# IBM SPSS Neural Networks 24

**注释**

使用本信息及其支持的产品之前，请阅读第 17 页的『声明』中的信息。

**产品信息**

此版本适用于 IBM® SPSS® Statistics V24.0.0 及所有后续发布和修订，除非在新版本中另有说明。

# 目录

# 第 1 章 Neural Networks 简介

因为 Neural networks 的强大性、灵活性和易用性，Neural networks 是很多预测数据挖掘应用程序的首选工具。预测神经网络在基础过程复杂的应用程序中特别有用，例如:

- 预测消费者需求以组织生产与交付成本。
- 预测对直接邮寄营销作出响应的概率以确定应给邮寄列表上的哪个家庭发送优惠。
- 给申请人评分以确定为申请人延长贷款的风险。
- 检测保险理赔数据集中的欺骗性交易。

从模型预测结果可以与目标变量的已知值进行比较的意义上来说，用于预测应用程序的 Neural Networks，例如**多层感知器（MLP）**和**径向基函数（RBF）**网络是受监督的。Neural Networks 选项允许您拟合 MLP 和 RBF 网络并保存结果模型以供评分。

## Neural Networks 是什么?

术语**神经网络**应用于关系松散的系列模型，并具有大型参数空间和灵活结构的特征，大脑机能研究递减。随着系列增长，大部分新模型经设计用于非生物学应用程序，虽然大量相关术语反映其起源。

神经网络的特定定义随其所应用于的字段而变化。没有任何单个定义包括整个模型系列，现在，考虑以下描述 [1]:

神经网络为大量平行分布的处理器，并具有存储经验知识及供使用的自然特性。其在两方面与大脑类似:

- 网络通过学习过程获取知识。
- 称为突触权重的中间神经元连接力度用于存储知识。

为讨论为何此定义可能过于限制，请参阅 [2]。

为使用此定义区分神经网络与传统统计方法，未述部分与定义的实际内容同样重要。例如，传统线性回归模型可通过最小平方方法获取知识并在回归系数存储知识。在此意义下，其为神经网络。实际上，您可以证明线性回归为特定神经网络的特殊个案。但是，线性回归具有严格模型结构和在学习数据之前施加的一组假设。

比较而言，以上定义规定模型结构和假设相关最小需求。因此，神经网络可以接近多种统计模型，并无需您预先假设因变量和自变量间的特定关系。相反，关系表在学习过程中确定。因变量和自变量间的线性关系适合，神经网络结果应接近线性回归模型的结果。如果非线性关系更适合，神经网络将自动接近“正确”模型结构。

此灵活性的平衡指神经网络的突触权重不可轻松解释。因此，如果您正试图解释生成因变量和自变量间关系的基础过程，最好使用更传统的统计模型。但是，如果模型的可解释性并不重要，您可以使用神经网络更快获取良好模型结果。

1. Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, 2nd ed. New York: Macmillan College Publishing.

2. Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

## Neural Networks 结构

尽管 Neural Networks 对模型结构和假设施加最小需求，但是对理解一般**网络体系结构**非常有用。多层感知器（MLP）或径向基函数（RBF）网络是一个将目标变量（也称为输出）的预测误差最小化的预测变量函数（也称为输入或自变量）。

请考虑随产品一起提供的 *bankloan.sav* 数据集，在其中您想在众多贷款申请者中标识潜在拖欠者。应用到该问题的 MLP 或 RBF 网络是一个将预测拖欠贷款的误差最小化的测量函数。下图对关联此函数的形式非常有用。

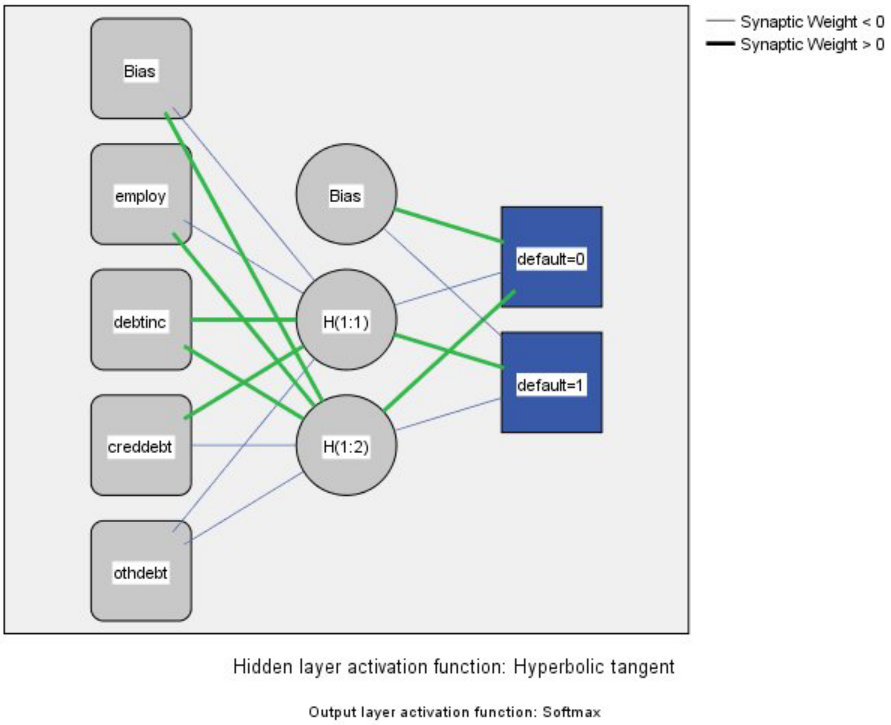


*图 1. 有一个隐藏层的前馈体系结构*

此结构称为**前馈体系结构**，因为网络中的连接未经任何反馈循环就从输入层转到了输出层。本图中:

- **输入层**包含预测变量。
- **隐藏层**包含无法观察的节点或单元格。每个隐藏单元格的值都是某个预测变量函数；函数的确切形式部分取决于网络类型，部分取决于用户可控制的规格。
- **输出层**包含响应。由于欠贷历史是一个有两种类别的分类变量，它可以重新编码为两个指示符变量。每个输出单位是隐藏单位的某些函数。同样，函数的确切形式部分取决于网络类型，还有部分取决于用户可控制的规格。

MLP 网络允许第二个隐藏层；在这种情况下，第二个隐藏层的每个单元格都是第一个隐藏层单元格的一个函数，并且每个响应都是第二个隐藏层单元格的一个函数。

# 第 2 章 多层感知器

“多层感知器” (MLP) 过程会根据预测变量的值来生成一个或多个因变量（目标变量）的预测模型。

**示例。** 以下是使用 MLP 过程的两种情况:

银行信贷员需要能够找到预示有可能拖欠贷款的人的特征，然后使用这些特征来识别信用风险的高低。使用以往客户的样本，她可以训练多层感知器，用以往客户的坚持样本来验证分析，然后再用网络将潜在客户按高或低信用风险分类。

医院系统注重跟踪接受心肌梗塞（MI 或“心脏病发作”）治疗的病人的成本与住院时间。获取这些测量的精确估计值有助于管理部门在病人接受治疗时正确管理现有床位。使用接受 MI 治疗的病人样本的治疗记录，管理员可以训练网络以预测成本和住院时间。

数据注意事项

**因变量。** 因变量可以是:

- *名义 (Nominal).* 当变量值表示不具有内在等级的类别时，该变量可以作为名义变量；例如，雇员任职的公司部门。名义变量的示例包括地区、邮政编码和宗教信仰。
- *有序 (Ordinal).* 当变量值表示带有某种内在等级的类别时，该变量可以作为有序变量；例如，从十分不满意到十分满意的服务满意度水平。有序变量的示例包括表示满意度或可信度的态度分数和优先选择评分。
- *刻度 (Scale).* 当变量值表示带有有意义的度规的已排序类别时，该变量可以作为刻度（连续）变量对待，以便在值之间进行合适的距离比较。刻度变量的示例包括以年为单位的年龄和以千美元为单位的收入。

  该过程假设已经将适当的测量级别分配给所有因变量，但您可以通过在源变量列表中右键单击该变量并从弹出菜单中选择测量级别暂时更改变量的测量级别。

变量列表中每个变量旁的图标标识测量级别和数据类型:

*表 1. 测量级别图标*

| | 数值 | 字符串 | 日期 | 时间 |
|---|---|---|---|---|
| 刻度（连续） |  | n/a |  |  |
| 有序 |  |  |  |  |
| 名义 |  |  |  |  |

**预测变量。** 预测变量可指定为因子（分类）或协变量（刻度）。

**类别变量编码。** 该过程使用使用一个 *c* 编码在过程期间临时重新编码分类预测变量和因变量。如果存在 *c* 分类变量，那么该变量存储为 c 向量，第一个类别表示为（1、0、...、0）、下一个类别表示为（0、1、0、...、0）、...、最后一个类别表示为（0、0、...、0、1）。

此编码方案增加了突触权重的数目并会导致培训减速，但是多数“压缩”编码方法通常导致较差的拟合神经网络。如果您的网络培训进行很慢，尝试通过将类似的类别组合起来或删除具有极少见类别的个案以减少分类预测变量中的类别数目。

所有 *c* 之一的编码以培训数据为基础，即使已经定义检验或坚持样本（请参阅第 5 页的『分区』）。因此，如果检验或坚持样本包含培训数据中不存在的预测变量类别个案，那么那些个案不用于该过程或评分。如果检验或坚持样本包含培训数据中不存在的因变量类别个案，那么那些个案已经用于该过程，但可能被评分。

**重新调整。** 在缺省情况下，将重新调整刻度因变量和协变量以改善网络培训。基于培训数据执行所有重标度，即使已经定义检验或坚持样本（请参阅第 5 页的『分区』）。也就是说，根据重标度的类型，仅使用培训数据计算平均值、标准差、协变量或因变量的最小值或最大值。如果您指定一个变量以定义分区，这些协变量或因变量在培训样本、检验样本或坚持样本之间具有相似分布将至关重要。

**频率权重。** 该过程忽略频率权重。

**复制结果。** 如果您想准确复制您的结果，除了使用相同过程设置以外，还可以使用针对随机数字生成器的相同初始化值、相同数据顺序和相同变量顺序。有关此问题的更多详细信息，请参阅以下内容:

- **随机数字生成器。** 该过程在分区随机分配、突触权重初始化的随机子样本、体系结构自动选择的随机子样本、用于权重初始化和体系结构自动选择的模拟加强算法之间使用随机数字生成器。想要以后再次生成相同的随机结果，在每次运行多层感知器过程之前使用随机数字生成器的相同初始化值。
- **个案顺序。** 在线和袖珍型批处理培训方法（请参阅第 6 页的『培训』）明显取决于个案顺序；然而，甚至批处理培训取决于个案顺序，因为突触权重初始化包含数据集的子样本。

  要使顺序的影响降至最低程度，可随机个案等级排序的顺序。想要验证给定解的稳定性，您可能想要通过以不同随机顺序排序的案例来得到多个不同的解。在文件非常大的情况，可使用以不同随机顺序排序的个案样本运行多次。
- **变量顺序。** 由于变量顺序改变时分配了不同模式的初始值，结果可能会受到因子和协变量列表中的变量顺序的影响。因为个案顺序影响，您可能要尝试不同变量顺序（只需在因子或协变量列表中拖放）以评估给出解的稳定性。

创建多层感知器网络

从菜单中选择:

**分析 > Neural Networks > 多层感知器...**

1. 选择至少一个因变量。
2. 至少选择一个因子或协变量。

根据需要，在变量选项卡上您可以更改重标度协变量的方法。选项为:

- **标准化。** 减去平均值并除以标准差，(*x*–mean)/*s*。
- **标准化。** 减去平均值并除以范围，(*x*–min)/(max–min)。标准化值介于 0 和 1 之间。
- **调整标准化。** 减去最小值并除以范围所得到的调整版本，[2*(*x*–min)/(max–min)]–1。调整后的标准化值介于 –1 和 1 之间。
- **无。** 无协变量重标度。

具有未知测量级别的字段

当数据集中的一个或多个变量（字段）的测量级别未知时，将显示测量级别警告。由于测量级别会影响该过程的计算结果，因此所有变量必须都定义有测量级别。

**扫描数据。** 读取活动数据集中的数据，并分配缺省测量级别给任何具有当前未知测量级别的字段。如果数据集较大，该过程可能需要一些时间。

**手动分配。** 打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框将测量级别分配给这些字段。您也可以在数据编辑器的变量视图中分配测量级别。

由于测量级别对该过程很重要，因此您无法访问运行该过程的对话框，除非所有字段均定义了测量级别。

## 分区

**分区数据集。** 此组指定将活动数据集划分为训练样本、检验样本或坚持样本的方法。**训练样本**包含用于训练神经网络的数据记录；数据集中的某些个案百分比必须分配给训练样本以获得一个模型。**检验样本**是一个用于跟踪训练过程中的错误以防止超额训练的独立数据记录集。强烈建议您创建一个训练样本，并且如果测试样本小于训练样本，网络训练通常最高效。**坚持样本**是另一个用于评估最终神经网络的独立数据记录集；坚持样本的误差给出一个模型预测能力的"真实"估计值，因为坚持个案不用于构建模型。

- **根据个案的相对数量随机分配个案。** 指定随机分配到每个样本（训练、检验和坚持）的个案的相对数量（比率）。**%** 列根据您已经指定的相对数量，报告将被分配到每个样本的个案的百分比。

  例如，指定 7、3、0 作为训练、检验和坚持样本的相对数量对应于 70%、30% 和 0%。指定 2、1、1 作为相对数量对应 50%、25% 和 25%；1、1、1 对应将数据集在训练、检验和坚持中分为相等的三部分。
- **使用分区变量分配个案。** 指定一个将活动数据集中的每个个案分配到训练、检验和坚持样本中的数值变量。变量为正值的个案被分配到训练样本中，值为 0 的个案被分配到检验样本中，而负值个案被分配到坚持样本中。具有系统缺失值的个案会从分析中排除。分区变量的任何用户缺失值始终视为有效。

*注*：使用分区变量将不能保证连续运行该过程会产生相同结果。请参阅主多层感知器主题中的"复制结果"。

## 体系结构

"体系结构"选项卡用于指定网络结构。该过程可以自动选择"最佳"体系结构，或者您也可以指定自定义体系结构。

体系结构自动选择构建具有一个隐藏层的网络。指定隐藏层中允许存在的最小或最大单位量，体系结构自动选择计算隐藏层中的"最佳"单位量。体系结构自动选择使用隐藏层和输出层的缺省激活函数。

自定义体系结构选择向您提供针对隐藏层和输出层的专业控制，并且当您预先知道需要什么体系结构或当您需要调整体系结构自动选择的结果时，其最有用。

隐藏层

隐藏层包含无法观察的网络节点（单位）。每个隐藏单位是一个输入权重总和的函数。该函数是激活函数，而且权重值由估计算法确定。如果网络包含第二个隐藏层，第二个层中的每个隐藏单位是第一个隐藏层中权重总和的函数。两个层使用相同激活函数。

*隐藏层数*。一个多层感知器可以有一个或两个隐藏层。

*激活函数 (Activation Function)*. 激活函数将某个层中的单位的权重总和"关联"到下一层的单位值。

- **双曲正切。** 此函数公式为：$\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$。其取实数值自变量并将其变换到 (–1, 1) 范围。使用体系结构自动选择时，此为隐藏层所有单位的激活函数。
- **Sigmoid。** 此函数公式为：$\gamma(c) = 1/(1+e^{-c})$。其取实数值自变量并将其变换到 (0, 1) 范围。

*单元格数*。可以明确指定或由估计算法自动确定每个隐藏层中的单元格数。

输出层

输出层包含目标（因）变量。

*激活函数 (Activation Function)*. 激活函数将某个层中的单位的权重总和"关联"到下一层的单位值。

- **恒等。** 此函数公式为: $\gamma(c) = c$。其取实数值自变量并将其原样返回。使用体系结构自动选择时，如果存在刻度因变量，则此为输出层中所有单位的激活函数。
- **Softmax。** 此函数公式为: $\gamma(c_k) = \exp(c_k)/\Sigma_j\exp(c_j)$。其取实数值自变量的向量，并将其变换到元素介于 (0, 1) 范围的向量且和为 1。只有所有因变量是分类变量时，才可以使用 Softmax。使用体系结构自动选择时，如果所有因变量是分类变量，此为输出层中所有单位的激活函数。
- **双曲正切。** 此函数公式为: $\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$。其取实数值自变量并将其变换到 (–1, 1) 范围。
- **Sigmoid。** 此函数公式为: $\gamma(c) = 1/(1+e^{-c})$。其取实数值自变量并将其变换到 (0, 1) 范围。

**刻度因变量重标度。** 至少选择一个刻度因变量时才可以使用这些控制。

- **标准化。** 减去平均值并除以标准差，$(x-\text{mean})/s$。
- **标准化。** 减去平均值并除以范围，$(x-\min)/(\max-\min)$。标准化值介于 0 和 1 之间。如果输出层使用 sigmoid 激活函数，则此为刻度因变量所需的重标度方法。修正值选项指定一个较小数字 $\epsilon$，并将其作为修正值应用于重标度公式中；此修正值确保所有重标度因变量值介于激活函数范围。具体来说，当 $x$ 取最小值和最大值时，未修正的公式中的值 0 和 1 将定义 sigmoid 函数的范围限制，但是不介于该范围之内。修正公式为 $[x-(\min-\epsilon)]/[(\max+\epsilon)-(\min-\epsilon)]$。指定大于等于 0 的数。
- **调整标准化。** 减去最小值并除以范围所得到的调整版本，$[2*(x-\min)/(\max-\min)]-1$。调整的标准化值介于 –1 和 1 之间。如果输出层使用双曲正切激活函数，则此为刻度因变量所需的重标度方法。修正值选项指定一个较小数字 $\epsilon$，并将其作为修正值应用于重标度公式中；此修正值确保所有重标度因变量值介于激活函数范围。具体来说，当 $x$ 取最小值和最大值时，未修正的公式中的值 –1 和 1 将定义双曲正切函数的范围限制，但是不介于该范围之内。修正公式为 $\{2*[(x-(\min-\epsilon))/((\max+\epsilon)-(\min-\epsilon))]\}-1$。指定大于等于 0 的数。
- **无。** 未对刻度因变量进行重标度。

## 培训

"培训"选项卡用于指定如何培训网络。培训的类型和优化算法确定哪个培训选项可用。

**培训类型。** 培训类型确定网络如何处理记录。从下列培训类型中选择:

- **批处理。** 只有传递所有培训数据记录之后才能更新突触权重；也就是说，批处理培训使用培训数据集中所有记录信息。批处理培训通常为首选方法，因为它直接使总误差最小；然而，批处理培训可能需要多次更新权重，直至满足其中一条中止规则，因此可能需要传递数据多次。其对于"较小"数据集最有用。
- **在线。** 在每一个培训数据记录之后更新突触权重；也就是说，在线培训一次使用一个记录信息。在线培训连续获取记录并更新权重，直至满足其中一条中止规则。如果一次使用所有记录，而且不满足任何中止规则，那么该过程通过循环数据记录继续。对于与预测变量相关的"较大"数据集，在线培训要优于批处理；也就是说，如果有许多记录和输入，并且其值之间不相互独立，那么在线培训可以比批处理培训更快获取一个合理答案。
- **袖珍型批处理。** 将培训数据记录划分到大小近似相等的组中，然后在传递一组之后更新突触权重；也就是说，袖珍型批处理培训使用一组记录信息。然后，如果需要，该过程循环数据组。袖珍型批处理培训提供

介于批处理培训和在线培训之间的折中方法，它可能最适于“中型”数据集。该过程可以自动确定每个袖珍型批处理培训记录的数目，或者您可以指定一个大于 1 并小于或等于内存中存储的最大个案数的整数。您可以在选项选项卡上设置将存储到内存的最大个案数。

**优化算法。** 这是一种用于估计突触权重的方法。

- **调整的共轭梯度。** 使用共轭梯度方法对齐的假设仅应用于批处理培训类型，所以此方法不适用于在线培训或袖珍型批处理培训。
- **梯度下降。** 此方法需与在线培训或袖珍型批处理培训共同使用；也可以与批处理培训共同使用。

**培训选项。** 该培训选项允许您细微调整优化算法。您一般无需更改这些设置，除非网络出现估计问题。

调整的共轭梯度算法的培训选项包括:

- **初始 Lambda 值。** 针对调整的共轭梯度算法的 lambda 参数初始值。指定大于 0 并小于 0.000001 的数。
- **初始西格玛值。** 针对调整的共轭梯度算法的西格玛参数初始值。指定大于 0 并小于 .0001 的数。
- **间隔中心点和间隔偏移量。** 间隔中心点（$a_0$）和间隔偏移量（$a$）定义间隔 $[a_0\ a,\ a_0+a]$ ,并且在使用模拟加强时，在其间随机生成权重向量。模拟加强用于取出局部最小值，目标是利用优化算法找到全局最小值。此方法用于权重初始化和体系结构自动选择。 指定间隔中心点数目且该数大于间隔偏移量 0。

梯度下降算法的培训选项包括:

- **最初学习率。** 针对梯度下降算法的学习率初始值，较高的学习率表明在可能转为不稳定的代价下，网络培训较快。指定大于 0 的数。
- **学习率的较低边界。** 针对梯度下降算法的学习率较低边界。此设置仅应用于在线和袖珍型批处理培训。指定大于 0 并小于初始学习率的数。
- **动能。** 针对梯度下降算法的初始动能参数。该动能项有助于阻止过高学习率引起的不稳定性。指定大于 0 的数。
- **时程学习率减少。** 梯度递减与在线培训或袖珍型批处理培训一起使用时，时程数（$p$）或培训样本的数据传递需要将初始学习率降低到学习率的较低边界。这使您能控制学习率衰减因子 $\beta = (1/p\ K)*\ln(\eta_0/\eta_{low})$，其中 $\eta_0$ 是初始学习率，$\eta_{low}$ 是学习率的较低极限，$K$ 是培训数据集中袖珍型批处理（或针对在线培训的培训记录数目）的总数目。指定大于 0 的整数。

## 输出

**网络结构。** 显示与神经网络有关的摘要信息。

- **描述。** 显示与神经网络有关的信息，包括因变量、输入和输出单位数目、隐藏层和单位数目及激活函数。
- **图表。** 将神经网络图表作为不可编辑图表显示。请注意，随着协变量数目和因子级别的增加，图表变得更加难于解释。
- **突触权重。** 显示表明给定层中的单位与以下层中的单位之间关系的系数估计值。突触权重以培训样本为基础，即使活动数据集已划分为培训数据、检验数据和坚持数据。请注意，突触权重数目会变得非常大，而且这些权重一般不用于解释网络结果。

**网络性能。** 显示用于确定模型是否“良好”的结果。*注*: 该组中的图表以培训样本和检验样本组合为基础，或者如果不存在检验样本，则只以培训样本为基础。

- **模型摘要。** 显示分区和整体神经网络结果的摘要，包括错误、相对错误或不正确预测的百分比、用于终止培训的中止规则和培训时间。

  恒等、sigmoid 或双曲正切激活函数应用于输出层时，错误为平方和误差。softmax 激活函数应用于输出层时，则为交叉熵错误。

显示相对错误或不正确预测的百分比取决于因变量测量级别。如果任何因变量具有刻度测量级别，则显示平均整体相对错误（相对于平均值模型）。如果所有因变量都为分类变量，则显示不正确预测的平均百分比。也针对单个因变量显示相对错误或不正确预测的百分比。

- **分类结果。** 分区和整体显示每个分类因变量的分类表。每个表针对每个因变量类别给出正确或错误分类的个案数目。也报告正确分类的总体个案百分比。
- **ROC 曲线。** 显示每个分类因变量的 ROC（Receiver Operating Characteristic）曲线。其也显示一个给定每个曲线下区域的表格。对于给定因变量，ROC 图表针对每个类别显示一条曲线。如果因变量有两个类别，那么每条曲线将该类别视为正态与其它类别。如果因变量有两个多类别，那么每条曲线将该类别视为正态与所有其它类别的汇总。
- **累积增益图。** 显示每个分类因变量的累积增益图。每个因变量类别的曲线的显示与 ROC 曲线相同。
- **效益图。** 显示每个分类因变量的效益图。每个因变量类别的曲线的显示与 ROC 曲线相同。
- **观察预测图。** 显示每个因变量的观察预测值图表。针对分类因变量，显示每个响应类别的预测拟概率的复式箱图，并且观察响应类别为聚类变量。针对刻度因变量，显示散点图。
- **残差分析图。** 显示每个刻度因变量的残差分析值图表。残差和预测值之间不存在可见模式。此图表仅针对刻度因变量生成。

**个案处理摘要。** 显示个案处理摘要表，其通过培训、检验和坚持样本整体总结分析中包含和排除的个案数。

**自变量重要性分析。** 执行敏感度分析，其计算确定神经网络的每个预测变量的重要性。分析以培训样本和检验样本组合为基础，或者如果不存在检验样本，则只以培训样本为基础。此操作创建一个显示每个预测变量的重要性和标准化重要性的表和图表。请注意，如果存在大量预测变量或个案，敏感度分析需要进行大量计算并且很费时。

## 保存

保存选项卡用于将预测变量另存为数据集中的变量。

- **保存各因变量的预测值或类别。** 此操作保存刻度因变量的预测值和分类因变量的预测类别。
- **保存各因变量的预测拟概率或类别。** 此操作保存分类因变量的预测拟概率。针对第一个 *n* 类别保存单个变量，其中在*要保存的类别列*已指定 **n**。

**保存的变量名称。** 自动名称生成确保能保存您的所有工作。无需先删除数据编辑器中保存的变量，自定义名称允许您放弃/替换上一次运行的结果。

概率和拟概率

具有 softmax 激活和交叉熵错误的分类因变量将拥有每个类别的预测值，其中每个预测值为个案属于类别的概率。

具有平方和误差的分类因变量将拥有每个类别的预测值，但预测值不能理解为概率。该过程保存这些预测拟概率，即使某些预测拟概率小于 0 或大于 1，或给定因变量的和不为 1。

基于拟概率创建 ROC、累积增益图和效益图（请参阅第 7 页的『输出』）。如果任何拟概率小于 0 或大于 1，或给定变量的和不为 1，首先会将其重标度为介于 0 和 1 之间且和为 1。通过除以它们的和来重标度拟概率。例如，如果一个个案具有三个分类因变量的预测拟概率 0.50、0.60、0.40，那么每个拟概率除以和 1.50 得 0.33、0.40 和 .27。

如果任何一个拟概率为负，那么在进行以上重标度之前，将最小数的绝对值添加到所有拟概率中。例如，如果拟概率为 -0.30、0.50 和 1.30，那么每个值先加 0.30 得 0.00、0.80 和 1.60。 然后，用每个新值除以和 2.40 得 0.00、0.33 和 0.67。

## 导出

导出选项卡用于将每个因变量的突触权重估算保存到 XML（PMML）文件中。您可以使用该模型文件以应用模型信息到其他数据文件用于评分目的。如果已经指定拆分文件，此选项不可用。

## 选项

**用户缺失值。** 要在分析中包含个案，因子必须具有有效值。通过这些控制可以决定是否将用户缺失值在因子变量和分类因变量中视为有效值。

**中止规则。** 这些是确定何时终止培训神经网的规则。培训至少继续一个数据传递。可以按照以下已在列举顺序中检查的条件终止培训。按中止规则定义，一步对应于在线和袖珍型批处理方法的数据传递以及一个批处理方法的迭代。

- **误差未减少情况下的最大步骤数。** 检查误差减少之前的步骤数。指定步骤数之后如果没有减少，那么培训停止。指定一个大于 0 的整数。您也可以指定用于计算错误的数据样本。如果其存在，**自动选择**将使用检验样本，否则将使用培训样本。请注意,批处理培训保证在每次数据传递之后减少培训样本错误；因此，如果检验样本存在，此选项只适用于批处理培训。**培训和检验数据**检查每个样本的错误；此选项仅在检验样本存在时适用。

  *注*: 每个数据传递完成之后，在线和袖珍型批处理培训需要一个额外数据传递以计算培训错误。额外数据传递可以明显减慢培训，所以一般推荐您提供检验样本，并在任何一个个案中选择**自动选择**。
- **最长培训时间。** 选择是否指定运行算法的最大分钟数。指定大于 0 的数。
- **最长培训时程。** 允许的最大时程数（数据传递）。如果超过最大时程数，则停止培训。指定大于 0 的整数。
- **培训错误中的最小相对变化。** 如果与前一步相比,培训错误中相对变化小于标准值，则培训停止。指定一个大于 0 的数。针对在线和袖珍型批处理培训，如果只有检验数据用于计算错误，忽略此标准。
- **培训误差率中的最小相对变化。** 如果培训错误与空模型错误的比率小于标准值，则培训停止。空模型预测所有因变量的平均值。指定一个大于 0 的数。针对在线和袖珍型批处理培训，如果只有检验数据用于计算错误，忽略此标准。

**内存中存储的最大个案数。**这控制以下多层感知器算法内的设置。指定大于 1 的整数。

- 在体系结构自动选择中，用于确定网络体系结构的样本的大小为 min(1000,*memsize*)，其中 *memsize* 是内存中存储的最大个案数。
- 在具有自动计算袖珍型批处理数的袖珍型批处理培训中，袖珍型批处理数为 min(max(*M*/10,2),*memsize*)，其中 *M* 是培训样本中的个案数。

# 第 3 章 径向基函数

径向基函数 (RBF) 过程会根据预测变量的值来生成一个或多个因变量（目标变量）的预测模型。

**示例。** 电信提供商按照服务用途模式划分客户群，将客户分类成四组。RBF 网络使用人口统计学数据预测组成员身份，这样可以使公司为各个潜在客户自定义服务。

数据注意事项

**因变量。** 因变量可以是:

- *名义 (Nominal).* 当变量值表示不具有内在等级的类别时，该变量可以作为名义变量；例如，雇员任职的公司部门。名义变量的示例包括地区、邮政编码和宗教信仰。
- *有序 (Ordinal).* 当变量值表示带有某种内在等级的类别时，该变量可以作为有序变量；例如，从十分不满意到十分满意的服务满意度水平。有序变量的示例包括表示满意度或可信度的态度分数和优先选择评分。
- *刻度 (Scale).* 当变量值表示带有有意义的度规的已排序类别时，该变量可以作为刻度（连续）变量对待，以便在值之间进行合适的距离比较。刻度变量的示例包括以年为单位的年龄和以千美元为单位的收入。

  该过程假设相应的测量级别已指定给所有因变量，尽管您可通过右键单击源变量列表中的变量并从弹出菜单中选择测量级别，以临时更改变量测量级别。

变量列表中每个变量旁的图标标识测量级别和数据类型:

*表 2. 测量级别图标*

| | 数值 | 字符串 | 日期 | 时间 |
|---|---|---|---|---|
| 刻度（连续） |  | n/a |  |  |
| 有序 |  |  |  |  |
| 名义 |  |  |  |  |

**预测变量。** 预测变量可指定为因子（分类）或协变量（刻度）。

**类别变量编码。** 该过程使用使用一个 *c* 编码在过程期间临时重新编码分类预测变量和因变量。如果存在 *c* 分类变量，那么该变量存储为 c 向量，第一个类别表示为（1、0、...、0）、下一个类别表示为（0、1、0、...、0）、...、最后一个类别表示为（0、0、...、0、1）。

此编码方案增加突触权重的数目并导致培训减速，但是多数“压缩”编码方法通常导致较差的拟合神经网络。如果您的网络培训进行很慢，尝试通过将类似的类别组合起来或删除具有极少见类别的个案以减少分类预测变量中的类别数目。

所有 *c* 之一的编码以培训数据为基础，即使已经定义检验或坚持样本（请参阅第 13 页的『分区』）。因此，如果检验或坚持样本包含培训数据中不存在的预测变量类别个案，那么那些个案不用于该过程或评分。如果检验或坚持样本包含培训数据中不存在的因变量类别个案，那么那些个案已经用于该过程，但可能被评分。

**重新调整。** 在缺省情况下，重新调整刻度因变量和协变量以改善网络培训。基于培训数据执行所有重标度，即使已经定义检验或坚持样本（请参阅第 13 页的『分区』）。也就是说，根据重标度的类型，仅使用培训数据计算平均值、标准差、协变量或因变量的最小值或最大值。如果您指定一个变量以定义分区，这些协变量或因变量在培训样本、检验样本或坚持样本之间具有相似分布将至关重要。

**频率权重。** 该过程忽略频率权重。

**复制结果。** 如果您想准确复制您的结果，除了使用相同过程设置以外，使用针对随机数字生成器的相同初始化值和相同数据顺序。有关此问题的更多详细信息，请参阅以下内容:

- **随机数字生成器。** 该过程在分区随机分配期间使用随机数字生成器。想要以后再次生成相同的随机结果，在每次运行径向基函数过程之前使用随机数字生成器的相同初始化值。请参阅 了解逐步操作说明。
- **个案顺序。** 结果也取决于数据顺序因为两步聚类算法用于确定径向基函数。

  要使顺序的影响降至最低程度，可随机个案等级排序的顺序。想要验证给定解的稳定性，您可能想要通过以不同随机顺序排序的案例来得到多个不同的解。在文件非常大的情况，可使用以不同随机顺序排序的个案样本运行多次。

创建一个径向基函数网络

从菜单中选择:

**分析 > Neural Networks > 径向基函数...**

1. 选择至少一个因变量。
2. 至少选择一个因子或协变量。

根据需要，在变量选项卡上您可以更改重标度协变量的方法。选项为:

- **标准化。** 减去平均值并除以标准差，($x$–mean)/$s$。
- **标准化。** 减去平均值并除以范围，($x$–min)/(max–min)。标准化值介于 0 和 1 之间。
- **调整标准化。** 减去最小值并除以范围所得到的调整版本，[2*($x$–min)/(max–min)]–1。调整后的标准化值介于 –1 和 1 之间。
- **无。** 无协变量重标度。

具有未知测量级别的字段

当数据集中的一个或多个变量（字段）的测量级别未知时，将显示测量级别警告。由于测量级别会影响该过程的计算结果，因此所有变量必须都定义有测量级别。

**扫描数据。** 读取活动数据集中的数据，并分配缺省测量级别给任何具有当前未知测量级别的字段。如果数据集较大，该过程可能需要一些时间。

**手动分配。** 打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框将测量级别分配给这些字段。您也可以在数据编辑器的变量视图中分配测量级别。

由于测量级别对该过程很重要，因此您无法访问运行该过程的对话框，除非所有字段均定义了测量级别。

## 分区

**分区数据集。** 此组指定将活动数据集划分为训练样本、检验样本或坚持样本的方法。**训练样本**包含用于训练神经网络的数据记录；数据集中的某些个案百分比必须分配给训练样本以获得一个模型。**检验样本**是一个用于跟踪训练过程中的错误以防止超额训练的独立数据记录集。强烈建议您创建一个训练样本，并且如果测试样本小于训练样本，网络训练通常最高效。**坚持样本**是另一个用于评估最终神经网络的独立数据记录集；坚持样本的误差给出一个模型预测能力的"真实"估计值，因为坚持个案不用于构建模型。

- **根据个案的相对数量随机分配个案。** 指定随机分配到每个样本（训练、检验和坚持）的个案的相对数量（比率）。**%** 列根据您已经指定的相对数量，报告将被分配到每个样本的个案的百分比。

  例如，指定 7、3、0 作为训练、检验和坚持样本的相对数量对应于 70%、30% 和 0%。指定 2、1、1 作为相对数量对应 50%、25% 和 25%；1、1、1 对应将数据集在训练、检验和坚持中分为相等的三部分。
- **使用分区变量分配个案。** 指定一个将活动数据集中的每个个案分配到训练、检验和坚持样本中的数值变量。变量为正值的个案被分配到训练样本中，值为 0 的个案被分配到检验样本中，而负值个案被分配到坚持样本中。具有系统缺失值的个案会从分析中排除。分区变量的任何用户缺失值始终视为有效。

## 体系结构

"体系结构"选项卡用于指定网络结构。该过程创建一个有隐藏"径向基函数"层的神经网络；通常，不需要更改这些设置。

**隐藏层中的单位数。** 选择隐藏单位数有三种方式。

1. **在某个自动计算范围内查找最佳单位数。** 该过程自动计算范围的最小值和最大值并在该范围内查找最佳隐藏单位数。

   如果已定义检验样本，则该过程使用检验数据标准：隐藏单位的最佳数量是指在检验数据中造成最少错误的数量。如果未定义检验样本，则该过程使用贝叶斯信息标准 (BIC)：隐藏单位的最佳数量是指对培训数据造成最少 BIC 的数量。
2. **在某个指定范围内查找最佳单位数。** 您可以提供自己的范围，并且该过程会在那个范围内查找"最佳"隐藏单位数。和以前一样，该范围中最佳隐藏单位数通过使用检验数据标准或 BIC 准则来确定。
3. **使用指定的单位数。** 您可以覆盖某个范围的使用并直接指定特定数量的单位。

**隐藏层激活函数。** 隐藏层激活函数是径向基函数，它将某个层中的单位"关联"到下一层的单位值。对于输出层，激活函数是恒等函数，因此输出单位仅仅是隐藏单位的权重总和。

- **标准化径向基函数。** 使用 softmax 激活函数以使所有隐藏单位的激活都标准化合计为 1。
- **一般径向基函数。** 使用指数激活函数，因此隐藏单位激活是作为输入函数的高斯"增加"。

**隐藏单位中的重叠。** 重叠因子是应用到径向基函数宽度的乘数。重叠因子的自动计算值为 1+0.1*d*，其中 *d* 是输入单位数（所有因子类别数量和协变量数量之和）。

## 输出

**网络结构。** 显示与神经网络有关的摘要信息。

- **描述。** 显示与神经网络有关的信息，包括因变量、输入和输出单位数目、隐藏层和单位数目及激活函数。
- **图表。** 将神经网络图表作为不可编辑图表显示。请注意，随着协变量数目和因子级别的增加，图表变得更加难于解释。

- **突触权重。** 显示表明给定层中的单位与以下层中的单位之间关系的系数估计值。突触权重以培训样本为基础，即使活动数据集已划分为培训数据、检验数据和坚持数据。请注意，突触权重数目会变得非常大，而且这些权重一般不用于解释网络结果。

**网络性能。** 显示用于确定模型是否"良好"的结果。*注*: 该组中的图表以培训样本和检验样本组合为基础，或者如果不存在检验样本，只以培训样本为基础。

- **模型摘要。** 显示分区和整体神经网络结果摘要，包括错误、相对错误或不正确预测的百分比和培训时间。

  误差为平方和误差。除此之外，显示相对错误或不正确预测的百分比取决于因变量测量级别。如果任何因变量具有刻度测量级别，则显示平均整体相对错误（相对于平均值模型）。如果所有因变量都为分类变量，则显示不正确预测的平均百分比。也针对单个因变量显示相对错误或不正确预测的百分比。
- **分类结果。** 显示每个分类因变量的分类表。每个表针对每个因变量类别给出正确或错误分类的个案数目。也报告正确分类的总体个案百分比。
- **ROC 曲线。** 显示每个分类因变量的 ROC（Receiver Operating Characteristic）曲线。其也显示一个给定每个曲线下区域的表格。对于给定因变量，ROC 图表针对每个类别显示一条曲线。如果因变量有两个类别，那么每条曲线将该类别视为正态与其它类别。如果因变量有两个多类别，那么每条曲线将该类别视为正态与所有其它类别的汇总。
- **累积增益图。** 显示每个分类因变量的累积增益图。每个因变量类别的曲线的显示与 ROC 曲线相同。
- **效益图。** 显示每个分类因变量的效益图。每个因变量类别的曲线的显示与 ROC 曲线相同。
- **观察预测图。** 显示每个因变量的观察预测值图表。针对分类因变量，显示每个响应类别的预测拟概率的复式箱图，并且观察响应类别为聚类变量。针对刻度因变量，显示散点图。
- **残差分析图。** 显示每个刻度因变量的残差分析图。残差和预测值之间不存在可见模式。此图表仅针对刻度因变量生成。

**个案处理摘要。** 显示个案处理摘要表，其通过培训、检验和坚持样本整体总结分析中包含和排除的个案数。

**自变量重要性分析。** 执行敏感度分析，其计算确定神经网络的每个预测变量的重要性。分析以培训样本和检验样本组合为基础，或者如果不存在检验样本，只以培训样本为基础。此操作创建一个显示每个预测变量的重要性和标准化重要性的表和图表。请注意，如果存在大量预测变量和个案，敏感度分析需要进行大量计算并且费时。

## 保存

保存选项卡用于将预测变量另存为数据集中的变量。

- **保存各因变量的预测值或类别。** 此操作保存刻度因变量的预测值和分类因变量的预测类别。
- **为各因变量保存预测拟概率。** 此操作保存分类因变量的预测拟概率。针对第一个 *n* 类别保存单个变量，其中在*要保存的类别*列已指定 *n*。

**保存的变量名称。** 自动名称生成确保能保存您的所有工作。无需先删除数据编辑器中保存的变量，自定义名称允许您放弃或替换上一次运行的结果。

概率和拟概率

预测拟概率无法解释为概率，因为径向基函数过程使用输出层的平方和误差和恒等激活函数。即使存在小于 0 或大于 1 的预测拟概率，或给定因变量的和不为 1，该过程仍将保存这些预测拟概率。

基于拟概率创建 ROC、累积增益图和效益图（请参阅第 13 页的『输出』）。如果任何拟概率小于 0 或大于 1，或给定变量的和不为 1，首先会将其重标度为介于 0 和 1 之间且和为 1。通过除以它们的和来重标度拟概率。例如，如果一个个案具有三个分类因变量的预测拟概率 0.50、0.60、0.40，那么每个拟概率除以和 1.50 得 0.33、0.40 和 .27。

如果任何一个拟概率为负，那么在进行以上重标度之前，将最小数的绝对值添加到所有拟概率中。例如，如果拟概率为 -0.30、0.50 和 1.30，那么每个值先加 0.30 得 0.00、0.80 和 1.60。然后，用每个新值除以和 2.40 得 0.00、0.33 和 0.67。

## 导出

导出选项卡用于将每个因变量的突触权重估算保存到 XML（PMML）文件中。您可以使用该模型文件以应用模型信息到其他数据文件用于评分目的。如果已经指定拆分文件，此选项不可用。

## 选项

**用户缺失值。** 要在分析中包含个案，因子必须具有有效值。通过这些控制可以决定是否将用户缺失值在因子变量和分类因变量中视为有效值。

# 声明

本信息是为在美国提供的产品和服务编写的。本资料的其他语言版本可以从 IBM 获取。但是，您可能需要拥有该语言的产品副本或产品版本才能访问这些资料。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您当前所在区域的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。不过，用户应自行负责评估和验证任何非 IBM 产品、程序或服务的工作情况。

IBM 公司可能已拥有或正在申请与本文档内容有关的各项专利。您获得该文档并不意味着授予您任何这些专利许可。您可以将书面许可查询函件发送至:

*IBM Director of Licensing*
*IBM Corporation*
*North Castle Drive, MD-NC119*
*Armonk, NY 10504-1785*
*US*

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往:

*Intellectual Property Licensing*
*Legal and Intellectual Property Law*
*IBM Japan Ltd.*
*19-21, Nihonbashi-Hakozakicho, Chuo-ku*
*Tokyo 103-8510, Japan*

International Business Machines Corporation“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某种特定用途的保证。某些管辖区域在某些交易中不允许免除明示或暗含的保证。 因此本条款可能不适用于您。

本信息可能含有技术误差或排版错误。此处的信息会定期进行更改；这些更改会体现在本出版物的新版本中。IBM 可以随时对本资料中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的， 不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的: (i) 允许在独立创建的程序和其他程序（包括本程序）之间进行信息交换，以及 (ii) 允许对已经交换的信息进行相互使用，请与下列地址联系:

*IBM Director of Licensing*
*IBM Corporation*

*North Castle Drive, MD-NC119*
*Armonk, NY 10504-1785*
*US*

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

本资料中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际软件许可协议或任何同等协议中的条款提供。

所引用的性能数据和客户示例只用于阐述说明。根据具体配置和操作条件，实际性能结果可能有所不同。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

有关 IBM 未来方向或意向的声明均可能未经通知即变更或撤销，并且仅代表目标和目的。

本信息包含日常业务运营中使用的数据和报告的示例。为了尽可能详尽地对其进行说明，示例中包含了人员的姓名、公司、品牌和产品的名称。所有这些名称都是虚构的，如果与实际人员或公司企业有任何类似则纯属巧合。

版权许可:

本信息包括源语言形式的样本应用程序，这些样本说明不同操作平台上的编程方法。如果是为按照在编写样本程序的操作平台上的应用程序编程接口 (API) 进行应用程序的开发、使用、经销或分发为目的，您可以任何形式对这些样本程序进行复制、修改、分发，而无须向 IBM 付费。这些示例并未在所有条件下作全面测试。因此，IBM 不能担保或暗示这些程序的可靠性、可维护性或功能。本样本程序仍然是“按现状”提供的，不附有任何种类的保证。对于因使用样本程序所引起的任何损害，IBM 概不负责。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明:

© （贵公司的名称）（年）。此部分代码是根据 IBM Corp. 公司的样本程序衍生出来的。

© Copyright IBM Corp. _（输入年份）_. All rights reserved.

## 商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域注册的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。Web 站点 www.ibm.com/legal/copytrade.shtml 上的“Copyright and trademark information”部分中提供了 IBM 商标的最新列表。

Adobe、Adobe 徽标、PostScript 和 PostScript 徽标是 Adobe Systems Incorporated 在美国和/或其他国家或地区的注册商标或商标。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国、其他国家或地区或两者的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国、其他国家或地区或两者的商标。

UNIX 是 The Open Group 在美国和其他国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Oracle 及/或其分支机构的商标和注册商标。

# 索引

Printed in China