Naive Bayes Algorithms

The Naive Bayes model is an old method for classification and predictor selection that is enjoying a renaissance because of its simplicity and stability.

Notation

The following notation is used throughout this chapter unless otherwise stated:

J_0	Total number of predictors.
X	Categorical predictor vector **X'** = (X_1, ..., X_J), where J is the number of predictors considered.
M_j	Number of categories for predictor X_j.
Y	Categorical target variable.
K	Number of categories of Y.
N	Total number of cases or patterns in the training data.
N_k	The number of cases with Y= k in the training data.
N^j_{mk}	The number of cases with Y= k and X_j=m in the training data.
π_k	The probability for Y= k.
p^j_{mk}	The probability of X_j=m given Y= k.

Naive Bayes Model

The Naive Bayes model is based on the conditional independence model of each predictor given the target class. The Bayesian principle is to assign a case to the class that has the largest posterior probability. By Bayes' theorem, the posterior probability of Y given **X** is:

$$P(Y=k|\mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x}|Y=k)P(Y=k)}{\sum_{i=1}^{K} P(\mathbf{X}=\mathbf{x}|Y=i)\,P(Y=i)}$$

Let X_1, ..., X_J be the J predictors considered in the model. The Naive Bayes model assumes that X_1, ..., X_J are conditionally independent given the target; that is:

$$P\left(\mathbf{X}=\mathbf{x}\,|Y=k) = \prod_{j=1}^{J} P\left(X_j = x_j|\,Y=k\right)\right.$$

These probabilities are estimated from training data by the following equations:

$$\pi_k = P(Y=k) = \frac{N_k+\lambda}{N+K\lambda}$$

$$p^j_{mk} = P(X_j=m|Y=k) = \frac{N^j_{mk}+f}{\Sigma_{l=1}^{M_j} N^j_{lk} + M_j f}$$

Where N_k is calculated based on all non-missing Y, N^j_{mk} is based on all non-missing pairs of X_J and Y, and the factors λ and *f* are introduced to overcome problems caused by zero or very small cell counts. These estimates correspond to Bayesian estimation of the multinomial probabilities with Dirichlet priors. Empirical studies suggest $\lambda = f = \frac{1}{N}$(Kohavi et al., 1997).

A single data pass is needed to collect all the involved counts.

For the special situation in which J = 0; that is, there is no predictor at all, $P(Y = k|\mathbf{X} = \mathbf{x}) = P(Y = k)$. When there are empty categories in the target variable or categorical predictors, these empty categories should be removed from the calculations.

Preprocessing

Missing Values

A predictor is ignored if every value is missing or if it has only one observed category. A case is ignored if the value of the target variable or the values of all predictors are missing. For each case missing some, but not all, of the values of the predictors, only the predictors with nonmissing values are used to predict the case, as suggested in (Kohavi et al., 1997).

This implies the following equation:

$$P(\mathbf{X} = \mathbf{x}_i|Y = y_i) = \prod_{\{j:x_{ji}\text{not missing}\}} P(X_j = x_{ji}|Y = y_i)$$

This also implies the following equation for B(J) in average log-likelihood calculations:

$$B(J) = -\frac{1}{N'}\sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \prod_{\{j:x_{ji}\text{not missing}\}} p^j_{x_{ji}k}\right)$$

Where the log() term for case *i* is ignored if all the values of the predictors considered in the model are missing, and *N'* is the total number of terms that are not ignored in the sum. For more information, see “Average Log-likelihood” on p. 4.

Continuous Variables

The Naive Bayes model assumes that the target and predictor variables are categorical. If there are continuous variables, they need to be discretized. There are many ways to discretize a continuous variable; the simplest is to divide the domain of a variable into equal width bins. This method performs well with the Naive Bayes model while no obvious improvement is found when complex methods are used (Hsu et al., 2000).

Sometimes the equal width binning method may produce empty bins. In this case, empty bins are eliminated by changing bin boundary points. Let $b_1 < b_2 < \ldots b_n$ be the bin boundary points produced by the equal width binning method. The two end bins $(-\infty, b_1]$ and (b_n, ∞) are non-empty by design. Suppose that bin $(b_i, b_{i+1}]$ is empty, and suppose that the closest left

non-empty bin has right boundary point b_j (< b_i) and the closest right non-empty bin has left boundary point b_k (> b_i). Then empty bins are eliminated by deleting all boundary points from b_j to b_k, and setting a new boundary point at $(b_j+b_k)/2$.

Feature Selection

Given a total of J_0 predictors, the goal of feature selection is to choose a subset of J predictors using the Naive Bayes model (Natarajan and Pednault, 2001). This process has the following steps:

- Collect the necessary summary statistics to estimate all possible model parameters.
- Create a sequence of candidate predictor subsets that has an increasing number of predictors; that is, each successive subset is equal to the previous subset plus one more predictor.
- From this sequence, find the "best" subset.

Collect Summary Statistics

One pass through the training data is required to collect the total number of cases, the number of cases per category of the target variable, and the number of cases per category of the target variable for each category of each predictor.

Create the Sequence of Subsets

Start with an initial subset of predictors considered vital to the model, which can be empty. For each predictor not in the subset, a Naive Bayes model is fit with the predictor plus the predictors in the subset. The predictor that gives the largest average log-likelihood is added to create the next larger subset. This continues until the model includes the user-specified:

- Exact number of predictors

or

- Maximum number of predictors

Alternatively, the maximum number of predictors, J_{Max}, may be automatically chosen by the following equation:

$$J_{\text{Max}} = \min\left\{J_{\text{Must}} + \min\left\{100, \max\left(20, \frac{J_0}{5}\right)\right\}, J_0\right\}$$

where J_{Must} is the number of predictors in the initial subset.

Find the "Best" Subset

If you specify an exact number of predictors, the final subset in the sequence is the final model. If you specify a maximum number of predictors, the "best" subset is determined by one of the following:

- A test data criterion based on the average log-likelihood of the test data.
- A pseudo-BIC criterion that uses the average log-likelihood of the training data and penalizes overly complex models using the number of predictors and number of cases. This criterion is used if there are no test data.

Smaller values of these criteria indicate "better" models. The "best" subset is the one with the smallest value of the criterion used.

Test Data Criterion

$$Q(J) = -\bar{l}_{\text{Test}}(J)$$

Where $\bar{l}_{\text{Test}}(J)$ is the average log-likelihood for test data.

Pseudo-BIC Criterion

$$Q(J) = -\bar{l}_{\text{Train}}(J) + \tfrac{1}{2} J \tfrac{\log(N)}{N}$$

Where J denotes the number of predictors in the model, and $-\bar{l}_{\text{Train}}(J)$ is the average log-likelihood for training data.

Average Log-likelihood

The average (conditional) log-likelihood for data $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ with J predictors is

$$\begin{aligned}
\bar{l}(J) &= \tfrac{1}{N}\log L = \tfrac{1}{N}\sum_{i=1}^{N}\log P(Y = y_i|\mathbf{X} = \mathbf{x}_i) \\
&= \tfrac{1}{N}\sum_{i=1}^{N}\log P(Y = y_i) + \frac{1}{N}\sum_{i=1}^{N}\log P(\mathbf{X} = \mathbf{x}_i|Y = y_i) \\
&- \tfrac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K} P(\mathbf{X} = \mathbf{x}_i|Y = k)\, P(Y = k)\right) \\
&= \tfrac{1}{N}\sum_{k=1}^{K} N_k \log(\pi_k) + \frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{m=1}^{M_j} N_{mk}^{j}\log\left(p_{mk}^{j}\right) - \frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\pi_k\prod_{j=1}^{J} p_{x_{ji}k}^{j}\right)
\end{aligned}$$

Let

$$A(J) = \tfrac{1}{N}\sum_{k=1}^{K} N_k \log(\pi_k) + \frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{m=1}^{M_j} N_{mk}^{j}\log\left(p_{mk}^{j}\right)$$

$$B(J) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\pi_k\prod_{j=1}^{J}p_{x_{ji}k}^{j}\right)$$

then

$$\overline{l}(J) = A(J) + B(J)$$

Note: for the special situation in which J = 0; that is, there are no predictors,
$\overline{l}(J) = \frac{1}{N}\sum_{i=1}^{N}\log P(Y = y_i) = \frac{1}{N}\sum_{k=1}^{K}N_k\log(\pi_k)$

Calculation of average log-likelihood by sampling

When adding each predictor to the sequence of subsets, a data pass is needed to calculate B(J). When the data set is small enough to fit in the memory, this is not a problem. When the data set cannot fit in memory, this can be costly. The Naive Bayes model uses simulated data to calculate B(J). Other research has shown that this approach yields good results . The formula for B(J) can be rewritten as, for a data set of *m* cases:

$$B(J) = -\frac{1}{m}\sum_{i=1}^{m}\log\left(\sum_{k=1}^{K}\pi_k\prod_{j=1}^{J}p_{x_{ji}k}^{j}\right)$$

By default *m* = 1000.

Classification

The target category with the highest posterior probability is the predicted category for a given case.

$$\hat{y}(\mathbf{x}) = \underset{k}{\text{argmax}}\{P(Y = k|\mathbf{X} = \mathbf{x})\} = \underset{k}{\text{argmax}}\{P(\mathbf{X} = \mathbf{x}|Y = k)P(Y = k)\}$$

Ties are broken in favor of the target category with greater prior probability π_k.

When cases being classified contain categories of categorical predictors that did not occur in the training data, these new categories are treated as missing.

Classification Error

If there is test data, the error equals the misclassification ratio of the test data. If there is no test data, the error equals the misclassification ratio of the training data.

References

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53, 370–418.

Becker, B., R. Kohavi, and D. Sommerfield. 2001. Visualizing the Simple Bayesian Classifier. In: *Information Visualization in Data Mining and Knowledge Discovery,* U. Fayyad, G. Grinstein, and A. Wierse, eds. San Francisco: Morgan Kaufmann Publishers, 237–249.

Domingos, P., and M. J. Pazzani. 1996. Beyond Independence: conditions for the optimality of the simple Bayesian classifier. In: *Machine Learning: Proceedings of the Thirteenth International Conference,* L. Saitta, ed., 105–112.

Hsu, C., H. Huang, and T. Wong. 2000. Why Discretization Works for Naive Bayesian Classifiers. In: *Proceedings of the 17th International Conference on Machine Learning,* San Francisco: MorganKaufman, 399–406.

Kohavi, R., and D. Sommerfield. 1995. Feature subset selection using the wrapper model: Overfitting and dynamic search space topology. In: *The First International Conference on Knowledge Discovery and Data Mining,* Menlo Park, California: AAAIPress, 192–197.

Kohavi, R., B. Becker, and D. Sommerfield. 1997. Improving Simple Bayes. In: *Proceedings of the European Conference on Machine Learning,* , 78–87.

Natarajan, R., and E. Pednault. 2001. Using Simulated Pseudo Data to Speed Up Statistical Predictive Modeling from Massive Data Sets. In: *SIAM First International Conference on Data Mining,* .

Bibliography

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53, 370–418.

Becker, B., R. Kohavi, and D. Sommerfield. 2001. Visualizing the Simple Bayesian Classifier. In: *Information Visualization in Data Mining and Knowledge Discovery,* U. Fayyad, G. Grinstein, and A. Wierse, eds. San Francisco: Morgan Kaufmann Publishers, 237–249.

Domingos, P., and M. J. Pazzani. 1996. Beyond Independence: conditions for the optimality of the simple Bayesian classifier. In: *Machine Learning: Proceedings of the Thirteenth International Conference,* L. Saitta, ed., 105–112.

Hsu, C., H. Huang, and T. Wong. 2000. Why Discretization Works for Naive Bayesian Classifiers. In: *Proceedings of the 17th International Conference on Machine Learning,* San Francisco: MorganKaufman, 399–406.

Kohavi, R., and D. Sommerfield. 1995. Feature subset selection using the wrapper model: Overfitting and dynamic search space topology. In: *The First International Conference on Knowledge Discovery and Data Mining,* Menlo Park, California: AAAIPress, 192–197.

Kohavi, R., B. Becker, and D. Sommerfield. 1997. Improving Simple Bayes. In: *Proceedings of the European Conference on Machine Learning,* , 78–87.

Natarajan, R., and E. Pednault. 2001. Using Simulated Pseudo Data to Speed Up Statistical Predictive Modeling from Massive Data Sets. In: *SIAM First International Conference on Data Mining,* .