha}{1-\alpha}\right) \min_{f\in\mathcal{F}}\{R(f) - R(f^*) + \frac{1}{\varepsilon}\frac{c(f)\,log2}{n}\} + O\left(n^{-1}\right) \tag{14.3}$$

where $\alpha = \frac{\varepsilon b^2}{1-2b^2\varepsilon/3}$.

14.1 Maximum Likelihood Estimation

The focus of this lecture is to consider another approach to learning based on maximum likelihood estimation. Consider the classical signal plus noise model:

$$Y_i = f\left(\frac{i}{n}\right) + W_i, i = 1, \cdots, n \tag{14.4}$$

where W_i are iid zero-mean noises. Furthermore, assume that $W_i \sim P(w)$ for some known density $P(w)$. Then

$$Y_i \sim P\left(y - f\left(\frac{i}{n}\right)\right) \equiv P_{f_i}(y) \tag{14.5}$$

[1] This content is available online at <http://cnx.org/content/m16276/1.2/>.

since $Y_i - f\left(\frac{i}{n}\right) = W_i$.

A very common and useful loss function to consider is

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\left(-logP_{f_i}(Y_i)\right). \tag{14.6}$$

Minimizing $\hat{R}_n$ with respect to f is equivalent to maximizing

$$\frac{1}{n}\sum_{i=1}^{n} logP_{f_i}(Y_i) \tag{14.7}$$

or

$$\prod_{i=1}^{n} P_{f_i}(Y_i). \tag{14.8}$$

Thus, using the negative log-likelihood as a loss function leads to maximum likelihood estimation. If the W_i are iid zero-mean Gaussian r.v.s then this is just the squared error loss we considered last time. If the W_i are Laplacian distributed e.g. $P(w) \propto e^{-|w|}$, then we obtain the absolute error, or L_1, loss function. We can also handle non-additive models such as the Poisson model

$$Y_i \sim P(y|f(i/n)) = e^{-f(i/n)}\frac{[f(i/n)]^y}{y!}. \tag{14.9}$$

In this case

$$-logP(Y_i|f(i/n)) = f(i/n) - Y_i log(f(i/n)) + \text{constant} \tag{14.10}$$

which is a very different loss function, but quite appropriate for many imaging problems.

Before we investigate maximum likelihood estimation for model selection, let's review some of the basic concepts. Let Θ denote a parameter space (e.g., $\Theta = R$), and assume we have observations

$$Y_i \overset{iid}{\sim} P_{\theta^*}(y), \quad i = 1, \cdots, n \tag{14.11}$$

where $\theta^* \in \Theta$ is a parameter determining the density of the $\{Y_i\}$. The ML estimator of θ^* is

$$\begin{aligned} \hat{\theta}_n &= \underset{\theta\in\Theta}{argmax} \textstyle\prod_{i=1}^{n} P_\theta(Y_i) \\ &= \underset{\theta\in\Theta}{argmax} \textstyle\sum_{i=1}^{n} logP_\theta(Y_i) \\ &= \underset{\theta\in\Theta}{argmin} \textstyle\sum_{i=1}^{n} -logP_\theta(Y_i). \end{aligned} \tag{14.12}$$

$\hat{\theta}$ maximizes the expected log-likelihood. To see this, let's compare the expected log-likelihood of θ^* with any other $\theta \in \Theta$.

$$\begin{aligned} E\left[logP_{\theta^*}(Y) - logP_\theta(Y)\right] &= E\left[log\frac{P_{\theta^*}(Y)}{P_\theta(Y)}\right] \\ &= \int log\frac{P_{\theta^*}(y)}{P_\theta(y)} P_{\theta^*}(y)\, dy \\ &= K(P_\theta, P_{\theta^*}) \quad \text{the KL divergence} \\ &\geq 0 \quad \text{with equality iff } P_{\theta^*} = P_\theta. \end{aligned} \tag{14.13}$$

Why?

$$\begin{aligned} -E\left[log\frac{P_{\theta^*}(y)}{P_\theta(y)}\right] &= E\left[log\frac{P_\theta(y)}{P_{\theta^*}(y)}\right] \\ &\leq logE\left[\frac{P_\theta(y)}{P_{\theta^*}(y)}\right] \\ &= log\int P_\theta(y)\,dy = 0 \\ &\Rightarrow K(P_\theta, P_{\theta^*}) \geq 0 \end{aligned} \tag{14.14}$$

On the other hand, since $\hat{\theta}_n$ maximizes the likelihood over $\theta \in \Theta$, we have

$$\sum_{i=1}^{n} log\frac{P_{\theta^*}(Y_i)}{P_{\hat{\theta}_n}(Y_i)} = \sum_{i=1}^{n} logP_{\theta^*}(Y_i) - logP_{\hat{\theta}_n}(Y_i) \leq 0. \tag{14.15}$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} log\frac{P_{\theta^*}(Y_i)}{P_{\hat{\theta}_n}(Y_i)} - K\left(P_{\hat{\theta}_n}, P_{\theta^*}\right) + K\left(P_{\hat{\theta}_n}, P_{\theta^*}\right) \leq 0 \tag{14.16}$$

or re-arranging

$$K\left(P_{\hat{\theta}_n}, P_{\theta^*}\right) \leq \left|\frac{1}{n}\sum_{i=1}^{n} log\frac{P_{\theta^*}(Y_i)}{P_{\hat{\theta}_n}(Y_i)} - K\left(P_{\hat{\theta}_n}, P_{\theta^*}\right)\right|. \tag{14.17}$$

Notice that the quantity

$$\frac{1}{n}\sum_{i=1}^{n} log\frac{P_{\theta^*}(Y_i)}{P_\theta(Y_i)} \tag{14.18}$$

is an empirical average whose mean is $K(P_\theta, P_{\theta^*})$. By the law of large numbers, for each $\theta \in \Theta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} log\frac{P_{\theta^*}(Y_i)}{P_\theta(Y_i)} - K(P_\theta, P_{\theta^*})\right| \overset{a.s.}{\to} 0. \tag{14.19}$$

If this also holds for the sequence $\{\hat{\theta}_n\}$, then we have

$$K\left(P_{\hat{\theta}_n}, P_{\theta^*}\right) \leq \left|\frac{1}{n}\sum log\frac{P_{\theta^*}(Y_i)}{P_{\hat{\theta}_n}(Y_i)} - K\left(P_{\hat{\theta}_n}, P_{\theta^*}\right)\right| \to 0 \,\text{as}\, n \to \infty \tag{14.20}$$

which implies that

$$P_{\hat{\theta}_n} \to P_{\theta^*} \tag{14.21}$$

which often implies that

$$\hat{\theta}_n \to \theta^* \tag{14.22}$$

in some appropriate sense (e.g., point-wise or in norm).

Example 14.1: Gaussian Distributions

$$P_{\theta^*}(y) = \frac{1}{\sqrt{\pi}} e^{-(y-\theta^*)^2} \tag{14.23}$$

$$\Theta = \mathbb{R}, \quad \{Y_i\}_{i=1}^n \overset{iid}{\sim} P_{\theta^*}(y) \tag{14.24}$$

$$\begin{aligned} K(P_\theta, P_{\theta^*}) &= \int log \frac{P_{\theta^*}(y)}{P_\theta(y)} P_{\theta^*}(y)\, dy \\ &= \int \left[(y-\theta)^2 - (y-\theta^*)^2 \right] P_{\theta^*}(y)\, dy \\ &= E_{\theta^*}\left[(y-\theta)^2 \right] - E_{\theta^*}\left[(y-\theta^*)^2 \right] \\ &= E_{\theta^*}\left[Y^2 - 2Y\theta + \theta^2 \right] - 1/2 \\ &= (\theta^*)^2 + 1/2 - 2\theta^*\theta + \theta^2 - 1/2 \\ &= (\theta^* - \theta)^2 \end{aligned} \tag{14.25}$$

$$\Rightarrow \theta^* \text{ maximizes } E\left[log P_\theta(Y) \right] \text{ wrt } \theta \in \Theta \tag{14.26}$$

$$\begin{aligned} \hat{\theta}_n &= \underset{\theta}{argmax} \{ -\sum (Y_i - \theta)^2 \} \\ &= \underset{\theta}{argmin} \{ \sum (Y_i - \theta)^2 \} \\ &= \frac{1}{n} \sum_{i=1}^n Y_i \end{aligned} \tag{14.27}$$

14.1.1 Hellinger Distance

The KL divergence is not a distance function.

$$K(P_{\theta_1}, P_{\theta_2}) \neq K(P_{\theta_2}, P_{\theta_1}) \tag{14.28}$$

Therefore, it is often more convenient to work with the Hellinger metric,

$$H(P_{\theta_1}, P_{\theta_2}) = \left(\int \left(P_{\theta_1}^{\frac{1}{2}} - P_{\theta_2}^{\frac{1}{2}} \right)^2 dy \right)^{\frac{1}{2}}. \tag{14.29}$$

The Hellinger metric is symmetric, non-negative and

$$H(P_{\theta_1}, P_{\theta_2}) = H(P_{\theta_2}, P_{\theta_1}) \tag{14.30}$$

and therefore it is a distance measure. Furthermore, the squared Hellinger distance lower bounds the KL divergence, so convergence in KL divergence implies convergence of the Hellinger distance.

Proposition 1

$$H^2(P_{\theta_1}, P_{\theta_2}) \leq K(P_{\theta_1}, P_{\theta_2}) \tag{14.31}$$

Proof:

$$\begin{aligned}
H\left(P_{\theta_1}, P_{\theta_2}\right) &= \int\left(\sqrt{P_{\theta_1}(y)}-\sqrt{P_{\theta_2}(y)}\right)^2 dy \\
&= \int P_{\theta_1}(y)\,dy + \int P_{\theta_2}(y)\,dy - 2\int\sqrt{P_{\theta_1}(y)}\sqrt{P_{\theta_2}(y)}dy \\
&= 2 - 2\int\sqrt{P_{\theta_1}(y)}\sqrt{P_{\theta_2}(y)}dy, \quad \text{since } \int P_\theta(y)\,dy = 1\,\forall\theta \\
&= 2\left(1 - E_{\theta_2}\left[\sqrt{P_{\theta_1}(Y)/P_{\theta_2}(Y)}\right]\right) \\
&\leq 2log\left(E_{\theta_2}\left[\sqrt{P_{\theta_2}(Y)/P_{\theta_1}(Y)}\right]\right), \quad \text{since } 1-x \leq -logx \\
&\leq 2E_{\theta_2}\left[log\sqrt{P_{\theta_2}(Y)/P_{\theta_1}(Y)}\right], \quad \text{by Jensen's inequality} \\
&= E_{\theta_2}\left[log\left(P_{\theta_2}(Y)/P_{\theta_1}(Y)\right)\right] \;\equiv\; K\left(P_{\theta_1}, P_{\theta_2}\right)
\end{aligned} \tag{14.32}$$

Note that in the proof we also showed that

$$H\left(P_{\theta_1}, P_{\theta_2}\right) = 2\left(1 - \int\sqrt{P_{\theta_1}(y)}\sqrt{P_{\theta_2}(y)}dy\right) \tag{14.33}$$

and using the fact $logx \leq x-1$ again, we have

$$H\left(P_{\theta_1}, P_{\theta_2}\right) \leq -2log\left(\int\sqrt{P_{\theta_1}(y)}\sqrt{P_{\theta_2}(y)}dy\right). \tag{14.34}$$

The quantity inside the log is called the **affinity** between P_{θ_1} and P_{θ_2}:

$$A\left(P_{\theta_1}, P_{\theta_2}\right) = \int\sqrt{P_{\theta_1}(y)}\sqrt{P_{\theta_2}(y)}dy. \tag{14.35}$$

This is another measure of closeness between P_{θ_1} and P_{θ_2}.

Example 14.2: Gaussian Distributions

$$P_\theta(y) = \frac{1}{\pi}e^{-(y-\theta)^2} \tag{14.36}$$

$$\begin{aligned}
&-2log\int\sqrt{P_{\theta_1}(y)}\sqrt{P_{\theta_2}(y)}dy \\
&= -2log\int\left(\tfrac{1}{\sqrt{\pi}}e^{-(y-\theta_1)^2}\right)^{\frac{1}{2}}\left(\tfrac{1}{\sqrt{\pi}}e^{-(y-\theta_2)^2}\right)^{\frac{1}{2}}dy \\
&= -2log\left(\int\tfrac{1}{\sqrt{\pi}}e^{-\left[\frac{(y-\theta_1)^2}{2}+\frac{(y-\theta_2)^2}{2}\right]}dy\right) \\
&= -2log\left(\int\tfrac{1}{\sqrt{\pi}}e^{-\left[\left(y-\left(\frac{\theta_1+\theta_2}{2}\right)\right)^2+\left(\frac{\theta_1-\theta_2}{2}\right)^2\right]}dy\right) \\
&= -2loge^{-\left(\frac{\theta_1-\theta_2}{2}\right)^2} \\
&= \tfrac{1}{2}(\theta_1-\theta_2)^2
\end{aligned} \tag{14.37}$$

$$\begin{aligned}
&\Rightarrow \quad -2logA\left(P_{\theta_1}, P_{\theta_2}\right) = \tfrac{1}{2}(\theta_1-\theta_2)^2 \quad \text{for Gaussian distributions} \\
&\Rightarrow \quad H\left(P_{\theta_1}, P_{\theta_2}\right) \leq \tfrac{1}{2}(\theta_1-\theta_2)^2 \quad \text{for Gaussian.}
\end{aligned} \tag{14.38}$$

Example 14.3: Poisson Distributions
If $P_\theta(y) = e^{-\theta}\frac{\theta^y}{y!}, \theta \geq 0$, then

$$-2logA\left(P_{\theta_1}, P_{\theta_2}\right) = \left(\sqrt{\theta_1}-\sqrt{\theta_2}\right)^2. \tag{14.39}$$

Summary

$$Y_i \overset{iid}{\sim} P_{\theta^*} \tag{14.40}$$

1. Maximum likelihood estimator maximizes the empirical average

$$\frac{1}{n}\sum_{i=1}^{n} logP_\theta\left(Y_i\right) \tag{14.41}$$

   (our empirical risk is negative log-likelihood)
2. θ^* maximizes the expectation

$$E\left[\frac{1}{n}\sum_{i=1}^{n} logP_\theta\left(Y_i\right)\right] \tag{14.42}$$

   (the risk is the expected negative log-likelihood)
3.

$$\frac{1}{n}\sum_{i=1}^{n} logP_\theta\left(Y_i\right) \overset{a.s.}{\to} E\left[\frac{1}{n}\sum_{i=1}^{n} logP_\theta\left(Y_i\right)\right] \tag{14.43}$$

   so we expect some sort of concentration of measure.
4. In particular, since

$$\frac{1}{n}\sum_{i=1}^{n} log\frac{P_{\theta^*}\left(Y_i\right)}{P_\theta\left(Y_i\right)} \overset{a.s.}{\to} K\left(P_\theta, P_{\theta^*}\right) \tag{14.44}$$

   we might expect that $K\left(P_{\hat{\theta}_n}, P_{\theta^*}\right) \to 0$ for the sequence of estimates $\{P_{\hat{\theta}_n}\}_{n=1}^{\infty}$.

So, the point is that maximum likelihood estimator is just a special case of a loss function in learning. Due to its special structure, we are naturally led to consider KL divergences, Hellinger distances, and Affinities.

Chapter 15

Maximum Likelihood and Complexity Regularization[1]

15.1 Review : Maximum Likelihood Estimation

In the last lecture (Chapter 14), we have n i.i.d observations drawn from an unknown distribution

$$Y_i \overset{i.i.d.}{\sim} p_{\theta^*} \quad , \; i = \{1, ..., n\} \tag{15.1}$$

$$\text{where } \theta^* \in \Theta. \tag{15.2}$$

With **loss function** defined as $l(\theta, Y_i) = -logp_\theta(Y_i)$, the empirical risk is

$$\hat{R_n} = -\frac{1}{n}\sum_{i=1}^{n} logp_\theta(Y_i). \tag{15.3}$$

Essentially, we want to choose a distribution from the collection of distributions within the parameter space that minimizes the empirical risk,**i.e.,** we would like to select

$$p_{\hat{\theta}_n} \in \mathcal{P} = \{p_\theta\}_{\theta \in \Theta} \tag{15.4}$$

where

$$\hat{\theta_n} = arg\min_{\theta \in \Theta} \quad -\sum_{i=1}^{n} logp_\theta(Y_i). \tag{15.5}$$

The risk is defined as

$$R(\theta) = E[l(\theta, Y)] = -E[logp_\theta(Y)]. \tag{15.6}$$

Note that θ^* minimizes $R(\theta)$ over Θ.

$$\begin{aligned} \theta^* &= arg\min_{\theta \in \Theta} \quad -E[logp_\theta(Y)] \\ &= \underset{\theta \in \Theta}{argmin} \quad -\int logp_\theta(y) \cdot p_{\theta^*}(y)\, dy. \end{aligned} \tag{15.7}$$

[1]This content is available online at <http://cnx.org/content/m16275/1.2/>.

Finally, the excess risk of θ is defined as

$$R(\theta) - R(\theta^*) = \int log \frac{p_{\theta^*}(y)}{p_\theta(y)} p_{\theta^*}(y)\, dy \equiv K(p_\theta, p_{\theta^*}) . \tag{15.8}$$

We recognized that the excess risk corresponding to this loss function is simply the **Kullback-Leibler (KL) Divergence** or **Relative Entropy**, denoted by $K(p_{\theta_1}, p_{\theta_2})$. It is easy to see that $K(p_{\theta_1}, p_{\theta_2})$ is always non-negative and is zero if and only if $p_{\theta_1} = p_{\theta_2}$. KL divergence measures how different two probability distributions are and therefore is natural to measure convergence of the maximum likelihood procedures. However, $K(p_{\theta_1}, p_{\theta_2})$ is not a distance metric because it is not symmetric and does not satisfy the triangle inequality. For this reason, two other quantities play a key role in maximum likelihood estimation, namely **Hellinger Distance** and **Affinity**.

The Hellinger distance is defined as

$$H(p_{\theta_1}, p_{\theta_2}) = \left(\int \left(\sqrt{p_{\theta_1}(y)} - \sqrt{p_{\theta_2}(y)} \right)^2 dy \right)^{\frac{1}{2}} . \tag{15.9}$$

We proved that the squared Hellinger distance lower bounds the KL divergence:

$$\begin{aligned} H^2(p_{\theta_1}, p_{\theta_2}) &\leq K(p_{\theta_1}, p_{\theta_2}) \\ H^2(p_{\theta_1}, p_{\theta_2}) &\leq K(p_{\theta_2}, p_{\theta_1}) . \end{aligned} \tag{15.10}$$

The affinity is defined as

$$A(p_{\theta_1}, p_{\theta_2}) = \int \sqrt{p_{\theta_1} \cdot p_{\theta_2}(y)}\, dy . \tag{15.11}$$

we also proved that

$$H^2(p_{\theta_1}, p_{\theta_2}) \leq -2log\left(A(p_{\theta_1}, p_{\theta_2})\right) . \tag{15.12}$$

Example 15.1: Gaussian Distribution

Y is Gaussian with mean θ and variance σ^2.

$$p_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2}} . \tag{15.13}$$

First, look at

$$log \frac{p_{\theta_2}}{p_{\theta_1}} = \frac{1}{2\sigma^2} \left[\left(\theta_1^2 - \theta_2^2\right) - 2(\theta_1 - \theta_2) y \right] . \tag{15.14}$$

Then,

$$\begin{aligned}
K\left(p_{\theta_1}, p_{\theta_2}\right) &= E_{\theta_2}\left[log\,\tfrac{p_{\theta_2}}{p_{\theta_1}}\right] \\
&= \tfrac{\theta_1^2-\theta_2^2}{2\sigma^2} - \tfrac{2(\theta_1-\theta_2)}{2\sigma^2}\underbrace{\int y\cdot p_{\theta_2}(y)\,dy}_{E[Y]=\theta_2} \\
&= \tfrac{1}{2\sigma^2}\left(\theta_1^2+\theta_2^2-2\theta_1\theta_2\right) \quad = \tfrac{\left(\theta_1^2-\theta_2\right)^2}{2\sigma^2}. \\
-2log\,A\left(p_{\theta_1}, p_{\theta_2}\right) &= -2log\left(\int\left(\tfrac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(y-\theta_1)^2}{2\sigma^2}}\right)^{1/2}\cdot\left(\tfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\theta_2)^2}{2\sigma^2}}\right)^{1/2}dy\right) \\
&= -2log\left(\int\tfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\theta_1)^2}{4\sigma^2}-\frac{(y-\theta_2)^2}{4\sigma^2}}\,dy\right) \\
&= -2log\left(\int\tfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\left[\left(y-\frac{\theta_1+\theta_2}{2}\right)^2+\left(\frac{\theta_1-\theta_2}{2}\right)^2\right]}dy\right) \\
&= -2log\,e^{-\frac{\left(\frac{\theta_1-\theta_2}{2}\right)^2}{2\sigma^2}} \\
&= \tfrac{(\theta_1-\theta_2)^2}{4\sigma^2} \quad = \tfrac{1}{2}K\left(p_{\theta_1}, p_{\theta_2}\right) \quad \geq H^2\left(p_{\theta_1}, p_{\theta_2}\right).
\end{aligned} \tag{15.15}$$

15.2 Maximum likelihood estimation and Complexity regularization

Suppose that we have n i.i.d training samples, $\{X_i, Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} p_{XY}$.

Using conditional probability, p_{XY} can be written as

$$p_{XY}(x,y) = p_X(x)\cdot p_{Y|X=x}(y)\,. \tag{15.16}$$

Let's assume for the moment that p_X is completely unknown, but $p_{Y|X=x}(y)$ has a special form:

$$p_{Y|X=x}(y) = p_{f^*(x)}(y) \tag{15.17}$$

where $p_{Y|X=x}(y)$ is a known parametric density function with parameter $f^*(x)$.

Example 15.2: Signal-plus-noise observation model

$$Y_i = f^*(X_i) + W_i \quad , i = 1, ..., n \tag{15.18}$$

where $W_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0,\sigma^2\right)$ and $X_i \overset{i.i.d.}{\sim} p_X$.

$$p_{f^*(x)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-f^*(x))^2}{2\sigma^2}} \tag{15.19}$$

$Y|X = x \sim \text{Poisson}(f^*(x))$

$$p_{f^*(x)}(y) = e^{-f^*(x)}\frac{[f^*(x)]^y}{y!}\,. \tag{15.20}$$

The likelihood loss function is

$$\begin{aligned}
l\left(f(x), y\right) &= -log\,p_{XY}(X,Y) \\
&= -log\,p_X(X) - log\,p_{Y|X}(Y|X) \\
&= -log\,p_X(X) - log\,p_{f(X)}(Y)\,.
\end{aligned} \tag{15.21}$$

The **expected loss** is

$$\begin{aligned} E\left[l\left(f\left(X\right),Y\right)\right] &= E_X\left[E_{Y|X}\left[l\left(f\left(X\right),Y\right)|X=x\right]\right] \\ &= E_X\left[\,E_{Y|X}\left[-log\,p_X\left(x\right)-log\,p_{f(x)}\left(Y\right)|X=x\right]\,\right] \\ &= -E_X\left[\,log\,p_X\left(X\right)\,\right]-E_X\left[\,E_{Y|X}\left[\,log\,p_{f(x)}\left(Y\right)|X=x\,\right]\,\right] \\ &= -E_X\left[\,log\,p_X\left(X\right)\,\right]-E\left[\,log\,p_{f(X)}\left(Y\right)\,\right]. \end{aligned} \tag{15.22}$$

Notice that the first term is a constant with respect to f.

Hence, we define our risk to be

$$\begin{aligned} R\left(f\right) &= -E\left[\,log\,p_{f(X)}\left(Y\right)\,\right] \\ &= -E_X\left[\,E_{Y|X}\left[log\,p_{f(x)}\left(Y\right)|X=x\right]\,\right] \\ &= -\int\left(\int log\,p_{f(x)}\left(y\right)\cdot p_{f^*(x)}\left(y\right)\,dy\right)\,p_X\left(x\right)\,dx\,. \end{aligned} \tag{15.23}$$

The function f^* minimizes this risk since $f\left(x\right)=f^*\left(x\right)$ minimizes the integrand.

Our empirical risk is the negative log-likelihood of the training samples:

$$\hat{R}_n\left(f\right)=\frac{1}{n}\sum_{i=1}^{n}\,-log\,p_{f(X_i)}\left(Y_i\right). \tag{15.24}$$

The value $\frac{1}{n}$ is the **empirical** probability of observing $X=X_i$.

Often in function estimation, we have control over where we sample X. Let's assume that $\mathcal{X}=[0,1]^d$ and $\mathcal{Y}=\mathbf{R}$. Suppose we sample $\mathcal{X}$ uniformly with $n=m^d$ samples for some positive integer m (**i.e.,** ,take m evenly spaced samples in each coordinate).

Let x_i ,$i=1,...,n$ denote these sample points, and assume that $Y_i\sim p_{f^*(x_i)}\left(y\right)$. Then, our empirical risk is

$$\hat{R}_n\left(f\right)=\frac{1}{n}\sum_{i=1}^{n}l\left(f\left(x_i\right),Y_i\right)=\frac{1}{n}\sum_{i=1}^{n}\,-log\,p_{f(x_i)}\left(Y_i\right)\,. \tag{15.25}$$

Note that x_i is now a deterministic quantity.

Our risk is

$$\begin{aligned} R\left(f\right) &= -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\,E\left[log\,p_{f(x_i)}\left(Y_i\right)\right] \\ &= -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left[\int log\,p_{f(x_i)}\left(y_i\right)\cdot p_{f^*(x_i)}\left(y_i\right)\,dy_i\right]\,. \end{aligned} \tag{15.26}$$

The risk is minimized by f^*. However, f^* is not a unique minimizer. Any f that agrees with f^* at the point $\{x_i,Y_i\}$ also minimizes this risk.

Now, we will make use of the following vector and shorthand notation. The uppercase Y denotes a random variable, while the lowercase y and x denote deterministic quantities.

$$Y=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}\quad y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}\quad x=\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} \tag{15.27}$$

Then,

$p_f\left(Y\right)=\prod_{i=1}^{n}p\left(Y_i|f\left(x_i\right)\right)$ (random)

$p_f\left(y\right)=\prod_{i=1}^{n}p\left(y_i|f\left(x_i\right)\right)$ (deterministic) .

With this notation, the empirical risk and the true risk can be written as

$$\begin{aligned} \hat{R}_n(f) &= -\tfrac{1}{n} log p_f(Y) . \\ R(f) &= -\tfrac{1}{n} E[log p_f(Y)] \\ &= -\tfrac{1}{n} \int log p_f(y) \cdot p_{f^*}(y)\, dy . \end{aligned} \tag{15.28}$$

15.3 Error Bound

Suppose that we have a pool of candidate functions $\mathcal{F}$, and we want to select a function f from $\mathcal{F}$ using the training data. Our usual approach is to show that the distribution of $\hat{R}_n(f)$ concentrates about its mean as n grows. First, we assign a complexity $c(f) > 0$ to each $f \in \mathcal{F}$ so that $\sum 2^{-c(f)} \leq 1$. Then, apply the union bound to get a **uniform** concentration inequality holding for all models in $\mathcal{F}$. Finally, we use this concentration inequality to bound the expected risk of our selected model.

We will essentially accomplish the same result here, but avoid the need for explicit concentration inequalities and instead make use of the information-theoretic bounds.

We would like to select an $f \in \mathcal{F}$ so that the excess risk is small.

$$\begin{aligned} 0 &\leq R(f) - R(f^*) \\ &= \tfrac{1}{n} E[log p_{f^*}(Y) - log p_f(Y)] \\ &= \tfrac{1}{n} E\left[log \tfrac{p_{f^*}(Y)}{p_f(Y)}\right] \\ &\equiv \tfrac{1}{n} K(p_f, p_{f^*}) \end{aligned} \tag{15.29}$$

where

$$K(p_f, p_{f^*}) = \sum_{i=1}^{n} \underbrace{\left(\int log \frac{p_{f^*(x_i)}(y_i)}{p_{f(x_i)}(y_i)} \cdot p_{f^*(x_i)}(y_i)\, dy_i \right)}_{K\left(p_{f(x_i)}, p_{f^*(x_i)}\right)} \tag{15.30}$$

is again the KL divergence.

Unfortunately, as mentioned before, $K(p_f, p_{f^*})$ is not a true distance. So instead we will focus on the expected squared Hellinger distance as our measure of performance. We will get a bound on

$$\frac{1}{n} E\left[H^2(p_f(Y), p_{f^*}(Y))\right] = \frac{1}{n} \sum_{i=1}^{n} \left(\int \left(\sqrt{p_{f(x_i)}(y_i)} - \sqrt{p_{f^*(x_i)}(y_i)} \right)^2 dy_i \right) . \tag{15.31}$$

15.4 Maximum Complexity-Regularized Likelihood Estimation

Theorem 15.1: Li-Barron 2000, Kolaczyk-Nowak 2002
Let $\{x_i, Y_i\}_{i=1}^n$ be a random sample of training data with $\{Y_i\}$ independent,

$$Y_i \sim p_{f^*(x_i)}(y_i) \quad , i = 1, ..., n \tag{15.32}$$

for some unknown function f^*.

Suppose we have a collection of candidate functions $\mathcal{F}$, and complexities $c(f) > 0, f \in \mathcal{F}$, satisfying

$$\sum_{f\in\mathcal{F}} 2^{-c(f)} \leq 1. \tag{15.33}$$

Define the complexity-regularized estimator

$$\hat{f_n} \equiv arg\min_{f\in\mathcal{F}}\{-\frac{1}{n}\sum_{i=1}^{n} log\, p_f(Y_i) + \frac{2c(f)\,log2}{n}\}. \tag{15.34}$$

Then,

$$\begin{array}{rcl} \frac{1}{n}E\left[H^2\left(p_f(Y), p_{f^*}(Y)\right)\right] & \leq & -\frac{2}{n}E\left[log\left(A\left(p_f(Y), p_{f^*}(Y)\right)\right)\right] \\ & \leq & \min_{f\in\mathcal{F}}\{\frac{1}{n}K\left(p_f, p_{f^*}\right) + \frac{2c(f)log2}{n}\}. \end{array} \tag{15.35}$$

Before proving the theorem, let's look at a special case.

Example 15.3: Gaussian noise

Suppose $Y_i = f(x_i) + W_i \quad , W_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right)$.

$$p_{f(x_i)}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - f(x_i))^2}{2\sigma^2}}. \tag{15.36}$$

Using results from example 1 (Example 15.1: Gaussian Distribution), we have

$$\begin{array}{rcl} -2logA\left(p_{\hat{f}_n}(Y), p_{f^*}(Y)\right) & = & \sum_{i=1}^{n} -2logA\left(p_{\hat{f}_n(x_i)}(Y_i), p_{f^*(x_i)}(Y_i)\right) \\ & = & \sum_{i=1}^{n} -2log\int\sqrt{p_{\hat{f}_n(x_i)}(y_i)\cdot p_{f^*(x_i)}(y_i)}\,dy_i \\ & = & \frac{1}{4\sigma^2}\sum_{i=1}^{n}\left(\hat{f_n}(x_i) - f^*(x_i)\right)^2. \end{array} \tag{15.37}$$

Then,

$$-\frac{2}{n}E\left[log\, A\left(p_{\hat{f}_n}, p_{f^*}\right)\right] = \frac{1}{4\sigma^2 n}\sum_{i=1}^{n} E\left[\left(\hat{f_n}(x_i) - f^*(x_i)\right)^2\right]. \tag{15.38}$$

We also have,

$$\begin{array}{rcl} \frac{1}{n}K\left(p_f, p_{f^*}\right) & = & \frac{1}{n}\sum_{i=1}^{n}\frac{(f(x_i) - f^*(x_i))^2}{2\sigma^2} \\ -log\, p_f(Y) & = & \sum_{i=1}^{n}\frac{(Y_i - f(X_i))^2}{2\sigma^2}. \end{array} \tag{15.39}$$

Combine everything together to get

$$\hat{f_n} = \underset{f\in\mathcal{F}}{argmin}\{\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - f(X_i))^2}{2\sigma^2} + \frac{2c(f)\,log2}{n}\}. \tag{15.40}$$

The theorem tells us that

$$\frac{1}{4n}\sum_{i=1}^{n} E\left[\frac{\left(\hat{f}_n\left(x_i\right)-f^*\left(x_i\right)\right)^2}{\sigma^2}\right] \leq \min_{f\in\mathcal{F}}\{\frac{1}{n}\sum_{i=1}^{n}\frac{\left(f\left(x_i\right)-f^*\left(x_i\right)\right)^2}{2\sigma^2}+\frac{2c\left(f\right)log2}{n}\} \tag{15.41}$$

or

$$\frac{1}{n}\sum_{i=1}^{n} E\left[\left(\hat{f}_n\left(x_i\right)-f^*\left(x_i\right)\right)^2\right] \leq \min_{f\in\mathcal{F}}\{\frac{2}{n}\sum_{i=1}^{n}\left(f\left(x_i\right)-f^*\left(x_i\right)\right)^2+\frac{8\sigma^2c\left(f\right)log2}{n}\}. \tag{15.42}$$

Now let's come back to the proof.

Proof 15.1:

$$\begin{aligned} H^2\left(p_{\hat{f}_n}, p_{f^*}\right) &= \int\left(\sqrt{p_{\hat{f}_n}\left(y\right)}-\sqrt{p_{f^*}\left(y\right)}\right)^2 dy \\ &\leq -2log\underbrace{\left(\int\sqrt{p_{\hat{f}_n}\left(y\right)\cdot p_{f^*}\left(y\right)}\,dy\right)}_{affinity} \end{aligned} \tag{15.43}$$

$$\Rightarrow \tag{15.44}$$

$$E\left[H^2\left(p_{\hat{f}_n}, p_{f^*}\right)\right] \leq 2\,E\left[log\left(\frac{1}{\int\sqrt{p_{\hat{f}_n}\left(y\right)\cdot p_{f^*}\left(y\right)}\,dy}\right)\right]. \tag{15.45}$$

Now, define the theoretical analog of $\hat{f}_n$:

$$f_n = \underset{f\in\mathcal{F}}{argmin}\{\frac{1}{n}K\left(p_f, p_{f^*}\right)+\frac{2c\left(f\right)log2}{n}\}. \tag{15.46}$$

Since

$$\begin{aligned} \hat{f}_n &= \underset{f\in\mathcal{F}}{argmin}\{-\tfrac{1}{n}logp_f\left(Y\right)+\tfrac{2c(f)log2}{n}\} \\ &= \underset{f\in\mathcal{F}}{argmax}\{\tfrac{1}{n}\left(logp_f\left(Y\right)-2c\left(f\right)log2\right)\} \\ &= \underset{f\in\mathcal{F}}{argmax}\{\tfrac{1}{2}\left(logp_f\left(Y\right)-2c\left(f\right)log2\right)\} \\ &= \underset{f\in\mathcal{F}}{argmax}\{log\left(\sqrt{p_f\left(Y\right)}\cdot e^{-c(f)log2}\right)\} \\ &= \underset{f\in\mathcal{F}}{argmax}\{\sqrt{p_f\left(Y\right)}\cdot e^{-c(f)log2}\} \end{aligned} \tag{15.47}$$

we can see that

$$\frac{\sqrt{p_{\hat{f}_n}\left(Y\right)}e^{-c\left(\hat{f}_n\right)log2}}{\sqrt{p_{f_n}\left(Y\right)}e^{-c(f_n)log2}} \geq 1\,. \tag{15.48}$$

Then can write

$$\begin{aligned} E\left[H^2\left(p_{\overset{\wedge}{f_n}}, p_{f^*}\right)\right] &\leq 2\,E\left[log\left(\frac{1}{\int\sqrt{p_{\overset{\wedge}{f_n}}(y)\cdot p_{f^*}(y)}\,dy}\right)\right] \\ &\leq 2\,E\left[log\left(\frac{\sqrt{p_{\overset{\wedge}{f_n}}(Y)}e^{-c\left(\overset{\wedge}{f_n}\right)log2}}{\sqrt{p_{f_n}(Y)}e^{-c(f_n)log2}}\cdot\frac{1}{\int\sqrt{p_{\overset{\wedge}{f_n}}\cdot p_{f^*}}\,dy}\right)\right]. \end{aligned} \tag{15.49}$$

Now, simply multiply the argument inside the log by $\sqrt{\frac{p_{f^*}(Y)}{p_{f^*}(Y)}}$ to get

$$\begin{aligned} E\left[H^2\left(p_{\overset{\wedge}{f_n}}, p_{f^*}\right)\right] &\leq 2\,E\left[log\left(\frac{\sqrt{p_{f^*}(Y)}}{\sqrt{p_{f_n}(Y)}}\frac{\sqrt{p_{\overset{\wedge}{f_n}}(Y)}}{\sqrt{p_{f^*}(Y)}}\frac{e^{-c\left(\overset{\wedge}{f_n}\right)log2}}{e^{-c(f_n)log2}}\cdot\frac{1}{\int\sqrt{p_{\overset{\wedge}{f_n}}(y)\cdot p_{f^*}(y)}\,dy}\right)\right] \\ &= E\left[log\left(\frac{p_{f^*}(Y)}{p_{f_n}(Y)}\right)\right] + 2c\left(f_n\right)log2 \\ &\quad + 2E\left[log\left(\frac{\sqrt{p_{\overset{\wedge}{f_n}}(Y)}}{\sqrt{p_{f^*}(Y)}}\cdot\frac{e^{-c\left(\overset{\wedge}{f_n}\right)log2}}{\int\sqrt{p_{\overset{\wedge}{f_n}}(y)\cdot p_{f^*}(y)}\,dy}\right)\right] \\ &= K\left(p_{f_n}, p_{f^*}\right) + 2c\left(f_n\right)log2 \\ &\quad + 2E\left[log\left(\frac{\sqrt{p_{\overset{\wedge}{f_n}}(Y)}}{\sqrt{p_{f^*}(Y)}}\cdot\frac{e^{-c\left(\overset{\wedge}{f_n}\right)log2}}{\int\sqrt{p_{\overset{\wedge}{f_n}}(y)\cdot p_{f^*}(y)}\,dy}\right)\right]. \end{aligned} \tag{15.50}$$

The terms $K\left(p_{f_n}, p_{f^*}\right) + 2c\left(f_n\right)log2$ are precisely what we wanted for the upper bound of the theorem. So, to finish the proof we only need to show that the last term is non-positive. Applying Jensen's inequality, we get

$$2E\left[log\left(\frac{\sqrt{p_{\overset{\wedge}{f_n}}(Y)}}{\sqrt{p_{f^*}(Y)}}\cdot\frac{e^{-c\left(\overset{\wedge}{f_n}\right)log2}}{\int\sqrt{p_{\overset{\wedge}{f_n}}(y)\cdot p_{f^*}(y)}\,dy}\right)\right] \leq 2log\left(E\left[e^{-c\left(\overset{\wedge}{f_n}\right)log2}\cdot\frac{\sqrt{\frac{p_{\overset{\wedge}{f_n}}(Y)}{p_{f^*}(Y)}}}{\int\sqrt{p_{\overset{\wedge}{f_n}}(y)\cdot p_{f^*}(y)}\,dy}\right]\right). \tag{15.51}$$

Both Y and $\overset{\wedge}{f_n}$ are random, which makes the expectation difficult to compute. However, we can simplify the problem using the union bound, which eliminates the dependence on $\overset{\wedge}{f_n}$:

$$\begin{aligned} 2E\left[log\left(\frac{\sqrt{p_{\overset{\wedge}{f_n}}(Y)}}{\sqrt{p_{f^*}(Y)}}\cdot\frac{e^{-c\left(\overset{\wedge}{f_n}\right)log2}}{\int\sqrt{p_{\overset{\wedge}{f_n}}(y)\cdot p_{f^*}(y)}\,dy}\right)\right] &\leq 2log\left(E\left[\sum_{f\in\mathcal{F}}e^{-c(f)log2}\cdot\frac{\sqrt{\frac{p_f(Y)}{p_{f^*}(Y)}}}{\int\sqrt{p_f(y)\cdot p_{f^*}(y)}\,dy}\right]\right) \\ &= 2log\left(\sum_{f\in\mathcal{F}}2^{-c(f)}\frac{E\left[\sqrt{\frac{p_f(Y)}{p_{f^*}(Y)}}\right]}{\int\sqrt{p_f(y)\cdot p_{f^*}(y)}\,dy}\right) \\ &= 2log\left(\sum_{f\in\mathcal{F}}2^{-c(f)}\right) \\ &\leq 0. \end{aligned} \tag{15.52}$$

where the last two lines come from

$$E\left[\sqrt{\frac{p_f\left(Y\right)}{p_{f^*}\left(Y\right)}}\right]=\int\sqrt{\frac{p_f\left(y\right)}{p_{f^*}\left(y\right)}}\cdot p_{f^*}\left(y\right)\,dy=\int\sqrt{p_f\left(y\right)\cdot p_{f^*}\left(y\right)}\,dy \tag{15.53}$$

and

$$\sum_{f\in\mathcal{F}}2^{-c(f)}\leq 1. \tag{15.54}$$

Chapter 16

Denoising II: Adapting to Unknown Smoothness[1]

16.1 Review: Denoising in Smooth Function Spaces I - Method of Sieves

Suppose we make noisy measurements of a smooth function:

$$Y_i = f^*(x_i) + W_i, \;\; i = \{1, ..., n\}, \tag{16.1}$$

where

$$W_i \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right) \tag{16.2}$$

and

$$x_i = \left(\frac{i}{n}\right). \tag{16.3}$$

The unknown function f^* is a map

$$f^* : [0,1] \to \mathbf{R}. \tag{16.4}$$

In Lecture 4 (Chapter 5), we consider this problem in the case where f^* was Lipschitz on $[0,1]$. That is, f^* satisfied

$$|f^*(t) - f^*(s)| \leq L|t-s|, \quad \forall t, s \in [0,1] \tag{16.5}$$

where $L > 0$ is a constant. In that case, we showed that by using a piecewise constant function on a partition of $n^{\frac{1}{3}}$ equal-size bins Figure 16.1 we were able to obtain an estimator $\hat{f}_n$ whose mean square error was

$$E\left[\| f^* - \hat{f}_n \|^2\right] = O\left(n^{-\frac{2}{3}}\right). \tag{16.6}$$

[1] This content is available online at <http://cnx.org/content/m16268/1.2/>.

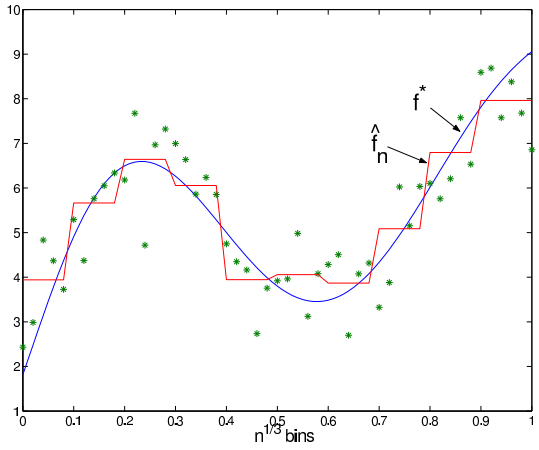


Figure 16.1: Example of the piecewise constant approximation of f^*

In this lecture we will use the Maximum Complexity-Regularized Likelihood Estimation result we derived in Lecture 14 (Chapter 15) to extend our denoising scheme in several important ways.

To begin with let's consider a broader class of functions.

16.2 Hölder Spaces

For $0 < \alpha < 1$, define the space of functions

$$H^\alpha(C_\alpha) = \{|f| < C_\alpha : \sup_{x,h} \frac{|f(x+h) - f(x)|}{|h|^\alpha} \leq C_\alpha\} \tag{16.7}$$

for some constant $C_\alpha < \infty$ and where $f \in L_\infty$.H^α above contains functions that are bounded, but less smooth than Lipschitz functions. Indeed, the space of Lipschitz functions can be defined as H^1 ($\alpha = 1$)

$$H^1(C_1) = \{|f| < C_1 : \sup_{x,h} \frac{|f(x+h) - f(x)|}{|h|} \leq C_1\} \tag{16.8}$$

for $C_1 < \infty$. Functions in H^1 are continuous, but those in $H^\alpha, \alpha < 1$, are not in general.

Let's also consider functions that are smoother than Lipschitz. If $\alpha = 1+\beta$, where $0 < \beta < 1$, then define

$$H^\alpha(C_\alpha) = \{f \in H^1(C_\alpha) : \frac{\partial f}{\partial x} \in H^\beta(C_\alpha)\}. \tag{16.9}$$

In other words, $H^\alpha, 1 < \alpha < 2$, contains Lipschitz functions that are also differentiable and their derivatives are Hölder smooth with smoothness $\beta = \alpha - 1$.

And finally, let

$$H^2(C_2) = \{f : \frac{\partial f}{\partial x} \in H^1(C_2)\} \tag{16.10}$$

contain functions that have continuous derivatives, but that are not necessarily twice-differentiable.

If $f \in H^\alpha(C_\alpha)$, $0 < \alpha \leq 2$, then we say that f is Hölder$-\alpha$ smooth with Hölder constant C_α. The notion of Hölder smoothness can also be extended to $\alpha > 2$ in a straightforward way.

Note: If $\alpha_1 < \alpha_2$ then

$$f \in H^{\alpha_2} \Rightarrow f \in H^{\alpha_1}. \tag{16.11}$$

Summarizing, we can describe Hölder spaces as follows. If $f^* \in H^\alpha(C_\alpha)$ for some $0 < \alpha \leq 2$ and $C_\alpha < \infty$, then

(i): $0 < \alpha \leq 1$ $\quad |f^*(t) - f^*(s)| \leq C_\alpha |t-s|^\alpha$

(ii): $1 < \alpha \leq 2$ $\quad \left| \frac{\partial f^*}{\partial x}(t) - \frac{\partial f^*}{\partial x}(s) \right| \leq C_\alpha |t-s|^{\alpha-1}$

Note that in general there is a natural relationship between the Hölder space containing the function and the approximation class used to estimate the function. Here we will consider functions which are Hölder$-\alpha$ smooth where $0 < \alpha \leq 2$ and work with piecewise linear approximations. If we were to consider smoother functions, $\alpha > 2$ we would need consider higher order approximation functions, i.e. quadratic, cubic, etc.

16.3 Denoising Example for Signal-plus-Gaussian Noise Observation Model

Now let's assume $f^* \in H^\alpha(C_\alpha)$ for some unknown $\alpha\,(0 < \alpha \leq 2)$; i.e. we don't know how smooth f^* is. We will use our observations

$$Y_i = f^*(x_i) + W_i, \;\; i = \{1, ..., n\}, \tag{16.12}$$

to construct an estimator $\hat{f}_n$. Intuitively, the smoother f^* is, the better we should be able to estimate it. Can we take advantage of extra smoothness in f^* if we don't know how smooth it is? The smoother f^* is, the more averaging we can perform to reduce noise. In other words for smoother f^* we should average over larger bins. Also, we will need to exploit the extra smoothness in our approximation of f^*. To that end, we will consider candidate functions that are piecewise linear functions on uniform partitions of $[0,1]$. Let

$$\mathcal{F}_k = \{ \; |f| \leq C : \quad \begin{array}{c} \text{f is piecewise linear on } \left[0, \frac{1}{k}\right), \left[\frac{1}{k}, \frac{2}{k}\right), ... \left[\frac{k-1}{k}, 1\right) \text{ and the} \\ \text{coefficients of each line segment are quantized to } \frac{1}{2} log n \text{ bits.} \end{array} \; \}. \tag{16.13}$$

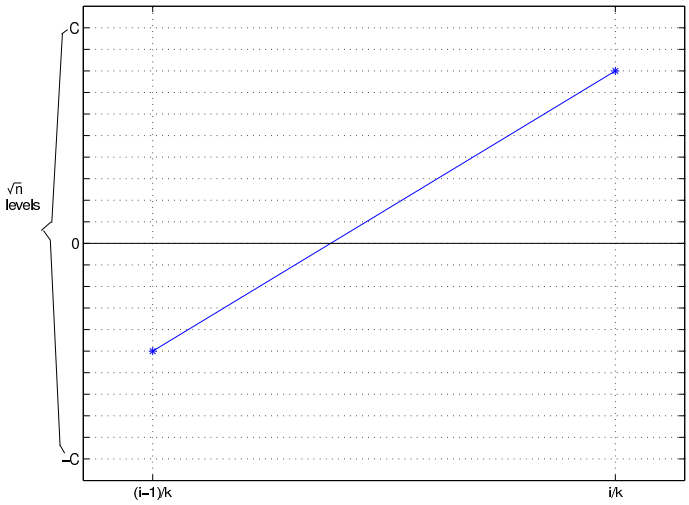


Figure 16.2: Example on the quantization of f on interval $\left[\frac{i-1}{k}, \frac{i}{k}\right)$

The start and end points of each line segment are each one of $\sqrt{n}$ discrete values, as indicated in Figure 16.2. Since each line may start at any of the $\sqrt{n}$ levels and terminate at any of the $\sqrt{n}$ levels, there are a total of n possible lines for each segment.

Given that there are k intervals we have

$$|\mathcal{F}_k| = n^k \Rightarrow log\,|\mathcal{F}_k| = klogn. \tag{16.14}$$

Therefore we can use $klogn$ bits to describe a function $f \in \mathcal{F}_k$.

Let

$$\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k. \tag{16.15}$$

Construct a prefix code for every $f \in \mathcal{F}$ by

$$\begin{aligned} &\text{(i)} \quad \text{Use } \underbrace{000\cdots1}_{k\,bits} \text{ to encode the smallest } k \text{ such that } f \in \mathcal{F}_k \\ &\text{(ii)} \quad \text{Use } klogn \text{ bits to encode which element of } \mathcal{F}_k \text{ we are considering.} \end{aligned} \tag{16.16}$$

Thus, if $f \in \mathcal{F}_k$, then the prefix code associated with f has codeword length

$$c\,(f) = k + klogn = k\,(1 + logn) \tag{16.17}$$

which satisfies the Kraft Inequality

$$\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1. \tag{16.18}$$

Now we will apply our complexity regularization result to select a function $\hat{f}_n$ from $\mathcal{F}$ and bound its risk. We are assuming Gaussian errors, so

$$-logp_f\left(Y_i\right) = \frac{\left(Y_i - f\left(\frac{i}{n}\right)\right)^2}{2\sigma^2} + constant. \tag{16.19}$$

We can ignore the constant term and so our empirical selection is

$$\hat{f}_n = \underset{f\in\mathcal{F}}{argmin}\{\frac{1}{n}\sum_{i=1}^{n}\frac{\left(Y_i - f\left(\frac{i}{n}\right)\right)^2}{2\sigma^2} + \frac{2c\left(f\right)log2}{n}\}. \tag{16.20}$$

We can compute $\hat{f}_n$ according to:

For $k = 1, ..., n$

$$\hat{f}_n^{(k)} = \underset{f\in\mathcal{F}_k}{argmin}\ \hat{R}_n\left(f\right) = \underset{f\in\mathcal{F}_k}{argmin}\frac{1}{n}\sum_{i=1}^{n}\frac{\left(Y_i - f\left(\frac{i}{n}\right)\right)^2}{2\sigma^2} \tag{16.21}$$

then select

$$\hat{k} = arg\ \underset{k=1,...,n}{min}\{\hat{R}_n\left(\hat{f}_n^{(k)}\right) + \frac{2k\left(1+logn\right)log2}{n}\} \tag{16.22}$$

and finally

$$\hat{f}_n = \hat{f}_n^{\left(\hat{k}\right)}. \tag{16.23}$$

Because the KL divergence and $-2log$ **affinity** simply reduce to squared error in the Gaussian case (Lecture 14) (Chapter 15), we arrive at a relatively simple bound on the mean square error of $\hat{f}_n$

$$\frac{1}{n}\sum_{i=1}^{n}E\left[\left(\hat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right)\right)^2\right] \leq \underset{f\in\mathcal{F}}{min}\{\frac{2}{n}\sum_{i=1}^{n}\left(f\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right)\right)^2 + \frac{8\sigma^2 c\left(f\right)log2}{n}\}. \tag{16.24}$$

The first term in the brackets above is related to the error incurred by approximating f^* by an element of $\mathcal{F}$. The second term is related to the estimation error involved with the model selection process.

Let's focus on the approximation error. First, suppose $f^* \in H^\alpha\left(C_\alpha\right)$ for $1 < \alpha \leq 2$. Let f_k^* be the "best" piecewise linear approximation to f^*, with k pieces on intervals $\left[0, \frac{1}{k}\right), \left[\frac{1}{k}, \frac{2}{k}\right), ... \left[\frac{k-1}{k}, 1\right)$. Consider the difference between f^* and f_k^* on one such interval, say $\left[\frac{i-1}{k}, \frac{i}{k}\right)$. By applying Taylor's theorem with remainder we have

$$f^*\left(t\right) = f^*\left(\frac{i}{k}\right) + \frac{\partial f^*}{\partial x}\left(t^{'}\right)\left(t - \frac{i}{k}\right) \tag{16.25}$$

for $t \in \left[\frac{i-1}{k}, \frac{i}{k}\right)$ and some $t^{'} \in \left[t, \frac{i}{k}\right]$. Define

$$f_k^*\left(t\right) \equiv f^*\left(\frac{i}{k}\right) + \frac{\partial f^*}{\partial x}\left(\frac{i}{k}\right)\left(t - \frac{i}{k}\right). \tag{16.26}$$

Note that $f_k^*(t)$ is not necessarily the best piecewise linear approximation to f^*, just good enough for our purposes. Then using the fact that $f^* \in H^\alpha(C_\alpha)$, for $t \in [i-1/k, i/k)$ we have

$$\begin{aligned} |f^*(t) - f_k^*(t)| &= \left| \frac{\partial f^*}{\partial x}(t') \left(t - \tfrac{i}{k}\right) - \frac{\partial f^*}{\partial x}\left(\tfrac{i}{k}\right)\left(t - \tfrac{i}{k}\right) \right| \\ &\leq \tfrac{1}{k} \left| \frac{\partial f^*}{\partial x}(t') - \frac{\partial f^*}{\partial x}\left(\tfrac{i}{k}\right) \right| \\ &\leq \tfrac{1}{k} C_\alpha \left| t' - \tfrac{i}{k} \right|^{\alpha-1} \\ &\leq \tfrac{1}{k} C_\alpha \left(\tfrac{1}{k}\right)^{\alpha-1} = C_\alpha k^{-\alpha}. \end{aligned} \tag{16.27}$$

So, for all $t \in [0,1]$

$$|f^*(t) - f_k^*(t)| \leq C_\alpha k^{-\alpha}. \tag{16.28}$$

Now let f_k be the element of $\mathcal{F}_k$ closest to f_k^* (f_k is the quantized version of f_k^*)

$$\begin{aligned} |f^*(t) - f_k(t)| &= |f^*(t) - f_k^*(t) + f_k^*(t) - f_k(t)| \\ &\leq |f^*(t) - f_k^*(t)| + |f_k^*(t) - f_k(t)| \\ &\leq C_\alpha k^{-\alpha} + \tfrac{1}{\sqrt{n}} \end{aligned} \tag{16.29}$$

since we used $\frac{1}{2} logn$ bits to quantize the endpoints of each line segment. Consequently,

$$\begin{aligned} |f^*(t) - f_k^*(t)|^2 &\leq |f^*(t) - f_k^*(t)|^2 + 2|f^*(t) - f_k^*(t)||f_k^*(t) - f_k(t)| + |f_k^*(t) - f_k(t)|^2 \\ &\leq C_\alpha^2 k^{-2\alpha} + 2C_\alpha \tfrac{k^{-\alpha}}{\sqrt{n}} + \tfrac{1}{n}. \end{aligned} \tag{16.30}$$

Thus it follows that

$$\min_{f \in \mathcal{F}_k} \{ \frac{2}{n} \sum_{i=1}^{n} (f(i/n) - f^*(i/n))^2 + \frac{8\sigma^2 c(f) log2}{n} \} \leq 2C_\alpha^2 k^{-2\alpha} + \frac{4C_\alpha k^{-\alpha}}{\sqrt{n}} + \frac{2}{n} + \frac{8\sigma^2 k(logn+1) log2}{n}. \tag{16.31}$$

The first and last terms dominate the above expression. Therefore, the upper bound is minimized when $k^{-2\alpha}$ and $\frac{k}{n}$ are balanced. This is accomplished by choosing $k = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$. Then it follows that

$$\min_{f \in \mathcal{F}_k} \{ \frac{2}{n} \sum_{i=1}^{n} \left(f\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right) \right)^2 + \frac{8\sigma^2 c(f) log2}{n} \} = O\left(n^{-\frac{2\alpha}{2\alpha+1}} logn \right). \tag{16.32}$$

If $\alpha = 2$ then we have

$$\frac{1}{n} \sum_{i=1}^{n} E\left[\left(\hat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right) \right)^2 \right] = O\left(n^{-\frac{4}{5}} logn \right). \tag{16.33}$$

If $f^* \in H^\alpha(C_\alpha)$ for $0 < \alpha \leq 1$, let f_k^* be the following piecewise constant approximation to f^*. Let

$$f_k^*(t) \equiv f^*\left(\frac{i}{n}\right) \text{ on interval } \left[\frac{i-1}{k}, \frac{i}{k} \right). \tag{16.34}$$

Then

$$\begin{aligned} |f^*(t) - f_k^*(t)| &= \left| f^*(t) - f^*\left(\tfrac{i}{n}\right) \right| \\ &\leq C_\alpha \left| t - \tfrac{i}{n} \right|^\alpha \\ &\leq C_\alpha k^{-\alpha}. \end{aligned} \tag{16.35}$$

Repeating the same reasoning as in the $1 < \alpha \leq 2$ case, we arrive at

$$\frac{1}{n}\sum_{i=1}^{n} E\left[\left(\hat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right)\right)^2\right] = O\left(n^{-\frac{2\alpha}{2\alpha+1}} log n\right) \tag{16.36}$$

for $0 < \alpha \leq 1$. In particular, for $\alpha = 1$ we get

$$\frac{1}{n}\sum_{i=1}^{n} E\left[\left(\hat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right)\right)^2\right] = O\left(n^{-\frac{2}{3}} log n\right) \tag{16.37}$$

within a logarithmic factor of the rate we had before (in Lecture 4 (Chapter 5)) for that case!

16.4 Summary

1. $\hat{f}_n$ can be computed by finding least-square line fits to the data on partitions of the form $\left[0, \frac{1}{k}\right), \left[\frac{1}{k}, \frac{2}{k}\right), ... \left[\frac{k-1}{k}, 1\right)$ for $k = 1, ..., n$, and then selecting the best fit by the $\hat{k}$ that gives the minimum of the complexity regularization criterion.
2. If $f^* \in H^\alpha(C_\alpha)$ for some $0 < \alpha \leq 2$, then

$$MSE\left(\hat{f}_n\right) = \frac{1}{n}\sum_{i=1}^{n} E\left[\left(\hat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right)\right)^2\right] = O\left(n^{-\frac{2\alpha}{2\alpha+1}} log n\right). \tag{16.38}$$

3. $\hat{f}_n$ automatically picks the optimal number of bins. Essentially $\hat{f}_n$ (indirectly) estimates the smoothness of f^* and produces a rate which is near minimax optimal ! ($n^{-\frac{2\alpha}{2\alpha+1}}$ is the best possible).
4. The larger α is the faster the convergence and the better the denoising !

Chapter 17

Nonlinear Approximation and Wavelet Analysis[1]

17.1 Review

In Lecture 4 (Chapter 5) and 15 (Chapter 16), we investigated the problem of denoising a smooth signal in additive white noise. In Lecture 4 (Chapter 5), we considered Lipschitz functions and showed that by filling constants on a uniform partition of width $n^{-1/3}$ we can achieve an $n^{-2/3}$ rate of MSE convergence.

In Lecture 15 (Chapter 16), we considered Holder-α smooth functions, and we demonstrated that by automatically selecting partition width and using polynomial fits we can obtain a MSE convergence rate of $n^{-2\alpha/2\alpha+1}$, substantially better when $\alpha > 1$. Also important is the fact that we don't need to know the value of α a priori. The estimator $\hat{f}_n$ is fundamentally different than its counterpart in Lecture 4 (Chapter 5).

In both cases $\hat{f}_n(t)$ is a linear function (polynomial on constant fit) of the data in each interval of the underlying partition. In Lecture 4 (Chapter 5), the partition was independent of the data, and so the overall estimator is a linear function of the data .

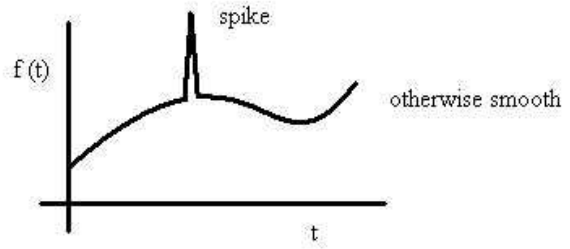
However, in Lecture 15 (Chapter 16) the partition itself was selected based on the data. Consequently, $\hat{f}_n(t)$ is a non-linear function of the data . Linear estimators (linear functions of the data) cannot adapt to unknown degrees of smoothness. In this lecture, we lay the groundwork for one more important extension in the denoising application - spatial adaptivity. That is, we would like to construct estimators that not only adapt to unknown degrees of global smoothness, but that also adapt to spatially varying degrees of smoothness.

We will focus on the approximation theoretic aspects of the problem in this lecture, considering tree-based approximations and wavelet expansions. In the next lecture (Chapter 21), we will apply these results to the denoising problem, this will bring us up to date with the current state-of-the-art in denoising and non-parametric estimation.

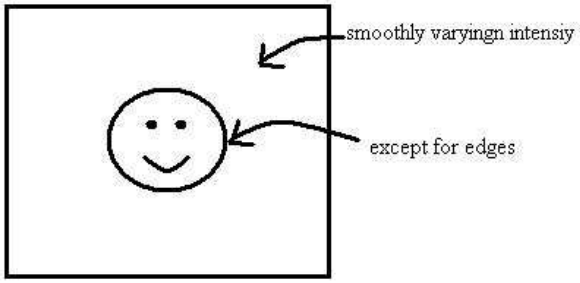
Recall that Holder spaces contain smooth functions that are well approximated with polynomials or piecewise polynomial functions. Holder spaces are quite large and contain many interesting signals. However, Holder spaces are still inadequate in many applications. Often, we encounter functions that are not smooth everywhere; they contain discontinuities, jumps, spikes, etc. Indeed, the "singularities" (or non-smooth points) can be the most interesting and informative aspects of the functions.

Example 17.1
Functions not smooth everywhere.

[1] This content is available online at <http://cnx.org/content/m16278/1.3/>.



(a)

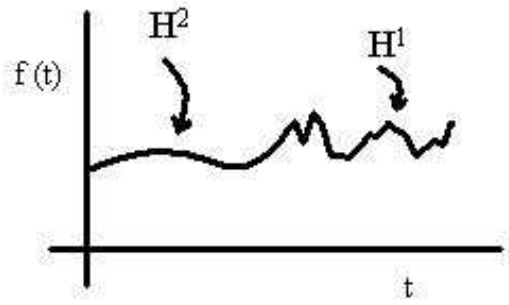


(b)

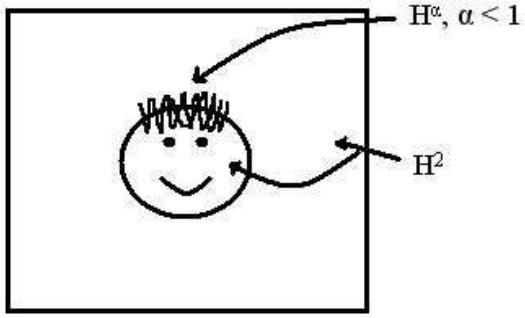
Figure 17.1: Example of functions not smooth everywhere. (a) 1-D Case (b) 2-D Case

Furthermore, functions of interest may possess different degrees of smoothness in different regions.

Example 17.2
Functions with different degrees of smoothness.



(a)



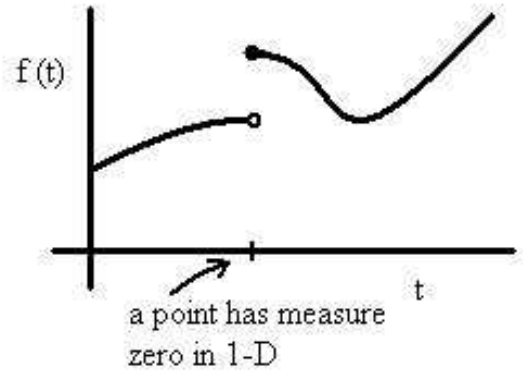
(b)

Figure 17.2: Example of functions having different degrees of smoothness. (a) 1-D Case (b) 2-D Case

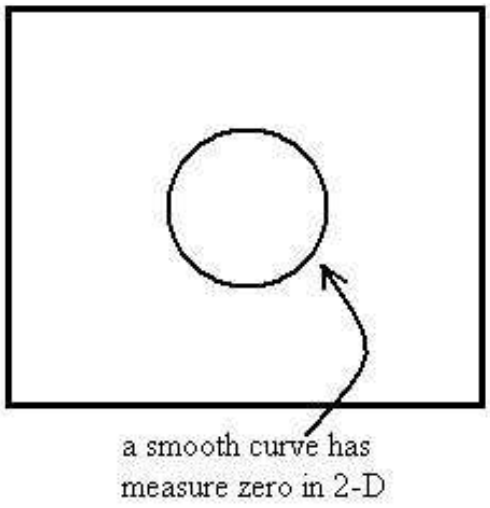
17.2 NonLinear Approximation via Trees

Let $B^{\alpha}(C_{\alpha})$ denote the set of all functions that are $H^{\alpha}(C_{\alpha})$ everywhere except on a set of measure zero. To simplify the notation, we won't explicitly identify the domain (e.g., $[0,1]$ or $[0,1]^d$); that will be clear from the context.

Example 17.3: Sets of measure zero



(a)



(b)

Figure 17.3: Sets of measure zero. (a) 1-D Case (b) 2-D Case

Let's consider a 1-D case first.

Let $f \in B^{\alpha}(C_{\alpha})$ and consider approximating f by a piecewise polynomial function on a uniform partition.

If f is Holder-α smooth everywhere, then by using an appropriate partition width k^{-1} and fitting degree $\lceil \alpha \rceil$ polynomials on each interval we have an approximation f_k satisfying

$$|f(t) - f_k(t)| \leq C_{\alpha} k^{-\alpha} \tag{17.1}$$

and

$$||f - f_k||_{L_2}^2 = O\left(k^{-2\alpha}\right). \tag{17.2}$$

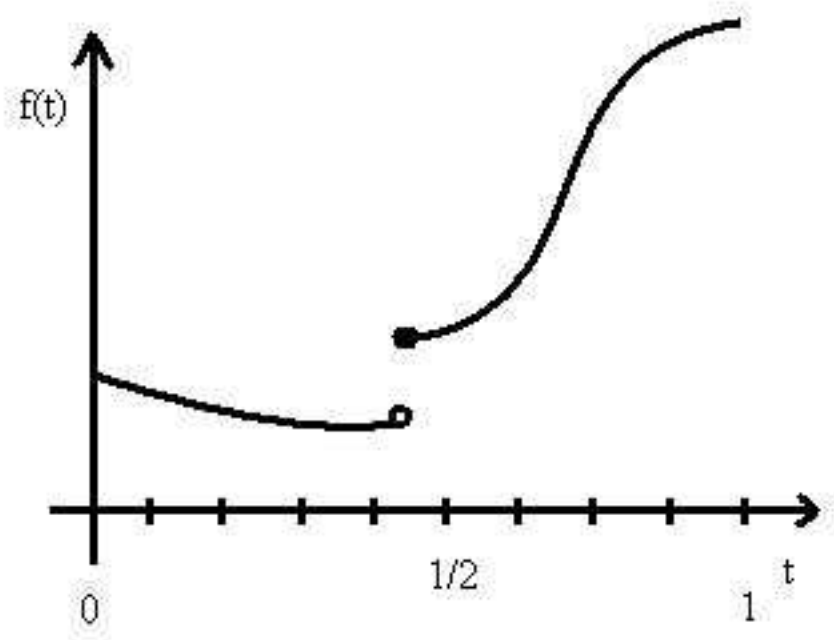


Figure 17.4: Smooth curve with a discontinuity.

However, if there is a discontinuity then for t in the interval containing the discontinuity the difference

$$|f(t) - f_k(t)| \tag{17.3}$$

will not be small.

Example 17.4
Suppose f is piecewise Lipschitz and f_k ia a piecewise constant.

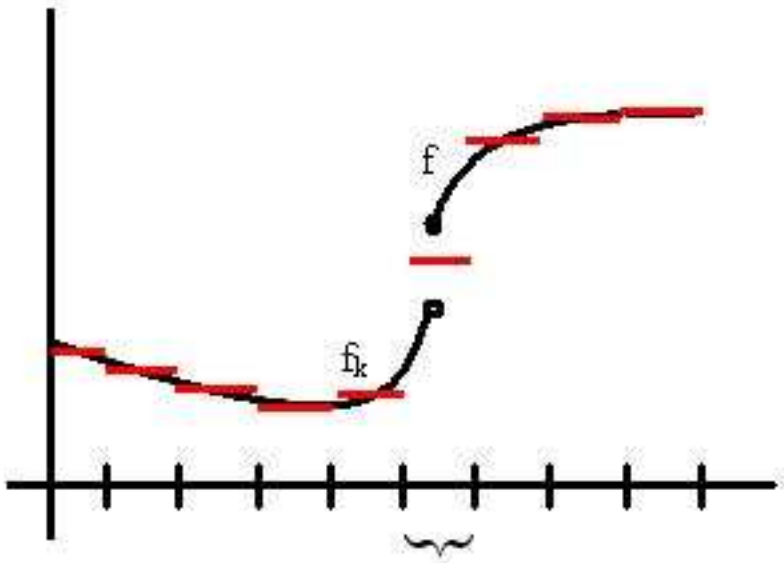


Figure 17.5

$$|f(t) - f_k(t)| \approx \Delta \tag{17.4}$$

where Δ is a constant equal to average of f on right and left side of discontinuity in this interval.

$$\Rightarrow ||f - f_k||_{L_2}^2 = O\left(k^{-1}\right) \tag{17.5}$$

where k^{-1} is the width of the interval. Notice this rate is quite slow.

This problem naturally suggests the following remedy: use very small intervals near discontinuities and larger intervals in smooth regions. Specifically, suppose we use intervals of width $k^{-2\alpha}$ to contain the discontinuities and the intervals of width k^{-1} elsewhere. Then accordingly piecewise polynomial approximation $\tilde{f}_k$ satisfies

$$||f - \tilde{f}_k||_{L_2}^2 = O\left(k^{-2\alpha}\right). \tag{17.6}$$

We can accomplish this need for "adaptive resolution" or "multiresolution" using recursive partitions and trees.

17.3 Recursive Dyadic Partitions

We discussed this idea already in our examination of classification trees. Here is the basic idea again, graphically.

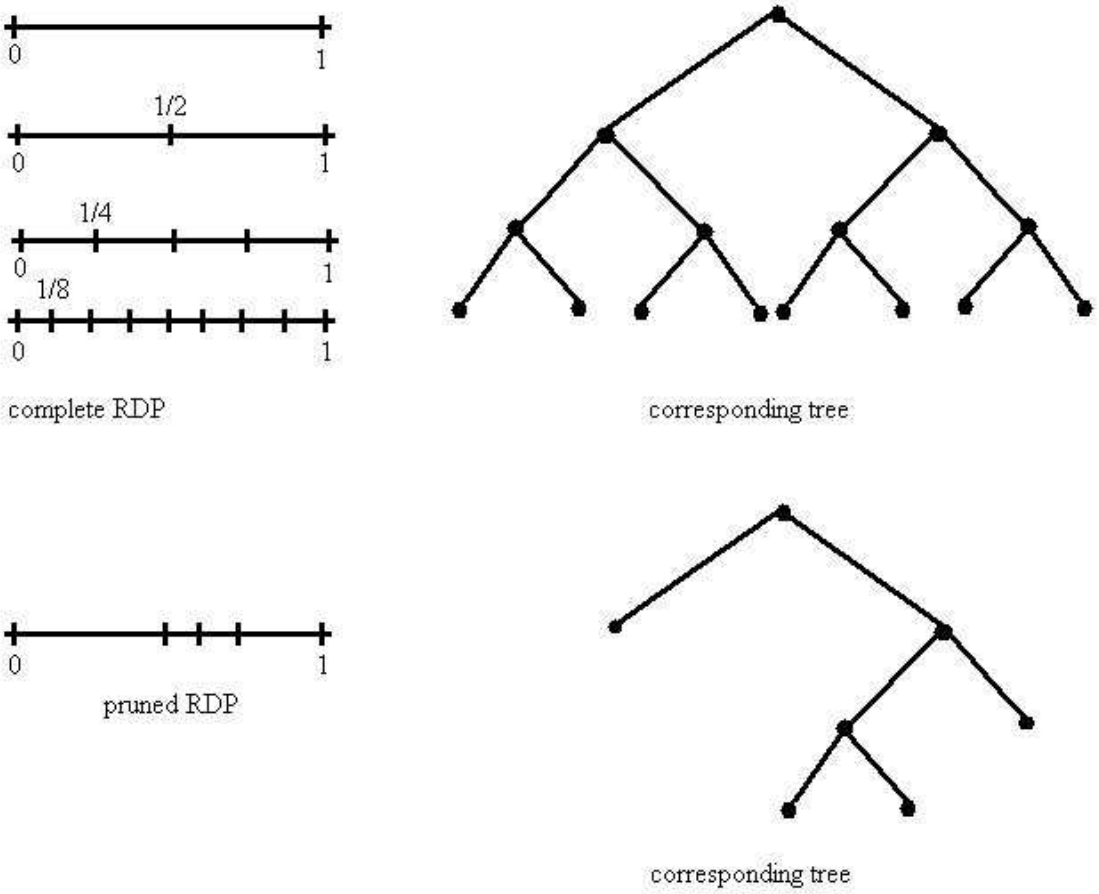


Figure 17.6: Complete and pruned RDP along with their correspnding tree structures.

Consider a function $f \in B^{\alpha}(C_{\alpha})$ that contains no more than m points of discontinuity, and is $H^{\alpha}(C_{\alpha})$ away from these points.

Lemma 17.1:

Consider a complete RDP with n intervals, then there exists an associated pruned RDP with $O(klogn)$ intervals, such that an associated piecewise degree $\lceil \alpha \rceil$ polynomial approximation $(\tilde{f})_k$, has a squared approximation error of $O\left(min\left(k^{-2\alpha}, n^{-1}\right)\right)$.

Proof:

Assume $n > k > m$. Divide $[0,1]$ into k intervals. If f is smooth on a particular interval I, then

$$|f(t) - \tilde{f}_k(t)| = O\left(k^{-2\alpha}\right) \forall t \in I. \tag{17.7}$$

In intervals that contain a discontinuity, recursively subdivide into two until the discontinuity is

contained in an interval of width n^{-1}. This process results in at most $log_2 n$ addition subintervals per discontinuity, and the squared approximation error is $O(k-2\alpha)$ on all of them accept the m intervals of width n^{-1} containing the discontinuities where the error is $O(1)$ at each point.

Thus, the overall squared L_2 norm is

$$||f-\tilde{f}_k||_{L_2}^2 = O\left(min\left(k^{-2\alpha}, n^{-1}\right)\right) \tag{17.8}$$

and there are at most $k + log_2 n$ intervals in the partition. Since k>m, we can upperbound the number of intervals by $2klog_2 n$.

Note that if the initial complete RDP has $n \approx k^{2\alpha}$ intervals, then the squared error is $O\left(k^{-2\alpha}\right)$.

Thus, we only incur a factor of $2\alpha logk$ additional leafs and achieve the same overall approximation error as in the $H^\alpha(C_\alpha)$ case. We will see that this is a small price to pay in order to handle not only smooth functions, but also piecewise smooth functions.

17.4 Wavelet Approximations

Let $f \in L^2([0,1]); \int f^2(t)\,dt < \infty$.

A wavelet approximation is a series of the form

$$f = c_o + \sum_{j\geq 0}\sum_{k=1}^{2^j} < f, \psi_{j,k} > \psi_{j,k} \tag{17.9}$$

where c_o is a constant $\left(c_o = \int_0^1 f(t)\,dt\right)$,

$$< f, \psi_{j,k} > = \int_0^1 f(t)\,\psi_{j,k}(t)\,dt \tag{17.10}$$

and the basis functions $\psi_{j,k}$ are orthonormal, oscillatory signals, each with an associated scale 2^{-j} and position $k2^{-j}$. $\psi_{j,k}$ is called the wavelet at scale 2^{-j} and position $k2^{-j}$.

Example 17.5: Haar Wavelets

$$\psi_{j,k}(t) = 2^{j/2}\left(\mathbf{1}_{\{t\in[2^{-j}(k-1),2^{-j}(k-1/2)]\}} - \mathbf{1}_{\{t\in[2^{-j}(k-1/2),2^{-j}k]\}}\right) \tag{17.11}$$

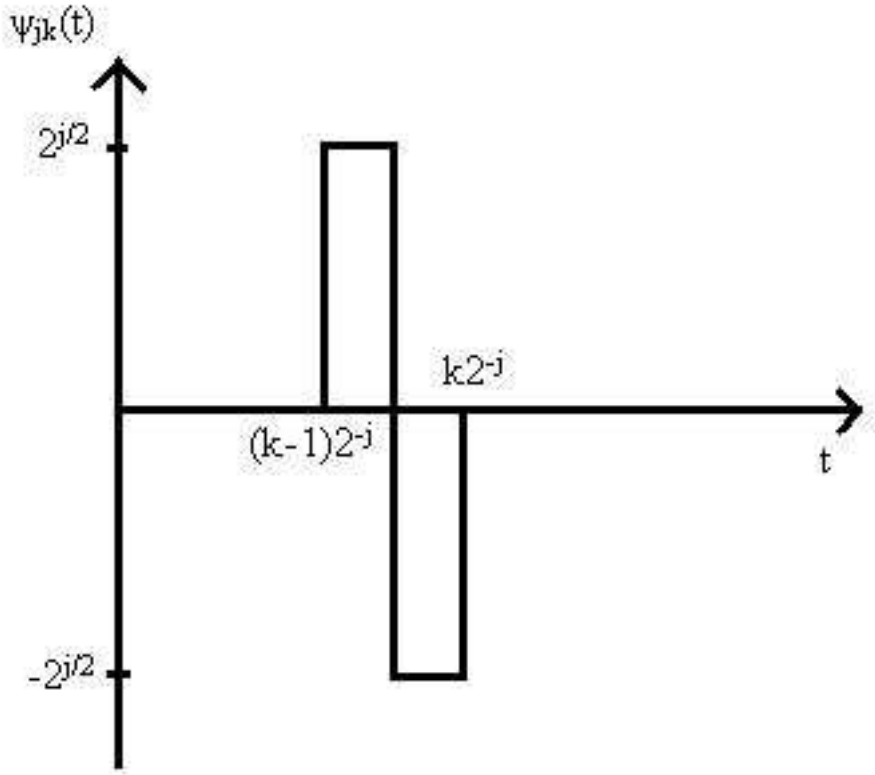


Figure 17.7: Haar Wavelet

$$\int_0^1 \psi_{j,k}(t)\,dt = 0 \tag{17.12}$$

$$\int_0^1 \psi_{j,k}^2(t)\,dt = \int_{(k-1)2^{-j}}^{k2^{-j}} 2^j dt = 1 \tag{17.13}$$

$$\int_0^1 \psi_{j,k}(t)\,\psi_{l,m}(t)\,dt = \delta_{j,l}.\delta_{k,m} \tag{17.14}$$

NOTE: If f is constant on $\left[2^{-j}(k-1), 2^{-j}k\right]$, then

$$\int f\psi_{j,k}(t) = 0. \tag{17.15}$$

Suppose f is piecewise constant with at most m discontinuities. Let

$$f_J = c_o + \sum_{j=0}^{J-1}\sum_{k=1}^{2^j} < f, \psi_{j,k} > \psi_{j,k}. \tag{17.16}$$

Then, f_J has at most mJ non-zero wavelet coefficients; i.e., $< f, \psi_{j,k} > = 0$ for all but mJ terms, since at most one Haar Wavelet at each scale senses each point of discontinuity. Said another way, all but at most m of the wavelets at each scale have support over constant regions of f.

f_J itself will be piecewise constant with discontinuities only possible occurring at end points of the intervals $\left[2^{-J}(k-1), 2^{-J}k\right]$. Therefore, in this case

$$||f - f_J||_{L_2}^2 = O\left(2^{-J}\right). \tag{17.17}$$

Daubechies wavelets are the extension of the Haar wavelet idea. Haar wavelets have one "vanishing moment":

$$\int_0^1 \psi_{j,k} = 0. \tag{17.18}$$

Daubechies wavelets are "smoother" basis functions with extra vanishing moments. The Daubechies-N wavelet has N vanishing moments.

$$\int_0^1 t^l \psi_{j,k} dt = 0 for l = 0, 1, ..., N-1. \tag{17.19}$$

The Daubechies-1 wavelet is just the Haar case.

If f is a piecewise degree $\leq N$ polynomial with at most m pieces, then using the Daubechies-N wavelet system.

$$||f - f_J||_{L_2}^2 = O\left(2^{-J}\right); \tag{17.20}$$

and

$$f_J(t) = c_o + \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} < f, \psi_{j,k} > \psi_{j,k}(t) \tag{17.21}$$

has at most $O(mJ)$ non-zero wavelet coefficients. f_J is called the Discrete Wavelet Transform (DWT) approximation of f. The key idea is the same as we saw with trees.

17.5 Sampled Data

We can also use DWT's to analyze and represent discrete, sampled functions. Suppose,

$$\underline{f} = [f(1/n), f(2/n), ..., f(n/n)] \tag{17.22}$$

then we can write $\underline{f}$ as

$$\underline{f} = c_o + \sum_{j=0}^{log_2 n - 1} \sum_{k=1}^{2^j} < \underline{f}, \underline{\psi}_{j,k} > \underline{\psi}_{j,k} \tag{17.23}$$

where

$$\underline{\psi}_{j,k} = [\psi_{j,k}(1), \psi_{j,k}(2), ..., \psi_{j,k}(n)] \tag{17.24}$$

is a discrete time analog of the continuous time wavelets we considered before. In particular,

$$\sum_{i=1}^{n} i^l \psi_{j,k}(i) = 0, l = 0, 1, ..., N-1 \tag{17.25}$$

for the Daubechies-N discrete wavelets.

$$< \underline{f}, \underline{\psi}_{j,k} > = \underline{f}^T \underline{\psi}_{j,k} \tag{17.26}$$

Thus, we also have an analogous approximation result: If $\underline{f}$ are samples from a piecewise degree $\leq N$ polynomial function with a finite number m of discontinuities, then $\underline{f}$ has $O(mJ)$ non-zero wavelet coefficients.

17.6 Approximating functions with wavelets

Suppose $f \in B^{\alpha}(C_{\alpha})$ and has a finite number of discontinuities. Let f_p denote piecewise degree-N ($N = \lceil \alpha \rceil$) polynomial approximation to f with $O(k)$ pieces; a uniform partition into k equal length intervals followed by addition splits at the points of discontinuity.

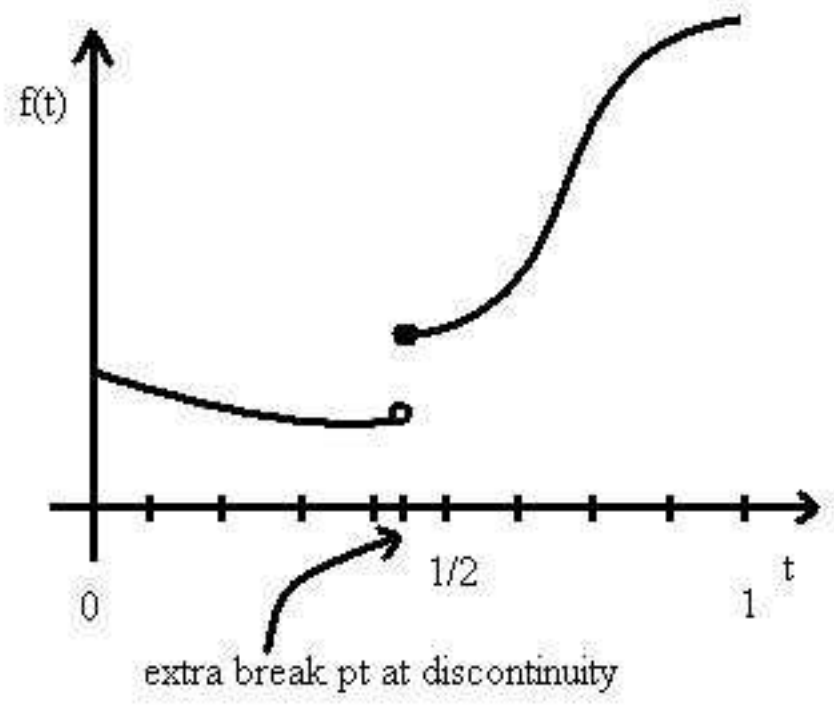


Figure 17.8

Then

$$|f(t) - f_p(t)|^2 = O\left(k^{(-2\alpha)}\right) \forall t \in [0,1] \tag{17.27}$$

$$\Rightarrow |f(i/n) - f_p(i/n)|^2 = O\left(k^{-2\alpha}\right) i = 1, ..., n \tag{17.28}$$

$$\Rightarrow 1/n ||\underline{f} - \underline{f}_p||^2_{L_2} = O\left(k^{-2\alpha}\right) \tag{17.29}$$

and $\underline{f}_p$ has $O(klog_2 n)$ non-zero coefficients according to our previous analysis.

17.7 Wavelets in 2-D

Suppose f is a 2-D image that is piecewise polynomial:

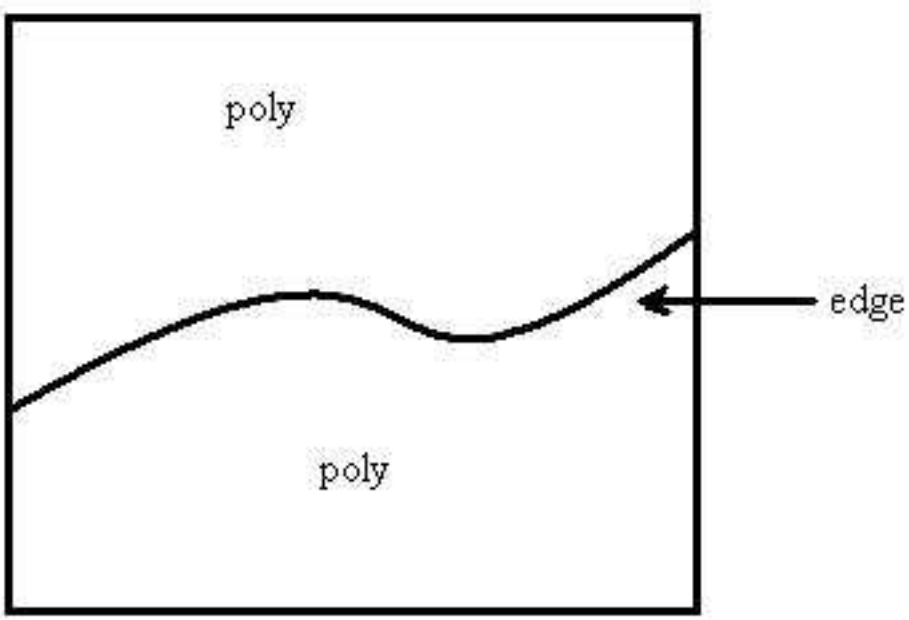


Figure 17.9

A pruned RDP of k squares decorated with polyfits gives

$$||f - f_k||_{L_2}^2 = O\left(k^{-1}\right). \tag{17.30}$$

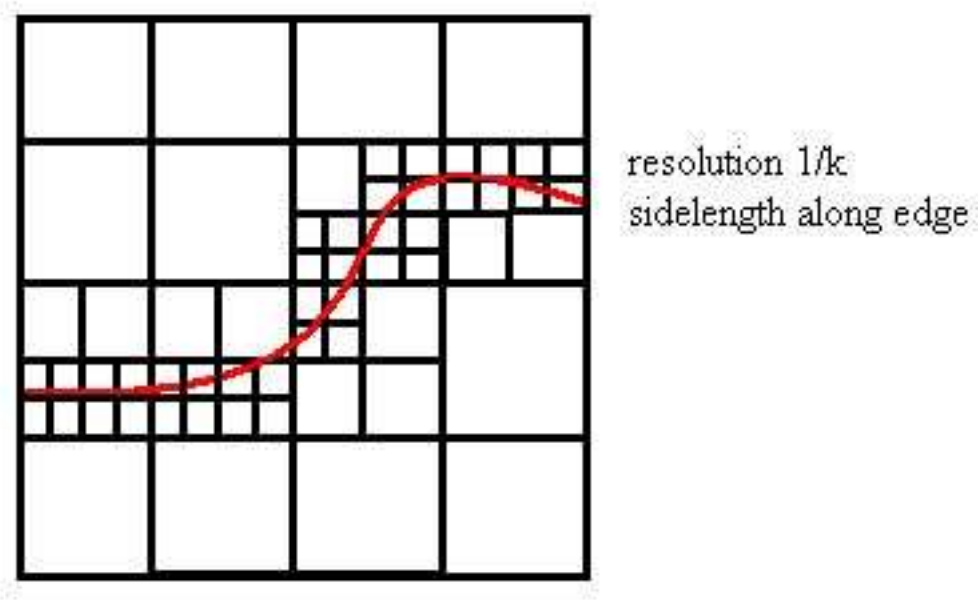


Figure 17.10

Let $\underline{f} = \left[f\left(i/k, j/k\right)_{i,j=1}^{n} \right.$ sample range.

$$f_n\left(t\right) = \sum_{i,j=1}^{k} f\left(i/k, j/kk\right) \mathbf{1}_{\{t \in [i-1/k, i/k) x [j-1/k, j/k)\}} \tag{17.31}$$

then

$$||f - f_n||_{L_2}^2 = O\left(k^{-1}\right) \tag{17.32}$$

$O\left(1\right)$ error on k of the k^2 pixels, near zero elsewhere. The DWT of $\underline{f}$ has $O\left(k\right)$ non-zero wavelet coefficients. $O\left(2^j\right)$ at scale $2^{-j}, j = 0, 1, ..., logn.$

Chapter 18

Vapnik-Chervonenkis Theory[1]

18.1 Review of Past Lecture

In our past lectures we considered collections of candidate function $\mathcal{F}$ that were either **finite** or **enumerable**. We then constructed penalties, usually codelengths, for each candidate $c(f)$, $f \in \mathcal{F}$, such that $\sum_{f\in\mathcal{F}} 2^{c(f)} \leq 1$ This allowed us to derive uniform concentration inequalities over the entire set $\mathcal{F}$ using the union bound. However, in many cases the collections $\mathcal{F}$ may be uncountably infinite. A simple example is the collection $\mathcal{F}$ of a single threshold classifier in 1-d having the form

$$f_t(x) = \mathbf{1}_{\{x\geq t\}} \tag{18.1}$$

and their complements

$$f_s(t) = \mathbf{1}_{\{x<s\}}. \tag{18.2}$$

Thus, $\mathcal{F}$ contains an uncountable number of classifiers, and we cannot apply the union bound argument in such cases.

18.2 Two Ways to Proceed

18.2.1 Discretize or Quantize the Collection

Example 18.1
To quantize $\mathcal{F}$

$$F_q = \{f, f(x) = \mathbf{1}_{\{x\leq 1/q i i\in\{0,1,\ldots,q\}\}}\} \tag{18.3}$$

q is positive, such that $\forall f_q \in \mathcal{F}_q$

$$\int |f - f_q| \leq c/q \tag{18.4}$$

if the density of x is bounded by $c > 0$. $q < n^{1/2}$.

[1]This content is available online at <http://cnx.org/content/m16284/1.2/>.

18.2.2 Identical Empirical Errors

Consider the fact that given only n training data, many of the classifiers in such a collection may produce identical empirical errors. Also, many $f \in \mathcal{F}$ will produce identical label assignments on the data. We will have at most 2^n unique labels.

f is uncountable, its interceptions are countable and bounded by 2^n. n intervals with 2 classifier per interval.

The number of distinct labeling assignments that a class $\mathcal{F}$ can produce on a set of n points is denoted

$$S(\mathcal{F}, n) \leq 2^n \tag{18.5}$$

The VC dimension is $logS(\mathcal{F}, n)$. Specifically, $VC(\mathcal{F}) = k$, where k is largest integer such that $S(\mathcal{F}, k) = 2^k$ Ex. $2n = 2^n$, $n = 2$, $VC(\mathcal{F}) = 2$.

Ex. Consider

$$\mathcal{F} = \{f : f(x) = \mathbf{1}_{\{x \geq t\}} or f(x) = \mathbf{1}_{\{x < t\}}, t \in [0,1]\} \tag{18.6}$$

Let q be a positive integer and

$$\mathcal{F}_q = \{f : f(x) = \mathbf{1}_{\{x \geq i/q\}} or f(x) = \mathbf{1}_{\{x < i/q\}}, i \in \{0,1,...,q\}\} \tag{18.7}$$

and,

$$|f_q| = 2(q+1). \tag{18.8}$$

Moreover, for any $f \in \mathcal{F}$ there exists an $f_1 \in \mathcal{F}_q$ such that

$$\int |f(x) - f_q(x)| dx \leq \int_{(i-1)/q}^{i/q} 1 dx = 1/q. \tag{18.9}$$

Now suppose we have n training data and suppose $f^* \in \mathcal{F}$. We know that in general, the minimum empirical risk classifier will converge to the Bayes classifier at the rate of $n^{-1/2}$ or slower. Therefore, it is unnecessary to drive the approximation error down faster than $n^{-1/2}$ So, we can restrict our attention of $\mathcal{F}_{n^{-1/2}}$ and, provided that the density of x is bound above. We have

$$min_{f \in \mathcal{F}_{n^{-1/2}}} R(f) - R(f^*) \leq C_{f_q} min \int |f^*(x) - f(x)| dx \leq c/n^{1/2}. \tag{18.10}$$

Vapnik-Chervonenkis theory is based not on explicitly quantizing the collection of candidate functions, but rather on recognizing that the richness of $\mathcal{F}$ is limited in a certain sense by the number of training data. Indeed, given n i.i.d. training data, there are at most 2^n different binary labelings. Therefore, any collection $\mathcal{F}$ may be divided into 2^n subsets of classifiers that are "equilvalent" with respect to the training data. In many cases a collection may not even be capable of producing 2^n different labellings.

18.3 Example

Consider $X = [0,1]$.

$$\mathcal{F} = \{f : f(x) = \mathbf{1}_{\{x \geq t\}} or f(x) = \mathbf{1}_{\{x < t\}} t \in [0,1]\} \tag{18.11}$$

Suppose we have n training data: $(x_1, ..., x_n) \in [0,1]$. With x^s denotes the location of each training point in [0,1]. Associated with each x is a label $y \in \{0,1\}$. Any classifier in $\mathcal{F}$ will label all points to the left of a number $t \in [0,1]$ as "1" or "0", and points to the right as "0" or "1", respectively. For $t \in [0, x_1)$, all points are either labelled "0" or "1". For $t \in (x_1, x_2)$, x_1 is labelled "0" or "1" and $x_2 ... x_n$ are label "1" or "0" and so on. We see that there are exactly $2n$ different labellings; far less than 2^n!

The number of different labellings that a class $\mathcal{F}$ can produce on a set of n training data is a measure of the "effective size" of $\mathcal{F}$. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{F}$ is proportional to the log of the effective size. Let $V(\mathcal{F}, n)$ denote the VC dimension of $\mathcal{F}$, typically a constant, independent of n. The VC inequality states that for all $f \in \mathcal{F}$

$$P\left(|\hat{R}_n(f) - R(f)| > \varepsilon\right) \leq 8e^{V(\mathcal{F},h)}e^{-n\varepsilon^2/32}. \tag{18.12}$$

This type of uniform concentration inequality can be used in a similar fashion to our use of Hoeffding's inequality plus union bound.

18.4 Hyperplane Classifiers

We will go into the details of VC Theory next lecture (Chapter 18), and the remainder of this lecture will introduce the key ideas with an example Consider the following setup. Let $X = [0,1]^d$, $Y = \{0,1\}$ Let

$$\mathcal{F} = \{f : f(x) = \mathbf{1}_{\{w^T x + w_0 > 0\}}\} \tag{18.13}$$

with w_0 and $w \in R^{d+1}$ This is the collection of all hyperplane classifiers. $\mathcal{F}$ is infinite and uncountable.

Suppose that we have n training data

$$\{X_i, Y_i\}_{i=1}^n. \tag{18.14}$$

There are at most $2\binom{n}{d}$ unique classifiers in $\mathcal{F}$ with respect to these data. To see this, consider d arbitrary data points $x_1, ..., x_{i_d}$, and let $w^T x + w_0 > 0$ be a hyperplane containing these points. To be specific, take the hyperplane with

$$||w_0 w|| = 1. \tag{18.15}$$

this hyperplane coincides with two possible classification rules:

$$f_1(x) = \mathbf{1}_{\{w^T x + w_0 > 0\}} \tag{18.16}$$

$$f_2(x) = \mathbf{1}_{\{w^T x + w_0 < 0\}} \tag{18.17}$$

Each d-tuple of training data produces two distinct classifiers, assuming the data are not co-linear. Thus, there are at most $2 * \binom{n}{d}$ unique classifiers in $\mathcal{F}$ with respect to the training data. (All other $f \in \mathcal{F}$ produce the same labels and empirical risk as one of the classifiers.) Let's enumerate the unique hyperplane classifiers $f_1, ..., f_{2*\binom{n}{d}}$, and let

$$\hat{f}_n = arg \min_{f \in \{f_1, ..., f_2\binom{n}{d}\}} \hat{R}_n(f) \tag{18.18}$$

and let

$$R^* = inf_{f \in \mathcal{F}} R(f) \tag{18.19}$$

and define

$$f^* = argmin_{f \in \mathcal{F}} R(f) \tag{18.20}$$

If multiple $f \in \mathcal{F}$ achieve R^*, pick f^* to be one of them in an arbitrary fixed number.

Theorem 18.1:
Assume that P_x has a density, but that the distribution of (x, y) is other arbitrary. If $n \geq d$ and $2d/n \leq \varepsilon \leq 1$ then

$$P\left(R\left(\hat{f}_n\right) - R(f) > \varepsilon\right) \leq e^{2d\varepsilon}\left(2\binom{n}{d} + 1\right)e^{-n\varepsilon^2/2}. \tag{18.21}$$

NOTE: The assumption that P_x has a density insures that no d+1 points are co-planar. This in turn, guarantees that there are exactly $2\binom{n}{d}$ unique classifier and that the $2\binom{n}{d}$ under consideration are fully representative of all possible classifiers in $\mathcal{F}$, with respect to the data.

Proof:
The proof is a specialization of the basic ingredients of VC Theory to the case at hand. Here we follow the proof in DGL '96. First we note that,

$$R\left(\hat{f}_n\right) - R(f^*) = R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right) + \hat{R}_n\left(\hat{f}_n\right) - R(f^*) \tag{18.22}$$

$$\leq R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right) + \hat{R}_n f^* - R(f^*) + d/n \tag{18.23}$$

and since $\hat{R}_n\left(\hat{f}_n\right) \leq \hat{R}_n(f) + d/n$ for any $f \in \mathcal{F}$

$$\leq max_{i=1,\ldots,2\binom{n}{d}}\left(R(f_i) - \left(\hat{R}_n\right)(f_i)\right) + \left(\hat{R}_n\right)(f^*) - R(f^*) + d/n. \tag{18.24}$$

Therefore, by the union bound:

$$P\left(R\left(\hat{f}_n\right) - R(f^*) > \varepsilon\right) \tag{18.25}$$

$$\leq \sum_{i=1}^{2\binom{n}{d}} P\left(R(f_i) - \hat{R}_n(f_i) > \varepsilon/2\right) + P\left(\hat{R}_n(f^*) - R(f^*) + d/n > \varepsilon/2\right). \tag{18.26}$$

We can bound the second term of the above bound using Chernoff's/Hoeffding's inequality:

$$P\left(\hat{R}_n(f^*) - R(f^*) > \varepsilon/2 - d/n\right) \tag{18.27}$$

$$\leq e^{-2n(\varepsilon/2 - d/n)^2} \tag{18.28}$$

$$\leq e^{2d\varepsilon}e^{-n\varepsilon^2/2}. \tag{18.29}$$

Next, let's bound one of the terms in the summation. For example, take

$$P\left(R(f_i) - \hat{R}_n(f_i) > (\varepsilon/2)\right). \tag{18.30}$$

Note that by symmetry all $2\binom{n}{d}$ terms will have identical bounds. Since the bounds are independent of $P_x y$.

Assume that f_1 is determined by the first d data points $x_14, ..., x_d$. By the smoothing property of expectations we can write,

$$P\left(R\left(f_i\right) - \hat{R}_n\left(f\right) > \varepsilon/2\right) = E\left[P\left(R\left(f_i\right) - \hat{R}_n\left(f_i\right) > \varepsilon/2 | x_1, ..., x_d\right)\right]. \tag{18.31}$$

From here, we will bound the conditional probability inside the expectation. Let $\left(X_1^{"}, Y_1^{"}\right), ..., \left(X_d^{"}, Y_d^{"}\right)$ be d additional random samples that are independent and identically distributed as the data $\left(X_1, Y_1\right), ..., \left(X_d, Y_d\right)$. $\{X_i^{"}, Y_i^{"}\}_{i=1}^d$ are often called the "ghost sample" since they are not actually observed. They are a fictious sample leads to a simple bound on the conditional probability. Define if $i \leq d$

$$\left(X_i^{'}, Y_i^{'}\right) = \left(X_i^{"}, Y_i^{"}\right) \tag{18.32}$$

or if $i > d$

$$\left(X_i^{'}, Y_i^{'}\right) = \left(X_i, Y_i\right). \tag{18.33}$$

That is, $\{X_i^{'}, Y_i^{'}\}_{i=1}^d$ agrees with our observed data on i>d, but the first d samples are replaced with the ghost sample. Then,

$$P\left(R\left(f_i\right) - \hat{R}_n\left(f_1\right) > \varepsilon/2 | x_1, ..., x_d\right) \tag{18.34}$$

$$\leq P\left(R\left(f_i\right) - 1/n \sum_{i=d+1}^{n} 1_{f_1(x_i) \neq y_i} > \varepsilon/2 | x_1, ..., x_d\right) \tag{18.35}$$

$$\leq P\left(R\left(f_i\right) - 1/n \sum_{1}^{n} 1_{f_1(x_i) \neq y_i} + d/n > \varepsilon/2 | x_1, ..., x_d\right) \tag{18.36}$$

$$= P\left(R\left(f_i\right) - \left(\hat{R}\right)_n^{'}\left(f_1\right) > t/2 - d/n | x_1, ..., x_d\right) \tag{18.37}$$

where,

$$\hat{R}_n^{'}\left(f_1\right) = 1/n \sum_{i=1}^{n} \mathbf{1}_{\{f_1\left(x_i^{'}\right) \neq y_i^{'}\}}. \tag{18.38}$$

Note that $n\left(\hat{R}\right)_n^{'}\left(f_1\right)$ is binomially distributed with mean $R\left(f_1\right)$ and it is independent of $x_1, ..., x_d$ Therefore,

$$P\left(R\left(f_i\right) - \hat{R}_n^{'}\left(f_1\right) > \varepsilon/2 - d/n | x_1, ..., x_d\right) \tag{18.39}$$

$$= P\left(R\left(f_i\right) - \hat{R}_n^{'}\left(f_1\right) > t/2 - d/n | x_1, ..., x_d\right) \tag{18.40}$$

$$\leq e^{-2n(\varepsilon/2 - d/n)^2} \tag{18.41}$$

$$\leq e^{2d\varepsilon} e^{-n\varepsilon^2/2}. \tag{18.42}$$

In conclusion,

$$P\left(R\left(\hat{f}_n\right) - R^* > \varepsilon\right) \tag{18.43}$$

$$\leq \sum_{i=1}^{2\binom{n}{d}} P\left(R(f)_i - \hat{R}_n\left(t_i\right) > \varepsilon/2\right) + P\left(\hat{R}_n\left(f^*\right) - R\left(f^*\right) + d/n > \varepsilon/2\right) \tag{18.44}$$

$$\leq 2\binom{n}{d} e^{2d\varepsilon} e^{-n\varepsilon^2/2} + e^{2d\varepsilon} e^{-n\varepsilon^2/2} \tag{18.45}$$

$$= e^{2d\varepsilon}\left(2\binom{n}{d} + 1\right) e^{-n\varepsilon^2/2}. \tag{18.46}$$

Lastly, Corollary If $n \geq d$, then

$$E\left[R\left(\hat{f}_n\right) - min_{f\in\mathcal{F}} R\left(f\right)\right] \leq \sqrt{2\left(d+1\right)\left(logn + 2\right)/n}. \tag{18.47}$$

Chapter 19

The Vapnik-Chervonenkis Inequality[1]

19.1 The Vapnik-Chervonenkis Inequality

The VC inequality is a powerful generalization of the bounds we obtained for the hyperplane classifier in the previous lecture (Chapter 21). The basic idea of the proof is quite similar. Before starting the inequality, we need to introduce the concept of shatter coefficients and VC dimension .

19.2 Shatter Coefficients

Let $\mathcal{A}$ be a collection of subsets of $\mathcal{R}^d$, definition : The n^{th} shatter coefficient of $\mathcal{A}$ is defined by

$$\mathcal{S}_{\mathcal{A}}(n) = \max_{x_1, ..., x_n \varepsilon \mathcal{R}^d} \left| \{ \{x_1, ..., x_n\} \bigcap A, A \varepsilon \mathcal{A} \} \right|. \tag{19.1}$$

The shatter coefficients are a measure of the richness of the collection $\mathcal{A}$. $\mathcal{S}_{\mathcal{A}}(n)$ is the largest number of different subsets of a set of n points that can be generated by intersecting the set with elements of $\mathcal{A}$.

Example 19.1
In 1-d, Let $\mathcal{A} = \{(-\infty, t], t \varepsilon \mathcal{R}\}$ Possible subsets of $\{x_1, ..., x_n\}$ generated by intersecting with sets of the form $(-\infty, t]$ are $\{x_1, ..., x_n\}$, $\{x_1, ..., x_{n-1}\}$, ..., $\{x_1\}$, ϕ. Hence $\mathcal{S}_d(n) = n + 1$.

Example 19.2
In 2-d, Let $\mathcal{A}$ = { all rectangles in $\mathcal{R}^2$}

Consider a set $\{x_1, x_2, x_3, x_4\}$ of training points. If we arrange the four points into the corner of a diamond shape. It's easy to see that we can find a rectangle in $\mathcal{R}^2$ to cover any subsets of the four points as the above picture, i.e. $\mathcal{S}_{\mathcal{A}}(4) = 2^4 = 16$.

Clearly, $\mathcal{S}_{\mathcal{A}}(n) = 2^n, n = 1, 2, 3$ as well.

However, for $n = 5, \mathcal{S}_{\mathcal{A}}(n) < 2^5$. This is because we can always select four points such that the rectangle, which just contains four of them, contains the other point. Consequently, we cannot find a rectangle classifier which contains the four outer points and does not contain the inner point as shown above.

Note the $\mathcal{S}_{\mathcal{A}} \leq 2^n$.

If $|\{\{x_1, ..., x_n\} \bigcap A, A \varepsilon \mathcal{A}\}| = 2^n$ then we say that $\mathcal{A}$ shatters x_1, ..., x_n.

[1] This content is available online at <http://cnx.org/content/m16283/1.2/>.

19.3 VC Dimension

Definition 19.1: The VC dimension
$V_{\mathcal{A}}$ of a collection of sets $\mathcal{A}$ is defined as the largest interger n such that $S_{\mathcal{A}}(n) = 2^n$.
Example
$\mathcal{A} = \{(-\infty, t]\ ; t \varepsilon \mathcal{R}\}, \mathcal{S}_{\mathcal{A}} = n+1$ hence $\mathcal{V}_{\mathcal{A}} = 1$.

Example
$\mathcal{A} = \{$ all rectangles in $\mathcal{R}^2\}$.
$\mathcal{S}_{\mathcal{A}} = 2^n, n = 1, 2, 3, 4$ and $\mathcal{S}_{\mathcal{A}} \leq 2^n, n = 4$, Hence $V_{\mathcal{A}} = 4$.
The VC dimension provides a useful bound on the growth of the shatter coefficients.

19.4 Sauer's Lemma:

Let $\mathcal{A}$ be a collection of set with VC dimension $V_{\mathcal{A}} < \infty$. Then $\forall n, \mathcal{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \begin{pmatrix} n \\ i \end{pmatrix}$, also $\mathcal{S}_{\mathcal{A}}(n) \leq (n+1)^{V_{\mathcal{A}}}, \forall n$.

19.5 VC Dimension and Classifiers

Let $\mathcal{F}$ be a collection of classifiers of the form $f : \mathcal{R}^d \to \{0, 1\}$ Define $\mathcal{A} = \{\{x : f(x) = 1\} \times \{0\} \bigcup \{x : f(x) = 0\} \times \{1\}, f \varepsilon \mathcal{F}\}$ In words, this is collection of subsets of $\mathcal{X} \times \mathcal{Y}$ for which on $f \varepsilon \mathcal{F}$ maps the features x to a label opposite of y. The size of $\mathcal{A}$ expresses the richness of $\mathcal{F}$. The larger $\mathcal{A}$ is the more likely it is that there exists an $f \varepsilon \mathcal{F}$ for which $R(f) = P(f(X) \neq Y)$ is close to the Bayes risk $R^* = P(f^*(X) \neq Y)$ where f^* is the Bayes classifier. The n^{th} shatter coefficient of $\mathcal{F}$ is defined as $\mathcal{S}_{\mathcal{F}}(n) = \mathcal{S}_{\mathcal{A}}(n)$ and the VC dimesion of $\mathcal{F}$ is defined as $V_{\mathcal{F}} = V_{\mathcal{A}}$.

Example 19.3
linear (hyperplane) classifiers in $\mathcal{R}^d$

Consider $d = 2$. Let n be the number of training points, it is easy to see that when $n = 1$, let $\mathcal{A}$ be as above. By using linear classifiers in $\mathcal{R}^2$, it is easy to see that we can assign 1 to all possible subsets $\{\{x_1\}, \phi\}$ and 0 to their complements. Hence $\mathcal{S}_{\mathcal{F}}(1) = 2$.

When $n = 2$, we can also assign 1 to all possible subsets $\{\{x_1, x_2\}$, $\{x_1\}$, $\{x_2\}$, $\phi\}$ and 0 to their complements, and vice versa. Hence $\mathcal{S}_{\mathcal{F}}(2) = 4 = 2^2$.

When $n = 3$, we can arrange arrange the point x_1, x_2, x_3(non-colinear) so that the set of linear classifiers shatters the three points, hence $\mathcal{S}_{\mathcal{F}}(3) = 8 = 2^3$

When $n = 4$, no matter where the points x_1, x_2, x_3, x_4 and what designated binary values y_1, y_2, y_3, y_4 are. It's clear that $\mathcal{A}$ does not shatter the four points. To see the claim, first observe that the four points will form a 4-gon (if the four points are co-linear, or if the three points are co-linear then clearly linear classifiers cannot shatter the points). The two points that belong to the same diagonal lines form 2 groups and no linear classifier can assign different values to the 2 groups. Hence $\mathcal{S}_{\mathcal{F}}(4) < 16 = 2^4$ and $V_{\mathcal{F}} = 3$.

We state here without proving it that in general the class of linear classifiers in $\mathcal{R}^d$ has $V_{\mathcal{F}} = d+1$.

19.6 The VC Inequality

Let $X_1, , ..., X_n$ be i.i.d. $\mathcal{R}^d$-valued random variables. Denote the common distribution of $X_i, 1 \leq i \leq n$ by $\mu(A) = P(X_1 \varepsilon A)$ for any subset $A \subset \mathcal{R}^d$. Similarly, define the empirical distribution $\mu_n(A) = \frac{1}{n}\sum_1^n 1_{\{X_i \varepsilon A\}}$.

Theorem 19.1: VC '71

For any probablilty measure μ and collection of subsets $\mathcal{A}$, and for any $\varepsilon > 0$.

$$P\left(\begin{array}{c} sup \\ A\varepsilon\mathcal{A} \end{array} |\mu_n(A) - \mu(A)| > \varepsilon \right) \leq 8\mathcal{S}_{\mathcal{A}}(n) e^{-n\varepsilon^2/32} \tag{19.2}$$

and

$$E\left[\begin{array}{c} sup \\ A\varepsilon\mathcal{A} \end{array} |\mu_n(A) - \mu(A)| \right] \leq 2\sqrt{\frac{log2\mathcal{S}_{\mathcal{A}}(n)}{n}} \tag{19.3}$$

Before giving a proof to the theorem. We present a Corollary.

Corollary 19.1:

Let $\mathcal{F}$ be a collection of classifiers of the form$f : \mathcal{R}^d \to \{0, 1\}$ with VC dimension $V_{\mathcal{F}} < \infty$, Let $R(f) = P(f(X) \neq Y)$ and $\hat{R}_n(f) = \frac{1}{n}\sum_1^n 1_{\{f(X_i) \neq Y_i\}}$, where $X_i, Y_i, 1 \leq i \leq n$ are i.i.d. with joint distributionP_{XY}.

Define

$\hat{f}_n = \begin{array}{c} argmin \\ f\varepsilon\mathcal{F} \end{array} \hat{R}_n(f)$.

Then

$$E\left[R\left(\hat{f}_n\right)\right] - \begin{array}{c} inf \\ f\varepsilon\mathcal{F} \end{array} R(f) \leq 4\sqrt{\frac{\mathcal{V}_{\mathcal{F}} log(n+1) + log2}{n}}. \tag{19.4}$$

Proof:

Let $\mathcal{A} = \{\{x : f(x) = 1\} \times \{0\} \bigcup \{x : f(x) = 0\} \times \{1\}, f \varepsilon \mathcal{F}\}$

Note that

$$P(f(X) \neq Y) = P((X, Y) \varepsilon A) := \mu(A) \tag{19.5}$$

where $A = \{x : f(x) = 1\} \times \{0\} \bigcup \{x : f(x) = 0\} \times \{1\}$.

Similarly,

$$\frac{1}{n}\sum_1^n 1_{\{f(X_i) \neq Y_i\}} = \frac{1}{n}\sum_1^n 1_{\{(X_i, Y_i) \varepsilon A\}} := \mu(A). \tag{19.6}$$

Therefore, according to the VC theorem.

$$\begin{aligned} E\left[\begin{array}{c} sup \\ f \varepsilon \mathcal{F} \end{array} \left|\hat{R}_n(f) - R(f)\right| \right] = E\left[\begin{array}{c} sup \\ A \varepsilon \mathcal{A} \end{array} |\mu_n(A) - \mu(A)| \right] &\leq 2\sqrt{\frac{log2\mathcal{S}_{\mathcal{A}}(n)}{n}} \\ &= 2\sqrt{\frac{log2\mathcal{S}_{\mathcal{F}}(n)}{n}} \end{aligned} \tag{19.7}$$

Since $V_{\mathcal{F}} < \infty, \mathcal{S}_{\mathcal{F}}(n) \leq (n+1)^{V_{\mathcal{F}}}$ and

$$E\left[\sup_{f \varepsilon \mathcal{F}} \left|\hat{R}_n(f) - R(f)\right|\right] \leq 2\sqrt{\frac{V_{\mathcal{F}} log(n+1) + log 2}{n}}. \tag{19.8}$$

Next, note that

$$\begin{aligned} R\left(\hat{f}_n\right) - \inf_{f \varepsilon \mathcal{F}} R(f) &= \left[R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right)\right] + \left[\hat{R}_n\left(\hat{f}_n\right) - \inf_{f \varepsilon \mathcal{F}} R(f)\right] \\ &= \left[R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right)\right] + \left[\sup_{f \varepsilon \mathcal{F}} \left(\hat{R}_n\left(\hat{f}_n\right) - R(f)\right)\right] \\ &\leq \left[R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right)\right] + \left[\sup_{f \varepsilon \mathcal{F}} \left(\hat{R}_n(f) - R(f)\right)\right] \\ &\leq 2 \sup_{f \varepsilon \mathcal{F}} \left|\hat{R}_n(f) - R(f)\right| \end{aligned} \tag{19.9}$$

Therefore,

$$\begin{aligned} E\left[R\left(\hat{f}_n\right)\right] - \inf_{f \varepsilon \mathcal{F}} R(f) &\leq 2E\left[\sup_{f \varepsilon \mathcal{F}} \left|\hat{R}_n(f) - R(f)\right|\right] \\ &\leq 4\sqrt{\frac{V_{\mathcal{F}} log(n+1) + log 2}{n}} \end{aligned} \tag{19.10}$$

Chapter 20

Applications of VC Bound[1]

20.1 Linear Classifiers

Suppose $\mathcal{F}$= {linear classifiers in $\mathbf{R}^d$}, then we have

$$V_{\mathcal{F}} = d+1, \quad \hat{f}_n = \underset{f\in\mathcal{F}}{argmin}\hat{R}_n(f) \tag{20.1}$$

$$E\left[R\left(\hat{f}_n\right)\right] - \underset{f\in\mathcal{F}}{inf}R(f) \leq 4\sqrt{\frac{(d+1)\,log\,(n+1)+log2}{n}}. \tag{20.2}$$

20.2 Generalized Linear Classifiers

Normally, we have a feature vector $X \in \mathbf{R}^d$. A hyperplane in $\mathbf{R}^d$ provides a linear classifier in $\mathbf{R}^d$. Nonlinear classifiers can be obtained by a straightforward generalization.

Let $\phi_1, \cdots, \phi_{d'}$, $d' \geq d$ be a collection of functions mapping $\mathbf{R}^d \to \mathbf{R}$. These functions, applied to a feature $X \in \mathbf{R}^d$, produce a generalized set of features, $\phi = (\phi_1(X), \phi_2(X), \cdots, \phi_{d'}(X))'$. For example, if $X = (x_1, x_2)'$, then we could consider $d' = S$ and $\phi = \left(x_1, x_2, x_1x_2, x_1^2, x_2^2\right)' \in \mathbf{R}^5$. We can then construct a linear classifier in the higher dimensional generalized feature space $\mathbf{R}^{d'}$.

The VC bounds immediately extend to this case, and we have for $\mathcal{F}'$ = { generalized linear classifiers based on maps $\phi : \mathbf{R}^d \to \mathbf{R}^{d'}$ },

$$E\left[R\left(\hat{f}_n\right)\right] - \underset{f\in\mathcal{F}'}{inf}R(f) \leq 4\sqrt{\frac{(d'+1)\,log\,(n+1)+log2}{n}}. \tag{20.3}$$

20.3 Half-Space Classifiers

Theorem 20.1: Steele '75, Dudley '78

Let $\mathcal{G}$be a finite-dimensional vector space of real-valued functions on $\mathbf{R}^d$. The class of sets $\mathcal{A} = \{\{x : g(x) \geq 0\} : g \in \mathcal{G}\}$ has VC dimension $\geq \dim(\mathcal{G})$.

[1]This content is available online at <http://cnx.org/content/m16262/1.2/>.

Proof:
It is sufficient to show that no set of $n = dim(\mathcal{G}) + 1$ points can be shattered by $\mathcal{A}$. Take any n points and for each $g \in \mathcal{G}$, define the vector $V_g = (g(x_1), \cdots, g(x_n))$.

The set $\{V_g : g \in \mathcal{G}\}$ is a linear subspace of $\mathbf{R}^n$ of dimension $\leq$ dim $(\mathcal{G}) = n - 1$. Therefore, there exists a non-zero vector $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbf{R}^n$ such that $\sum_{i=1}^{n} \alpha_i g(x_i) = 0$. We can assume that at least one of these α_i^S is negative (if all are positive, just negate the sum). We can then re-arrange this expression as $\sum_{i:\alpha_i \geq 0} \alpha_i g(x_i) = \sum_{i:\alpha_i < 0} -\alpha_i g(x_i)$.

Now suppose that there exists a $g \in \mathcal{G}$ such that the set $\{x : g(x) \geq 0\}$ selects precisely the x_i^S on the left-hand side above. Then all terms on the left are non-negative and all the terms on the right are non-positive. Since α is non-zero, this is a contradiction. Therefore, $x_1, \cdots, x_n$ cannot be shattered by sets in $\{x : g(x) \geq 0\}$, $g \in \mathcal{G}$. 6.375pt0.0pt6.375pt

Example
Consider half-spaces in $\mathbf{R}^d$ of the form $\mathcal{A} = \{x \in \mathbf{R}^d : x_i \geq b, i \in \{1, \cdots, d\}, b \in R\}$. Each half-space can be described by

$$g(x) = [0, \cdots, 0, 1, 0, \cdots, 0] \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} - b \tag{20.4}$$

$$\Rightarrow dim(\mathcal{G}) = d + 1, \quad V_{\mathcal{A}} \leq d + 1. \tag{20.5}$$

20.4 Tree Classifiers

Let

$$\mathcal{T}_k = \{recursive \quad rectangular \quad partitions \quad of \quad \mathbf{R}^d \quad with \quad k+1 \quad cells\} \tag{20.6}$$

Let $T \in \mathcal{T}_k$. Each cell of T results from splitting a rectangular region into two smaller rectangles parallel to one of the coordinate axes.

Example 20.1
$T \in \mathcal{T}_3$, $d = 2$.

Each additional split is analogous to a half-space set. Therefore, each additional split can potentially shatter $d + 1$ points. This implies that

$$V_{\mathcal{T}_k} \leq (d + 1)\, k. \tag{20.7}$$

Example 20.2
$d = 1$.

$k = 1$ split shatters two points.

$k = 2$ splits shatters three points < 4.

20.5 VC Bound for Tree Classifiers

$$\mathcal{F}_k = \{tree \quad classifiers \quad with \quad k+1 \quad leafs \quad on \quad \mathbf{R}^d\} \tag{20.8}$$

$$E\left[R\left(\hat{f}_n\right)\right] - \inf_{f\in\mathcal{F}_k} R(f) \leq 4\sqrt{\frac{(d+1)\,klogn + log2}{n}}. \tag{20.9}$$

Exercise 20.1 ***(Solution on p. 143.)***

How can we decide what dimension to choose for a generalized linear classifier?

How many leafs should be used for a classification tree?

20.6 Structural Risk Minimization (SRM)

SRM is simply complexity regularization using VC type bounds in place of Chernoff's bound or other concentration inequalities.

The basic idea is to consider a sequence of sets of classifiers $\mathcal{F}_1, \mathcal{F}_2, ...$, of increasing VC dimensions $V_{\mathcal{F}_1} \leq V_{\mathcal{F}_2} \leq ....$ Then for each $k = 1, 2, ...$ we find the minimum empirical risk classifier

$$\hat{f}_n^{(k)} = arg\min_{f\in\mathcal{F}_k} \hat{R}_n(f) \tag{20.10}$$

and then select the final classifier according to

$$\hat{k} = arg\min_{k\geq 1}\{\hat{R}_n\left(\hat{t}_n^{(k)}\right) + \sqrt{\frac{32V_{\mathcal{F}_k}\,(logn+1)}{n}}\} \tag{20.11}$$

and $\hat{f}_n \equiv \hat{f}_n^{\left(\hat{k}\right)}$ is the final choice.

The basic rational is that we know

$$R_n\left(\hat{f}_n^{(k)}\right) - \inf_{f\in\mathcal{F}_k} R(f) \leq C'\sqrt{\frac{V_{\mathcal{F}_k} logn}{n}} \tag{20.12}$$

where C' is a constant.

The end result is that

$$E\left[R\left(\hat{f}_n\right)\right] \leq \min_{k\geq 1}\{\min_{f\in\mathcal{F}_k} R(f) + 16\sqrt{\frac{V_{\mathcal{F}_k} logn + 4}{2n}}\} \tag{20.13}$$

analogous to our pervious complexity regularization results, except that codelengths are replaced by VC dimensions.

In order to prove the result we use the VC probability concentration bound and assume that $\triangle = \sum_{k\geq 1} V_{\mathcal{F}_k} < \infty$. This enables a union bounding argument and leads to a risk bound of the form given above.

20.7 Key Point of VC Theory

Complexity of classes depends on richness (shattering capability) relative to a set of n arbitrary points. This allows us to effectively "quantize" collections of functions in a slightly data-dependent manner.

20.8 Application to Trees

Let

$$\mathcal{F}_k = \{k \quad leaf \quad decision \quad trees \quad in \quad \mathbf{R}^d\}, \quad V_{\mathcal{F}_k} \leq (d+1)(k+1) \tag{20.14}$$

$$\hat{f}_n^{(k)} = arg \min_{f \in \mathcal{F}_k} \hat{R}_n(f) \tag{20.15}$$

$$\hat{k} = argmin_{k \geq 1} \left(\min_{f \in \mathcal{F}_k} R(f) + \sqrt{\frac{32(d+1)(k-1)(logn+1)}{n}} \right) \tag{20.16}$$

Then

$$\hat{f}_n = \hat{f}_n^{\left(\hat{k}\right)} \tag{20.17}$$

satisfies

$$E\left[R\left(\hat{f}_n\right)\right] \leq \min_{k \geq 1} \left(\min_{f \in \mathcal{F}_k} R(f) + 16\sqrt{\frac{(d+1)(k-1)logn+4}{2n}} \right) \tag{20.18}$$

compare with

$$E\left[R\left(\hat{f}_n\right)\right] \leq \min_{k \geq 1} \left(\min_{f \in dyadic \ k \ leaf \ trees} R(f) + \sqrt{\frac{(3k-1)log2 + \frac{1}{2}logn}{2n}} \right) \tag{20.19}$$

from Lecture 11 (Chapter 12).

Solutions to Exercises in Chapter 20

Solution to Exercise 20.1 (p. 141)
Complexity Regularization using VC bounds!

Chapter 21

Lower Performance Bounds for Estimators[1]

21.1 Lower Performance Bounds

In other modules, estimators/predictors are analyzed, in order to obtain upper bounds on their performance. These bounds are of the form:

$$\underset{f\in\mathcal{F}}{min}\mathbb{E}\left[d\left(\hat{f}_n,f\right)\right]\leq Cn^{-\gamma} \tag{21.1}$$

where $\gamma>0$. We would like to know if these bounds are tight, in the sense that there is no other estimator that is significantly better. To answer this, we need lower bounds like

$$\underset{\hat{f}_n}{inf}\underset{f\in\mathcal{F}}{sup}\mathbb{E}\left[d\left(\hat{f}_n,f\right)\right]\geq cn^{-\gamma} \tag{21.2}$$

We assume we have the following ingredients:

- ***:** Class of models, $\mathcal{F}\subseteq\mathcal{S}$. $\mathcal{F}$ is a class of models containing the "true" model and is a subset of some bigger class $\mathcal{S}$. E.g. $\mathcal{F}$ could be the class of Lipschitz density functions or distributions P_{XY} satisfying the box-counting condition.
- ***:** An observation model, $\mathcal{P}_f$, indexed by $f\in\mathcal{F}$. $\mathcal{P}_f$ denotes the distribution of the data under model f. E.g. in regression and classification, this is the distribution of $Z=(X_1,Y_1,\cdots,X_n,Y_n)\subseteq\mathcal{Z}$. We will assume that $\mathcal{P}_f$ is a probability measure on the measurable space $(\mathcal{Z},\mathcal{B})$.
- ***:** A performance metric $d(.,.)\geq 0$. If you have a model estimate $\hat{f}_n$, then the performance of that model estimate relative to the true model f is $d\left(\hat{f}_n,f\right)$. E.g.

$$\text{Regression:}\qquad d\left(\hat{f}_n,f\right)=||\hat{f}_n-f||_2=\left(\int\left(\hat{f}_n(x)-f(x)\right)^2dx\right)^{1/2} \tag{21.3}$$

$$\text{Classification:}\qquad d\left(\hat{f}_n,f\right)=R\left(\hat{G}_n\right)-R^*=\int_{\hat{G}_n\Delta G^*}|2\eta(x)-1|dP_X(x) \tag{21.4}$$

[1] This content is available online at <http://cnx.org/content/m17357/1.3/>.

As before, we are interested in the risk of a learning rule, in particular the maximal risk given as:

$$\underset{f\in\mathcal{F}}{sup}\mathbb{E}_f\left[d\left(\hat{f}_n,f\right)\right]=\underset{f\in\mathcal{F}}{sup}\int d\left(\hat{f}_n\left(Z\right),f\right)d\mathcal{P}_f\left(Z\right) \tag{21.5}$$

where $\hat{f}_n$ is a function of the observations Z and $\mathbb{E}_f$ denotes the expectation with respect to $\mathcal{P}_f$.

The main goal is to get results of the form

$$\mathcal{R}_n^*\triangleq\underset{\hat{f}_n}{inf}\underset{f\in\mathcal{F}}{sup}\mathbb{E}\left[d\left(\hat{f}_n,f\right)\right]\geq cs_n \tag{21.6}$$

where $c>0$ and $s_n\to 0$ as $n\to\infty$. The inf is taken over all estimators, i.e. all measurable functions $\hat{f}_n:\mathcal{Z}\to\mathcal{S}$.

Suppose we have shown that

$$\underset{n\to\infty}{lim\ inf}\ s_n^{-1}\mathcal{R}_n^*\geq c>0 \qquad \text{(A lower bound)} \tag{21.7}$$

and also that for a particular estimator $\overline{f}_n$

$$\underset{n\to\infty}{lim\ sup}\ s_n^{-1}\underset{f\in\mathcal{F}}{sup}\mathbb{E}_f\left[d\left(\overline{f}_n,f\right)\right]\leq C \tag{21.8}$$

$$\Rightarrow \underset{n\to\infty}{lim\ sup}\ s_n^{-1}\mathcal{R}_n^*\leq C, \tag{21.9}$$

We say that s_n is the optimal rate of convergence for this problem and that $\overline{f}_n$ attains that rate.

NOTE: Two rates of convergence Ψ_n and $\Psi_n^{'}$ are equivalent, i.e. $\Psi_n\equiv\Psi_n^{'}$ iff

$$0<\underset{n\to\infty}{lim\ inf}\frac{\Psi_n}{\Psi_n^{'}}\leq\underset{n\to\infty}{lim\ sup}\frac{\Psi_n}{\Psi_n^{'}}<\infty \tag{21.10}$$

21.1.1 General Reduction Scheme

Instead of directly bounding the expected performance, we are going to prove stronger probability bounds of the form

$$\underset{\hat{f}_n}{inf}\underset{f\in\mathcal{F}}{sup}\mathcal{P}_f\left(d\left(\hat{f}_n,f\right)\geq s_n\right)\geq c>0 \tag{21.11}$$

These bounds can be readily converted to expected performance bounds using Markov's inequality:

$$\mathcal{P}_f\left(d\left(\hat{f}_n,f\right)\geq s_n\right)\leq\frac{\mathbb{E}_f\left[d\left(\hat{f}_n,f\right)\right]}{s_n} \tag{21.12}$$

Therefore it follows:

$$\underset{\hat{f}_n}{inf}\underset{f\in\mathcal{F}}{sup}\mathbb{E}_f\left[d\left(\hat{f}_n,f\right)\right]\geq\underset{\hat{f}_n}{inf}\underset{f\in\mathcal{F}}{sup}s_n\mathcal{P}_f\left(d\left(\hat{f}_n,f\right)\geq s_n\right)\geq cs_n \tag{21.13}$$

21.1.1.1 First Reduction Step

Reduce the original problem to an easier one by replacing the larger class $\mathcal{F}$ with a smaller finite class $\{f_0, \cdots, f_M\} \subseteq \mathcal{F}$. Observe that

$$\underset{\hat{f}_n}{inf} \underset{f \in \mathcal{F}}{sup} \mathcal{P}_f \left(d\left(\hat{f}_n, f \right) \geq s_n \right) \geq \underset{\hat{f}_n}{inf} \underset{f \in \{f_0, \cdots, f_M\}}{sup} \mathcal{P}_f \left(d\left(\hat{f}_n, f \right) \geq s_n \right) \tag{21.14}$$

The key idea is to choose a finite collection of models such that the resulting problem is as hard as the original, otherwise the lower bound will not be tight.

21.1.1.2 Second Reduction Step

Next, we reduce the problem to a hypotheses test. Ideally, we would like to have something like

$$\underset{\hat{f}_n}{inf} \underset{f \in \mathcal{F}}{sup} \mathcal{P}_f \left(d\left(\hat{f}_n, f \right) \geq s_n \right) \geq \underset{\hat{f}_n}{inf} \underset{j \in \{0, \cdots, M\}}{sup} \mathcal{P}_{f_j} \left(\hat{h}_n \left(Z \right) \neq j \right) \tag{21.15}$$

The inf is over all measurable test functions

$$\hat{h}_n : \mathcal{Z} \to \{0, \cdots, M\} \tag{21.16}$$

and $\mathcal{P}_{f_j} \left(\hat{h}_n \left(Z \right) \neq j \right)$ denotes the probability that after observing the data, the test infers the wrong hypothesis.

This might not always be true or easy to show, but in certain scenarios it can be done. Suppose $d\left(.,.\right)$ is a semi-distance, i.e. it satisfies

(i): $d\left(f, g\right) = d\left(g, f\right) \geq 0$ (Symmetric)

(ii):

$$d\left(f, f\right) = 0 \tag{21.17}$$

(iii): $d\left(f, g\right) \leq d\left(h, f\right) + d\left(h, g\right)$ (Triangle inequality)

E.g. with $f, g : \mathbb{R}^d \to \mathbb{R}, d\left(f, g\right) \triangleq ||f - g||_2$.

Lemma 21.1:

Suppose $d\left(.,.\right)$ is a semi-distance. Also suppose that we have constructed $f_0, \cdots, f_M$ s.t. $d\left(f_j, f_k\right) \geq 2s_n, \forall j \neq k$. Take any estimator $\hat{f}_n$ and define the test: $\Psi^* \circ \hat{f}_n : \mathcal{Z} \to \{0, \cdots, M\}$ as

$$\Psi^* \left(\hat{f}_n \right) = \underset{j}{argmin} d\left(\hat{f}_n, f_j \right) \tag{21.18}$$

Then $\Psi^* \left(\hat{f}_n \right) \neq j$, implies $d\left(\hat{f}_n, f_j \right) \geq s_n$.

Suppose $\Psi^* \left(\hat{f}_n \right) \neq j$ [U+27FA] $\exists k \neq j : d\left(\hat{f}_n, f_k \right) \leq d\left(\hat{f}_n, f_j \right)$. Now

$$2s_n \leq d\left(f_j, f_k\right) \leq d\left(\hat{f}_n, f_j \right) + d\left(\hat{f}_n, f_k \right) \leq 2d\left(\hat{f}_n, f_j \right) \tag{21.19}$$

$$\Rightarrow d\left(\hat{f}_n, f_j\right) \geq s_n \tag{21.20}$$

The previous lemma implies that

$$\mathcal{P}_{f_j}\left(d\left(\hat{f}_n, f_j\right) \geq s_n\right) \geq \mathcal{P}_{f_j}\left(\Psi^*\left(\hat{f}_n\right) \neq j\right) \tag{21.21}$$

Therefore,

$$\begin{aligned} \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} \mathcal{P}_{f_j}\left(d\left(\hat{f}_n, f_j\right) \geq s_n\right) &\geq \inf_{\hat{f}_n} \max_{f\in\{f_0,\cdots,f_M\}} \mathcal{P}_{f_j}\left(d\left(\hat{f}_n, f_j\right) \geq s_n\right) \\ &\geq \inf_{\hat{f}_n} \max_{j\in\{0,\cdots,M\}} \mathcal{P}_{f_j}\left(\Psi^*\left(\hat{f}_n\right) \neq j\right) \\ &\geq \inf_{\hat{h}_n} \max_{j\in\{0,\cdots,M\}} \mathcal{P}_j\left(\hat{h}_n \neq j\right) \\ &\triangleq P_{e,M} \end{aligned} \tag{21.22}$$

The third step follows since we are replacing the class of tests defined by $\Psi^*\left(\hat{f}_n\right)$ by a larger class of ALL possible tests $\hat{h}_n$, and hence the inf taken over the larger class is smaller.

Now our goal throughout is going to be to find lower bounds for $P_{e,M}$.

So we need to construct $f_0, \cdots, f_M$ s.t. $d(f_j, f_k) \geq 2s_n$, $j \neq k$ and $P_{e,M} \geq c > 0$. Observe that this requires careful construction since the first condition necessitates that f_j and f_k are far from each other, while the second condition requires that f_j and f_k are close enough so that it is harder to distinguish them based on a given sample of data, and hence the probability of error $P_{e,M}$ is bounded away from 0.

We now try to lower bound the probability of error $P_{e,M}$. We first consider the case $M = 1$, corresponding to binary hypothesis testing.

M = 1: Let P_0 and P_1 denote the two probability measures, i.e. distributions of the data under models 0 and 1. Clearly if P_0 and P_1 are very "close", then it is hard to distinguish the two hypotheses, and so $P_{e,1}$ is large.

A natural measure between probability measures is the total variation , defined as:

$$V(P_0, P_1) = \sup_A |P_0(A) - P_1(A)| = \sup_A \left| \int_A p_0(Z) - p_1(Z)\, d\nu(Z) \right| \tag{21.23}$$

where p_0 and p_1 are the densities of P_0 and P_1 with respect to a common dominating measure ν and A is any subset of the domain. We will lower bound the probability of error $P_{e,1}$ using the total variation distance. But first, we establish the following lemma.

Lemma 21.2: Scheffe's lemma

$$\begin{aligned} V(P_0, P_1) &= \tfrac{1}{2}\int |p_0(Z) - p_1(Z)|\, d\nu(Z) = \tfrac{1}{2}\int |p_0 - p_1| \\ &= 1 - \int \min(p_0, p_1) \end{aligned} \tag{21.24}$$

Recall the definition of the total variation distance:

$$V(P_0, P_1) = \sup_A \left| \int_A p_0 - p_1 \right| \tag{21.25}$$

Observe that the set A maximizing the right hand side is given by either $\{Z \in \mathcal{Z} : p_0(Z) \geq p_1(Z)\}$ or $\{Z \in \mathcal{Z} : p_1(Z) \geq p_0(Z)\}$.

Let us pick $A_0 = \{Z \in \mathcal{Z} : p_0(Z) \geq p_1(Z)\}$. Then

$$V(P_0, P_1) = \int_{A_0} p_0 - p_1 = -\int_{A_0^c} p_0 - p_1 = \frac{1}{2} \int |p_0 - p_1| \tag{21.26}$$

For the second part, notice that

$$p_0(Z) - min(p_0(Z), p_1(Z)) = \begin{cases} 0 & \text{if } p_0(Z) \leq p_1(Z) \\ p_0(Z) - p_1(Z) & \text{if } p_0(Z) \geq p_1(Z) \end{cases} \tag{21.27}$$

Now consider

$$1 - \int min(p_0, p_1) = \int p_0(Z) - min(p_0(Z), p_1(Z)) = \int_{A_0} p_0(Z) - p_1(Z)\, d\nu(Z) = V(P_0, P_1) \tag{21.28}$$

We are now ready to tackle the lower bound on $P_{e,1}$. In this case, we consider all tests $\hat{h}_n(Z) : \mathcal{Z} \to \{0, 1\}$. Equivalently, we can define $\hat{h}_n(Z) = 1_A(Z)$, where A is any subset of the domain.

$$\begin{aligned} P_{e,1} = \inf_{\hat{h}_n} \max_{j \in \{0, \cdots, M\}} \mathcal{P}_j\left(\hat{h}_n \neq j\right) &\geq \inf_{\hat{h}_n} \left(\tfrac{1}{2} P_0\left(\hat{h}_n \neq 0\right) + P_1\left(\hat{h}_n \neq 1\right)\right) \\ &= \tfrac{1}{2} \inf_A P_0(1_A(Z) \neq 0) + P_1(1_A(Z) \neq 1) \\ &= \tfrac{1}{2} \inf_A P_0(A) + P_1(A^c) \\ &= \tfrac{1}{2} \inf_A 1 - (P_1(A) - P_0(A)) \\ &= \tfrac{1}{2}(1 - V(P_0, P_1)) \end{aligned} \tag{21.29}$$

So if P_0 is close to P_1, then $V(P_0, P_1)$ is small and the probability of error $P_{e,1}$ is large.

This is interesting, but unfortunately, it is hard to work with total variation, especially for multivariate distributions. Bounds involving the Kullback-Leibler divergence are much more convenient.

$$K(P_1 || P_0) = \int log \frac{p_1(Z)}{p_0(Z)} p_1(Z)\, d\nu(Z) = \int log \frac{p_1}{p_0} p_1 \tag{21.30}$$

The following Lemma relates total variation, affinity and KL divergence.

Lemma 21.3:

$1 - V(P_0, P_1) \geq \frac{1}{2} A^2(P_0, P_1) \geq \frac{1}{2} exp(-K(P_1 || P_0))$

For the first inequality,

$$\begin{aligned} A^2(P_0,P_1) &= \left(\int\sqrt{p_0p_1}\right)^2 \\ &= \left(\int\sqrt{min(p_0,p_1)\,max(p_0,p_1)}\right)^2 \\ &= \left(\int\sqrt{min(p_0,p_1)}\sqrt{max(p_0,p_1)}\right)^2 \\ &\leq \int min(p_0,p_1)\int max(p_0,p_1) \quad \text{by Cauchy-Schwarz inequality} \\ &= \int min(p_0,p_1)\left(2-\int min(p_0,p_1)\right) \quad \because \int min(p_0,p_1)+\int max(p_0,p_1)=\int p_0+\int p_1=2 \\ &\leq 2\int min(p_0,p_1) \\ &= 2(1-V(P_0,P_1)) \end{aligned} \tag{21.31}$$

For the second inequality,

$$\begin{aligned} A^2(P_0,P_1) &= \left(\int\sqrt{p_0p_1}\right)^2 \\ &= exp\left(log\left(\int\sqrt{p_0p_1}\right)^2\right) \\ &= exp\left(2log\left(\int\sqrt{p_0p_1}\right)\right) \\ &= exp\left(2log\left(\int\sqrt{\tfrac{p_0}{p_1}}p_1\right)\right) \\ &\geq exp\left(2\int log\left(\sqrt{\tfrac{p_0}{p_1}}\right)p_1\right) \quad \text{by Jensen's inequality} \\ &= exp\left(-\int log\left(\sqrt{\tfrac{p_1}{p_0}}\right)p_1\right) \\ &= exp\left(-K(P_1||P_0)\right) \end{aligned} \tag{21.32}$$

Putting everything together, we now have the following Theorem:

Theorem 21.1:
Let $\mathcal{F}$ be a class of models, and suppose we have observations Z distributed according to $\mathcal{P}_f$, $f\in\mathcal{F}$. Let $d\left(\hat{f}_n,f\right)$ be the performance measure of the estimator $\hat{f}_n(Z)$ relative to the true model f. Assume also $d(.,.)$ is a semi-distance. Let $f_0,f_1\in\mathcal{F}$ be s.t. $d(f_0,f_1)\geq 2s_n$. Then

$$\begin{aligned} \inf_{\hat{f}_n}\sup_{f\in\mathcal{F}}\mathcal{P}_f\left(d\left(\hat{f}_n,f\right)\geq s_n\right) &\geq \inf_{\hat{f}_n}\max_{j\in\{0,1\}}\mathcal{P}_{f_j}\left(d\left(\hat{f}_n,f_j\right)\geq s_n\right) \\ &\geq \tfrac{1}{4}exp\left(-K(P_{f_1}||P_{f_0})\right) \end{aligned} \tag{21.33}$$

How do we use this theorem?
Choose f_0,f_1 such that $K(P_1||P_0)\leq\alpha$, then $P_{e,1}$ is bounded away from 0 and we get a bound

$$\inf_{\hat{f}_n}\sup_{f\in\mathcal{F}}\mathcal{P}_f\left(d\left(\hat{f}_n,f\right)\geq s_n\right)\geq c>0 \tag{21.34}$$

or, after Markov's

$$\inf_{\hat{f}_n}\sup_{f\in\mathcal{F}}\mathbb{E}_f\left[d\left(\hat{f}_n,f\right)\right]\geq cs_n \tag{21.35}$$

To apply the theorem, we need to design f_0,f_1 s.t. $d(f_0,f_1)\geq 2s_n$ and $exp\left(-K(P_{f_1}||P_{f_0})\right)>0$. To reiterate, the design of f_0,f_1 requires careful construction so as to balance the tradeoff between

the first condition which requires f_0, f_1 to be far apart, and the second condition which requires f_0, f_1 to be close to each other.

Example

Lets use this theorem in a problem we are familiar with. Let $X \in [0,1]$ and $Y|X = x \sim$ Bernoulli $(\eta(x))$, where $\eta(x) = P(Y = 1|X = x)$.

Suppose $G^* = [t^*, 1]$. We proved that under these assumptions and an upper bound on the density of X, the Chernoff bounding technique yielded an expected error rate for ERM

$$\mathbb{E}\left[R\left(\hat{G}_n\right) - R^*\right] = O\left(\sqrt{\frac{logn}{n}}\right) \tag{21.36}$$

Is this the best possible rate?

Construct two models in the above class (denote it by $\mathcal{P}$), $P_{XY}^{(0)}$ and $P_{XY}^{(1)}$. For both take $P_X \sim$ Uniform $([0,1])$ and $\eta_{(0)} = 1/2 - a$, $\eta_{(1)} = 1/2 + a (a > 0)$, so $G_0^* = \varnothing$, $G_1^* = [0,1]$.

We are interested in controlling the excess risk

$$R\left(\hat{G}_n\right) - R(G^*) = \int_{\hat{G}_n \Delta G^*} |2\eta(x) - 1| dP_X(x) \tag{21.37}$$

Note that if the true underlying model is either $P_{XY}^{(0)}$ or $P_{XY}^{(1)}$, we have:

$$R_j\left(\hat{G}_n\right) - R_j\left(G_j^*\right) = \int_{\hat{G}_n \Delta G_j^*} |2\eta_j(x) - 1| dx = 2a \int_{\hat{G}_n \Delta G_j^*} dx = 2a d_\Delta\left(\hat{G}_n, G_j^*\right) \tag{21.38}$$

Proposition 1

$d_\Delta(.,.)$ is a semi-distance.

It suffices to show that $d(G_1, G_2) = d(G_2, G_1) \geq 0$, $d(G, G) = 0 \forall G$ and $d(G_1, G_2) \leq d(G_1, G_3) + d(G_3, G_2)$. The first two statements are obvious. The last one (triangle inequality) follows from the fact that $G_1 \Delta G_2 \subseteq (G_1 \Delta G_3) \cup (G_3 \Delta G_2)$.

Suppose this was not the case, then $\exists x : x \in G_1 \Delta G_2$ s.t. $x \notin G_1 \Delta G_3$ and $x \notin G_2 \Delta G_3$. In other words,

$$x \in (G_1 \Delta G_2) \cap (G_1 \Delta G_3)^c \cap (G_2 \Delta G_3)^c \tag{21.39}$$

Since $S \Delta T = (S \cap T^c) \cup (S^c \cap T)$, we have:

$$\begin{aligned} x &\in [(G_1 \cap G_2^c) \cup (G_1^c \cap G_2)] \cap [(G_1^c \cup G_3) \cap (G_1 \cup G_3^c)] \cap [(G_2^c \cup G_3) \cap (G_2 \cup G_3^c)] \\ &\in [G_1 \cap (G_1^c \cup G_3) \cap G_2^c \cap (G_2 \cup G_3^c)] \cup [G_1^c \cap (G_1 \cup G_3^c) \cap G_2 \cap (G_2^c \cup G_3)] \\ &\in [G_1 \cap G_3 \cap G_2 \cap G_3^c] \cup [G_1^c \cap G_3^c \cap G_2 \cap G_3] \\ &\in \varnothing, \text{ a contradiction} \end{aligned} \tag{21.40}$$

Lets look at the first reduction step:

$$\begin{aligned} \inf_{\hat{G}_n} \sup_{p \in \mathcal{P}} P\left(R\left(\hat{G}_n\right) - R(G^*) \geq s_n\right) &\geq \inf_{\hat{G}_n} \max_{j \in \{0,1\}} P_j\left(R_j\left(\hat{G}_n\right) - R_j\left(G_j^*\right) \geq s_n\right) \\ &= \inf_{\hat{G}_n} \max_{j \in \{0,1\}} P_j\left(d_\Delta\left(\hat{G}_n, G_j^*\right) \geq s_n/2a\right) \end{aligned} \tag{21.41}$$

So we can work out a bound on d_Δ and then translate it to excess risk.

Lets apply Theorem 1 (Theorem 21.1, p. 150). Note that $d_{\Delta}\left(G_0^*, G_1^*\right) = 1$ and let $P_0 \triangleq P^{(0)}_{X_1,Y_1,\cdots,X_n,Y_n}$ and $P_1 \triangleq P^{(1)}_{X_1,Y_1,\cdots,X_n,Y_n}$.

$$\begin{aligned} K\left(P_1||P_0\right) &= \mathbb{E}_1\left[log\frac{p^{(1)}_{X_1,Y_1,\cdots,X_n,Y_n}(X_1,Y_1,\cdots,X_n,Y_n)}{p^{(0)}_{X_1,Y_1,\cdots,X_n,Y_n}(X_1,Y_1,\cdots,X_n,Y_n)}\right] \\ &= \mathbb{E}_1\left[log\frac{p^{(1)}_{X_1,Y_1}(X_1,Y_1)\cdots p^{(1)}_{X_n,Y_n}(X_n,Y_n)}{P^{(0)}_{X_1,Y_1}(X_1,Y_1)\cdots p^{(0)}_{X_n,Y_n}(X_n,Y_n)}\right] \\ &= \sum_{i=1}^{n}\mathbb{E}_1\left[log\frac{p^{(1)}_{X_i,Y_i}(X_i,Y_i)}{p^{(0)}_{X_i,Y_i}(X_i,Y_i)}\right] \\ &= n\mathbb{E}_1\left[log\frac{p^{(1)}_{Y|X}(Y_1|X_1)}{p^{(0)}_{Y|X}(Y_1|X_1)}\right] \end{aligned} \tag{21.42}$$

Now $p^{(1)}_{Y|X}\left(Y_1 = 1|X_1\right) = 1/2 + a$ and $p^{(0)}_{Y|X}\left(Y_1 = 1|X_1\right) = 1/2 - a$. Also under model 1, $Y_1 \sim$ Bernoulli $(1/2 + a)$. So we get:

$$\begin{aligned} K\left(P_1||P_0\right) &= n\left[(1/2+a)\,log\tfrac{1/2+a}{1/2-a} + (1/2-a)\,log\tfrac{1/2-a}{1/2+a}\right] \\ &= n\left[2alog\left(1/2+a\right) - 2alog\left(1/2-a\right)\right] \\ &= 2nalog\tfrac{1/2+a}{1/2-a} \\ &\leq 2na\left(\tfrac{1/2+a}{1/2-a} - 1\right) \\ &= 4na^2\tfrac{1}{1/2-a} \end{aligned} \tag{21.43}$$

Let $a = 1/\sqrt{n}$ and $n \geq 16$, then $K\left(P_1||P_0\right) \leq 4n\frac{1}{n}\frac{1}{1/2-1/\sqrt{n}} \leq 16$.

Using Theorem 1 (Theorem 21.1, p. 150), since $d_{\Delta}\left(G_0^*, G_1^*\right) = 1$, we get:

$$\inf_{\hat{G}_n}\max_j P_j\left(d_{\Delta}\left(\hat{G}_n, G_j^*\right) \geq 1/2\right) \geq \frac{1}{4}e^{-16} \tag{21.44}$$

Taking $s_n = 1/\sqrt{n}$, this implies

$$\inf_{\hat{G}_n}\sup_{p\in\mathcal{P}} P\left(R\left(\hat{G}_n\right) - R\left(G^*\right) \geq 1/\sqrt{n}\right) \geq \frac{1}{4}e^{-16} \tag{21.45}$$

or, after Markov's inequality

$$\inf_{\hat{G}_n}\sup_{p\in\mathcal{P}}\mathbb{E}\left[R\left(\hat{G}_n\right) - R\left(G^*\right)\right] \geq \frac{1}{4}e^{-16}\frac{1}{\sqrt{n}} \tag{21.46}$$

Therefore, apart from the *logn* factor, ERM is getting the best possible performance.

Reducing the initial problem to a binary hypothesis testing does not always work. Sometimes we need M hypotheses, with $M \to \infty$ as $n \to \infty$. If this is the case, we have the following theorem:

Theorem 2 Let $M \geq 2$. $\{f_0, \cdots, f_M\} \in \mathcal{F}$ be such that

.: $d\left(f_j, f_k\right) \geq 2s_n$, where d is a semi-distance.

.: $\frac{1}{M}\sum_{j=1}^{M} K\left(P_j||P_0\right) \leq \alpha logM$, with $0 < \alpha < 1/8$.

Then

$$\begin{aligned} \inf_{\hat{f}_n}\sup_{f\in\mathcal{F}} P_f\left(d\left(\hat{f}_n, f\right) \geq s_n\right) &\geq \inf_{\hat{f}_n}\max_j P_j\left(d\left(\hat{f}_n, f_j\right) \geq s_n\right) \\ &\geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - 2\sqrt{\frac{\alpha}{logM}}\right) > 0 \end{aligned} \tag{21.47}$$

We will use this theorem to show that the estimator of Lecture 4 (Chapter 5) is optimal. Recall the setup of Lecture 4 (Chapter 5). Let

$$\mathcal{F} = \{f : |f(t) - f(s)| \leq L|t-s| \forall t, s\} \tag{21.48}$$

i.e. the class of Lipschitz functions with constant L. Let

$$x_i = i/n, \quad i = 1, \cdots, n \tag{21.49}$$

$$Y_i = f(x_i) + W_i \tag{21.50}$$

$\mathbb{E}[W_i] = 0, \mathbb{E}\left[W_i^2\right] = \sigma^2 < \infty, W_i, W_j$ are indepedent if $i \neq j$. In that lecture, we constructed an estimator $\hat{f}_n$ such that

$$\underset{f \in \mathcal{F}}{sup} \mathbb{E}\left[||\hat{f}_n - f||^2\right] = O\left(n^{-2/3}\right) \tag{21.51}$$

Is this the best we can do?

We are going to construct a collection $f_0, \cdots, f_M \in \mathcal{F}$ and apply Theorem 2. Notice that the metric of interest is $d\left(\hat{f}_n, f\right) = ||\hat{f}_n - f||$, a semi-distance. Let $W_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$. Let $m \in \mathbb{N}$, $h = 1/m$ and define

$$K(x) = \left(\frac{Lh}{2} - L|x|\right) \mathbb{I}_{|x| \leq h/2} = \frac{L}{2}|h - 2x| \mathbb{I}_{|x| \leq h/2} \tag{21.52}$$

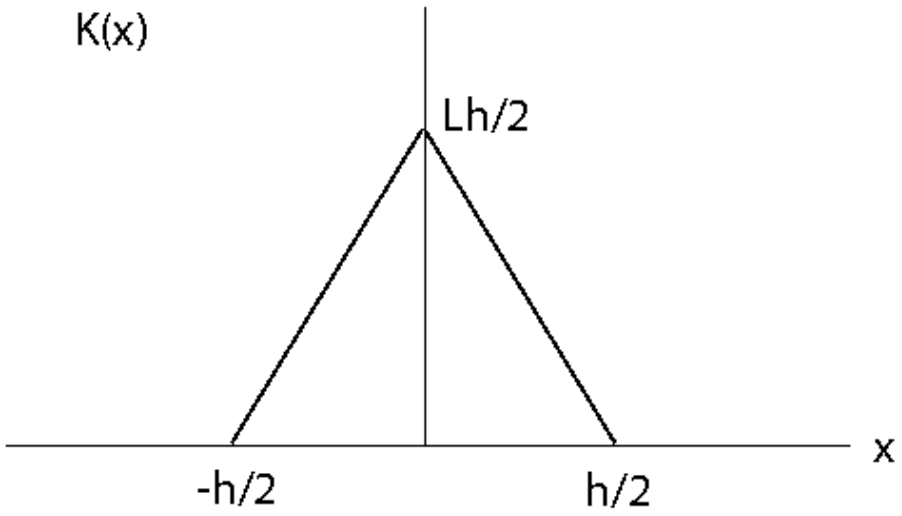


Figure 21.1

Note that $|K(a) - K(b)| \leq L|a - b|$, $\forall a, b$. The subclass we are going to consider are functions of the form

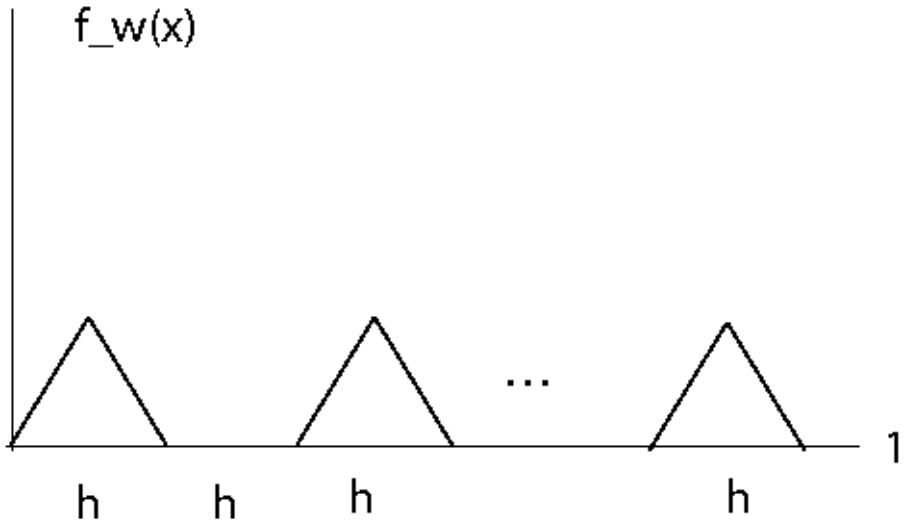


Figure 21.2

i.e. "bump" functions. Let $\Omega = \{0,1\}^m$ be the collection of binary vectors of length m, e.g. $w = (1,0,1,\cdots,0) \in \Omega$. Define

$$f_w(x) = \sum_{i=1}^{m} w_i K\left(x - \frac{h}{2}(2i-1)\right) \tag{21.53}$$

Note that for $w, w^{'} \in \Omega$,

$$\begin{aligned} d(f_w, f_{w^{'}}) = ||f_w - f_{w^{'}}|| &= \left(\int_0^1 \sum_{i=1}^m \left(w_i - w_i^{'}\right)^2 K^2\left(x - \tfrac{h}{2}(2i-1)\right)\right)^{1/2} \\ &= \sqrt{\rho(w,w^{'})}\sqrt{\int K^2(x)\,dx} \end{aligned} \tag{21.54}$$

where $\rho\left(w,w^{'}\right)$ is the Hamming distance, $\rho\left(w,w^{'}\right) = \sum_{i=1}^m |w_i - w_i^{'}|^2 = \sum_{i=1}^m |w_i - w_i^{'}|$. Now

$$\int K^2(x) = 2\int_0^{h/2} L^2 x^2 dx = 2L^2 \frac{h^3}{3\cdot 8} = \frac{L^2}{12}h^3 \tag{21.55}$$

so

$$d(f_w, f_{w^{'}}) = \sqrt{\rho(w,w^{'})}\frac{L}{\sqrt{12}}h^{3/2} \tag{21.56}$$

Since $|\Omega| = 2^n$, the number of functions in our class is 2^n. Turns out, we do not need to consider all functions $f_w, w \in \Omega$, but only a select few. Using all the functions leads to a looser lower bound of the form n^{-1}, which corresponds to the parametric rate. The problem under consideration is non-parametric, and hence we expect a slower rate of convergence. To get a tighter lower bound, the following result is of use:

Lemma 21.4: Varshamov-Gilbert '62
Let $m \geq 8$. There exists a subset $\{w^{(0)}, \cdots, w^{(M)}\}$ of Ω such that $w^{(0)} = (0, 0, \cdots, 0)$,

$$\rho\left(w^{(j)}, w^{(k)}\right) \geq \frac{m}{8}, \qquad \forall 0 \leq j < k \leq M \text{ and } M \geq 2^{m/8}. \tag{21.57}$$

What this lemma says is that there are many ($\sim 2^m$) sequences in Ω that are very different (i.e. $\rho\left(w^{(j)}, w^{(k)}\right) \sim m$). We are going to use the lemma to construct a useful set of hypotheses. Let $\{w^{(0)}, \cdots, w^{(M)}\}$ be the class of sequences in the lemma and define

$$f_j \triangleq f_{w^{(j)}}, \qquad j \in \{0, \cdots, M\} \tag{21.58}$$

We now need to look at the conditions of Theorem 2 and choose m appropriately.

First note that for $j \neq k$,

$$d\left(f_j, f_k\right) = \sqrt{\rho\left(w^{(j)}, w^{(k)}\right)} \frac{L}{\sqrt{12}} h^{3/2} \geq \sqrt{\frac{m}{8}} \frac{L}{\sqrt{12}} m^{-3/2} = \frac{L}{4\sqrt{6}} m^{-1} \tag{21.59}$$

Now let $P_j \triangleq P^{(j)}_{Y_1, \cdots, Y_m}, j \in \{0, \cdots, M\}$. Then

$$\begin{aligned} K\left(P_j || P_0\right) &= \mathbb{E}_j \left[log \frac{p^{(j)}_{Y_1, \cdots, Y_m}}{p^{(0)}_{Y_1, \cdots, Y_m}} \right] \\ &= \sum_{i=1}^n \mathbb{E}_j \left[log \frac{p^{(j)_{Y_i}}}{p^{(0)}_{Y_i}} \right] = \frac{1}{2\sigma^2} \sum_{i=1}^n f_j^2\left(x_i\right) \\ &\leq \frac{1}{2\sigma^2} \sum_{i=1}^n \left(\frac{Lh}{2}\right)^2 = \frac{L^2}{8\sigma^2} nh^2 = \frac{L^2}{8\sigma^2} nm^{-2} \end{aligned} \tag{21.60}$$

Now notice that $logM \geq \frac{m}{8} log2$ (from Lemma). We want to choose m such that

$$\frac{1}{M} \sum_{j=1}^M K\left(P_j || P_0\right) \leq \frac{L^2}{8\sigma^2} nm^{-2} < \alpha \frac{m}{8} log2 \leq \alpha logM \tag{21.61}$$

This gives

$$m > \left(\frac{L^2}{\alpha \sigma^2 log2}\right)^{1/3} n^{1/3} := C_0 n^{1/3} \tag{21.62}$$

so take $m = \lfloor C_0 n^{1/3} + 1 \rfloor$. Now

$$d\left(f_j, f_k\right) \geq \frac{L}{4\sqrt{6}} m^{-1} \geq 2\text{const } n^{-1/3} \qquad \text{for } n \geq n_0 \left(\text{const}\right) \tag{21.63}$$

Therefore,

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} P_f \left(||\hat{f}_n - f|| \geq \text{const } n^{-1/3} \right) \geq c > 0 \tag{21.64}$$

or,

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} P_f \left(||\hat{f}_n - f||^2 \geq \text{const } n^{-2/3} \right) \geq c > 0 \tag{21.65}$$

or after Markov's inequality,

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[||\hat{f}_n - f||^2 \right] \geq c \cdot \text{const } n^{-2/3} \tag{21.66}$$

Therefore, the estimator constructed in class attains the optimal rate of convergence.

Glossary

((Bayes' Risk)

The Bayes' risk is the infimum of the risk for all classifiers:

$$R^* = \inf_f R(f). \tag{3.4}$$

We can prove that the Bayes risk is achieved by the Bayes classifier.

B Bayes Classifier

The Bayes classifier is the following mapping:

$$f^*(x) = \begin{cases} 1, & \eta(x) \geq 1/2 \\ 0, & otherwise \end{cases} \tag{3.5}$$

where

$$\eta(x) \equiv P_{Y|X}(Y = 1|X = x). \tag{3.6}$$

Note that for any x, $f^*(x)$ is the value of $y \in \{0, 1\}$ that maximizes $P_{XY}(Y = y|X = x)$.

E Empirical Risk

Let $\{X_i, Y_i\}_{i=1}^n \overset{iid}{\sim} P_{XY}$ be a collection of training data. Then the empirical risk is defined as

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i). \tag{3.22}$$

Empirical risk minimization is the process of choosing a learning rule which minimizes the empirical risk; **i.e.,**

$$\hat{f}_n = \underset{f \in \mathcal{F}}{argmin} \hat{R}_n(f). \tag{3.23}$$

P Prefix Code

A code is called a prefix code if no codeword is a prefix of any other codeword.

Example: From Cover & Thomas '91Consider an alphabet of symbols, say A, B, C, and D and the codebooks below

---

This is an unsupported media type. To view, please see http://cnx.org/content/m16271/latest/

Figure 10.1

In the singular codebook we assign the same codeword to each symbol - a system that is obviously flawed! In the second case, the codes are not singular but the codeword 010 could represent B or CA or AD. Hence it is not a uniquely decodable codebook.

The third and fourth cases are both examples of uniquely decodable codebooks, but the fourth has the added feature that no codeword is a prefix of another. Prefix codes can be decoded from left to right since each codeword is "self-punctuating" - in this case with a zero to indicate the end of each word.

To design a uniquely decodable codebook in general is as challenging as the problem of selecting $c(f)$ to satisfy

$$\sum_{f \in \mathcal{F}} e^{-c(f)} < \infty. \tag{10.17}$$

However, prefix codes can often be easily designed or specified and they are inherently decodable. Moreover, prefix codes satisfy an important inequality called the Kraft Inequality .

T The VC dimension

$V_{\mathcal{A}}$ of a collection of sets $\mathcal{A}$ is defined as the largest interger n such that $S_{\mathcal{A}}(n) = 2^n$.

Example: $\mathcal{A} = \{(-\infty, t]\ ; t\, \varepsilon\, \mathcal{R}\}, \mathcal{S}_{\mathcal{A}} = n + 1$ hence $\mathcal{V}_{\mathcal{A}} = 1$.

Example: $\mathcal{A} = \{$ all rectangles in $\mathcal{R}^2\}$.

$\mathcal{S}_{\mathcal{A}} = 2^n, n = 1, 2, 3, 4$ and $\mathcal{S}_{\mathcal{A}} \leq 2^n, n = 4$, Hence $V_{\mathcal{A}} = 4$.

The VC dimension provides a useful bound on the growth of the shatter coefficients.

Bibliography

[1] T. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[2] L. Devroye, L. Gyrfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Spring, New York, 1996.

[3] S. M. Kay. *Fundamentals of Statistical Signal Processing.* Prentice Hall, 1993.

[4] E. L. Lehmann. *Theory of Point Estimation.* Wiley, New York, 1983.

[5] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I.* Wiley, New York, 1968.

[6] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

[7] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

Index of Keywords and Terms

Keywords are listed by the section with that keyword (page numbers are in parentheses). Keywords do not necessarily appear in the text of the page. They are merely associated with that section. *Ex.* apples, § 1.1 (1) **Terms** are referenced by the page they appear on. *Ex.* apples, 1

Attributions

Collection: *Statistical Learning Theory*
Edited by: Robert Nowak
URL: http://cnx.org/content/col10532/1.3/
License: http://creativecommons.org/licenses/by/2.0/

Module: "Basic Elements of Statistical Decision Theory and Statistical Learning Theory"
By: Robert Nowak
URL: http://cnx.org/content/m16263/1.3/
Pages: 1-6
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Elements of Statistical Learning Theory"
By: Robert Nowak
URL: http://cnx.org/content/m16269/1.2/
Pages: 7-9
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Introduction to Classification and Regression"
By: Robert Nowak
URL: http://cnx.org/content/m16272/1.2/
Pages: 11-20
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Introduction to Complexity Regularization"
By: Robert Nowak
URL: http://cnx.org/content/m16274/1.2/
Pages: 21-28
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "An Example of the Use of Sieves for Complexity Regularization in Denoising"
By: Robert Nowak
URL: http://cnx.org/content/m16261/1.3/
Pages: 29-34
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Plug-In Classifier and Histogram Classifier"
By: Robert Nowak
URL: http://cnx.org/content/m16280/1.2/
Pages: 35-41
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Probably Approximately Correct (PAC) Learning"
By: Robert Nowak
URL: http://cnx.org/content/m16282/1.2/
Pages: 43-48
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Chernoff's Bound and Hoeffding's Inequality"
By: Robert Nowak
URL: http://cnx.org/content/m16264/1.2/
Pages: 49-55
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Classification Error Bounds"
By: Robert Nowak
URL: http://cnx.org/content/m16265/1.2/
Pages: 57-60
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Error Bounds in Countably Infinite Spaces"
By: Robert Nowak
URL: http://cnx.org/content/m16271/1.2/
Pages: 61-66
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Complexity Regularization"
By: Robert Nowak
URL: http://cnx.org/content/m16266/1.2/
Pages: 67-70
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Decision Trees"
By: Robert Nowak
URL: http://cnx.org/content/m16287/1.2/
Pages: 71-83
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

Module: "Complexity Regularization for Squared Error Loss"
By: Robert Nowak
URL: http://cnx.org/content/m16267/1.2/
Pages: 85-90
Copyright: Robert Nowak
License: http://creativecommons.org/licenses/by/2.0/

About Connexions

Since 1999, Connexions has been pioneering a global system where anyone can create course materials and make them fully accessible and easily reusable free of charge. We are a Web-based authoring, teaching and learning environment open to anyone interested in education, including students, teachers, professors and lifelong learners. We connect ideas and facilitate educational communities.

Connexions's modular, interactive courses are in use worldwide by universities, community colleges, K-12 schools, distance learners, and lifelong learners. Connexions materials are in many languages, including English, Spanish, Chinese, Japanese, Italian, Vietnamese, French, Portuguese, and Thai. Connexions is part of an exciting new information distribution system that allows for **Print on Demand Books**. Connexions has partnered with innovative on-demand publisher QOOP to accelerate the delivery of printed course materials and textbooks into classrooms worldwide at lower prices than traditional academic publishers.