IBM SPSS Neural Networks 21

注意：使用本信息及其支持的产品之前，请阅读注意事项第 83 页码下的一般信息。

此版本适用于 IBM® SPSS® Statistics 21 及所有后续发布和修订，除非在新版本中另有说明。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。

前言

IBM® SPSS® Statistics 是一种用于分析数据的综合系统。神经网络 可选附加模块提供本手册中描述的其他分析方法。此 神经网络 附加模块必须与 SPSS Statistics Core 系统一起使用，并已完全集成到了该系统中。

关于 IBM Business Analytics

IBM Business Analytics 软件提供决策者赖以提高业务绩效的完整、一致和准确的信息。包括业务智能、预测分析、财务状况和战略管理以及分析应用程序在内的一整套产品组合让您即刻、清楚地了解当前绩效并依此采取行动，以及能够预测未来的成果。结合丰富的行业解决方案、被证明的实践经验和专业的服务，无论公司规模大小，都能促使其获得最高的产能、自信自觉地做出决定并得到更好的成绩。

作为产品组合的一部分，IBM SPSS Predictive Analytics 软件帮助公司预测未来实践并采取积极行动，促使其获得更好的业务成果。全世界的商业政府和学术客户依赖 IBM SPSS 技术，因其具有竞争力的优势，能够吸引、留住和发展客户，同时减少欺诈和减轻风险。通过将 IBM SPSS 软件融入日常运营中，公司成为具有预测性的企业，能够引导和自觉做出决策，以满足业务目标，实现可观的竞争优势。欲知更多信息或联系代表，请访问 http://www.ibm.com/spss。

技术支持

我们提供有“技术支持”以维护客户。客户可就 IBM Corp. 产品使用或某一受支持硬件环境的安装帮助寻求技术支持。如要联系技术支持，请访问 IBM Corp. 网站，网址为 http://www.ibm.com/support。在请求协助时，请准备好您和您组织的 ID 以及支持协议。

针对学生的技术支持

如果您是使用任何学生版 IBM SPSS 软件产品的学生，请访问我们专为学生提供的在线教育解决方案 (http://www.ibm.com/spss/rd/students/)页面。如果您是使用大学提供的 IBM SPSS 软件副本的学生，请联系所在大学的 IBM SPSS 产品协调员。

客户服务

如果对发货或帐户存在任何问题，请联系您当地的办事处。请先准备好您的序列号以供识别。

培训讲座

IBM Corp. 提供公开的以及现场的培训讲座。所有讲座都是以实践小组为特色的。讲座将定期在各大城市开展。有关这些讲座的更多信息，请前往 http://www.ibm.com/software/analytics/spss/training。

内容

附录

部分 I:
用户指南

章 1

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(Haykin, 1998)：

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。
- 称为键结值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，请参见 (Ripley, 1996)。

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。

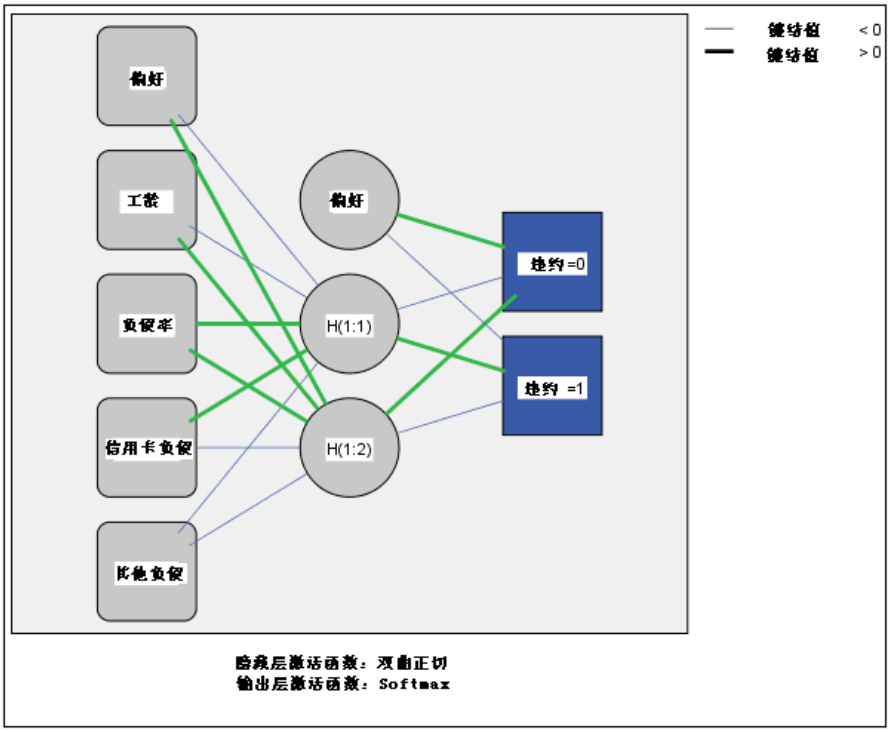
此灵活性的平衡指神经网络的键结值不可轻松解释。因此，如果您正试图解释生成因变量和自变量间关系的基础过程，最好使用更传统的统计模型。但是，如果模型的可解释性并不重要，您可以使用神经网络更快获取良好模型结果。

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欠贷者。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。

图片 1-1
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。每个隐藏单元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都是第一个隐藏层单元的一个函数，并且每个响应都是第二个隐藏层单元的一个函数。

章 2

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。

因变量。 因变量可以是：

- **标定.** 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.** 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.** 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上下文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：

	数值	字符串	日期	时间
尺度（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 矢量，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）。

此编码设计增加了键结值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（请参见 分区 第 7 页码）。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。 在默认情况下，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（请参见 分区 第 7 页码）。也就是说，根据重标度的类型，仅使用培训数据计算均值、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。有关此问题的详情，请参见以下内容：

- **随机数字生成器。** 该过程在分区随机分配、键结值初始化的随机子样本、自动体系结构选择的随机子样本、用于权重初始化和自动体系结构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。请参见 准备数据以进行分析第 31 页码 了解逐步操作说明。
- **个案顺序。** 在线和袖珍型批处理培训方法（请参见 培训 第 10 页码）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，因为键结值初始化包含数据集的子样本。

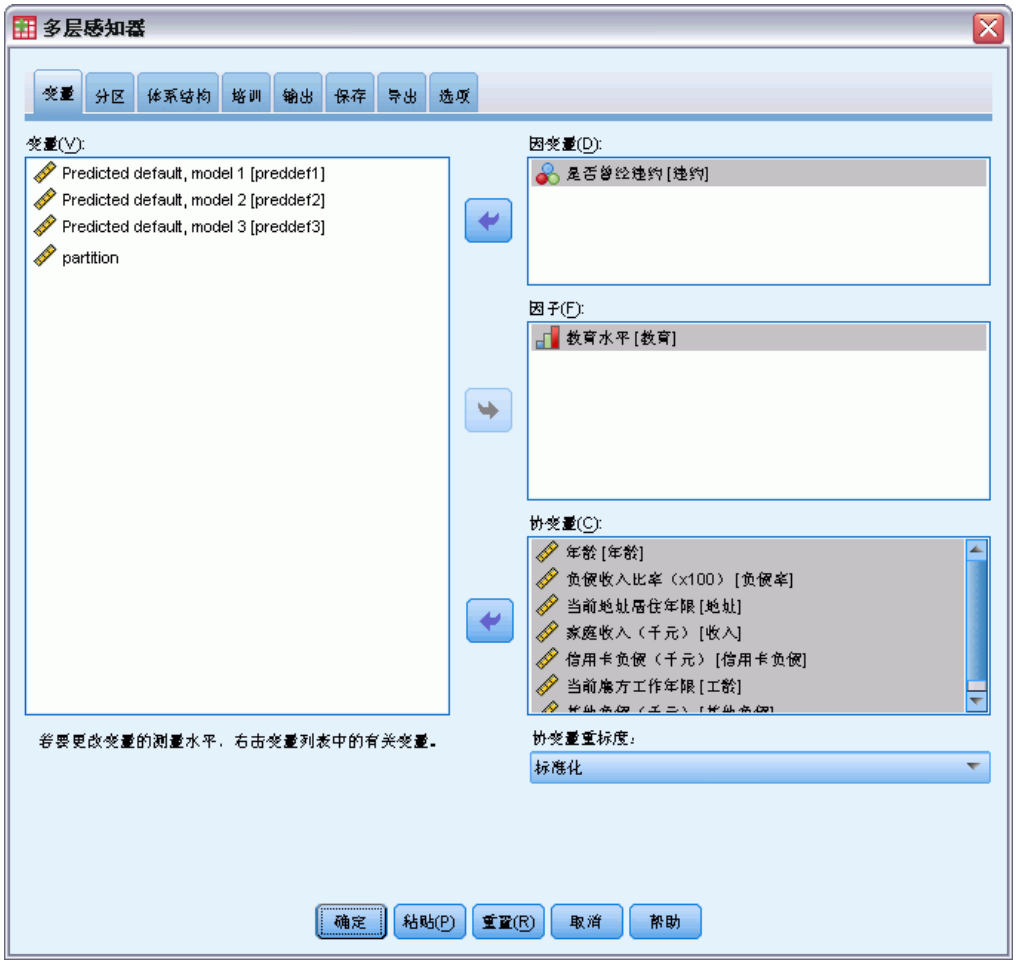
  要使顺序的影响降至最低程度，可随机排列个案的顺序。想要验证给定解的稳定性，您可能想要通过以不同随机顺序排序的案例来得到多个不同的解。在文件非常大的情况，可使用以不同随机顺序排序的个案样本运行多次。
- **变量顺序。** 由于变量顺序改变时分配了不同模式的初始值，结果可能会受到因子和协变量列表中的变量顺序的影响。因为个案顺序影响，您可能要尝试不同变量顺序（只需在因子或协变量列表中拖放）以评估给出解的稳定性。

创建多层感知器网络

从菜单中选择：

分析 > 神经网络 > 多层感知器...

图片 2-1
多层感知器：“变量”选项卡



- 选择至少一个因变量。
- 至少选择一个因子或协变量。

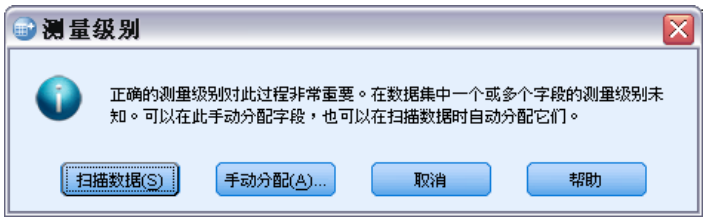
根据需要，在变量选项卡上您可以更改重标度协变量的方法。选项为：

- **标准化。** 减去均值并除以标准差，(x−均值)/s。
- **标准化。** 减去均值并除以范围，(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。

图片 2-2
测量级别警报

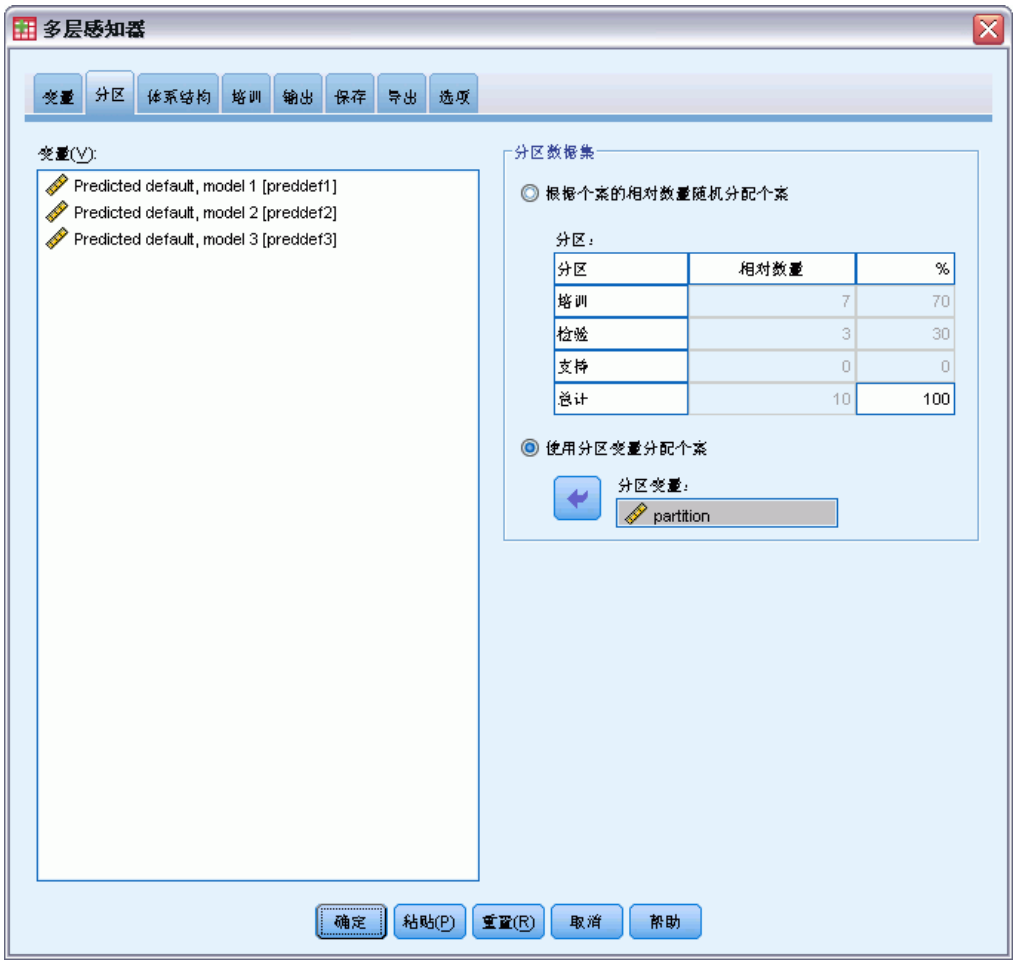


- **扫描数据。** 读取活动数据集中的数据，并分配默认测量级别给任何具有当前未知测量级别的字段。如果数据集较大，该过程可能需要一些时间。
- **手动分配。** 打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框将测量级别分配给这些字段。您也可以在数据编辑器的变量视图中分配测量级别。

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

图片 2-3
多层感知器：“分区”选项卡



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。

  注意：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。请参见主多层感知器主题中的“复制结果”。

体系结构

图片 2-4
多层感知器：体系结构选项卡

“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。

自动体系结构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，自动体系结构选择计算隐藏层中的“最佳”单位量。自动体系结构选择使用隐藏层和输出层的默认激活函数。

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自动体系结构选择的结果时，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之和的函数。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. 激活函数将某个层中的单位的加权和“关联”到下一层的单位值。

- **双曲正切。** 此函数格式：$\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$。其取实数值参数并将其变换到（-1、1）范围。使用自动体系结构选择时，此为隐藏层所有单位的激活函数。
- **Sigmoid。** 此函数格式：$\gamma(c) = 1/(1+e^{-c})$。其取实数值参数并将其变换到（0、1）范围。

单位数. 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数。

输出层

输出层包含目标（因）变量。

激活函数. 激活函数将某个层中的单位的加权和“关联”到下一层的单位值。

- **恒等。** 此函数格式：$\gamma(c) = c$。其取实数值参数并且其返回值保持不变。使用自动体系结构选择时，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。** 此函数格式：$\gamma(c_k) = \exp(c_k)/\Sigma_j\exp(c_j)$。其取实数值参数的矢量，并将其变换到元素介于（0、1）范围的矢量，和为 1。只有所有因变量是分类变量时，才可以使用 Softmax。使用自动体系结构选择时，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。** 此函数格式：$\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$。其取实数值参数并将其变换到（-1、1）范围。
- **Sigmoid。** 此函数格式：$\gamma(c) = 1/(1+e^{-c})$。其取实数值参数并将其变换到（0、1）范围。

尺度因变量重标度。 至少选择一个刻度因变量时才可以使用这些控制。

- **标准化。** 减去均值并除以标准差，(x-均值)/s。
- **标准化。** 减去均值并除以范围，(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 ε，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− ε)]/[(max+ ε)−(min− ε)]。请指定大于等于 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 ε，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− ε))/((max+ ε)−(min− ε))]}−1。指定一个大于或等于 0 的数字。
- **无。** 未对刻度因变量进行重标度。

培训

图片 2-5
多层感知器：“培训”选项卡

“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键结值；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键结值；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键结值；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 并小于或等于将存储到内存的个案的最大数目的整数。您可以在选项选项卡上设置将存储到内存的个案最大数目。

优化算法。 这是一种用于估计键结值的方法。

- **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 Sigma 值。** 针对调整的共轭梯度算法的 sigma 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矢量。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自动体系结构选择。指定间隔中心点数目且该数大于间隔偏移量 0。

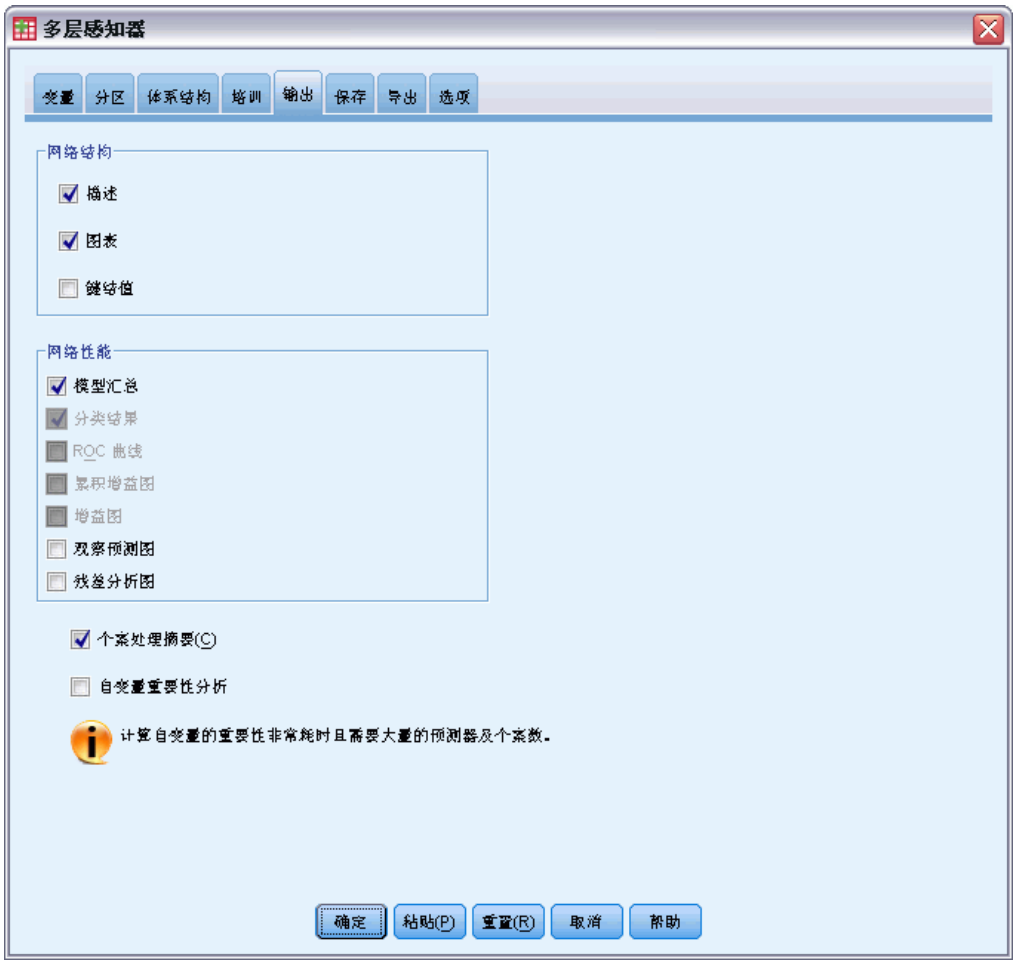
梯度下降算法的培训选项包括：

- **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。
- **学习率的较低极限。** 针对梯度下降算法的学习率较低极限。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极限。这使您能控制学习率衰减因子 $\beta = (1/pK)*\ln(\eta_0/\eta_{low})$，其中 η_0 是初始学习率，η_{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

图片 2-6
多层感知器：“输出”选项卡



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。
- **键结值。** 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。键结值以培训样本为基础，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，键结值数目会变得非常大，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。注意：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，错误为平方和错误。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。
- **增益图。** 显示每个分类因变量的增益图。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分群变量。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

图片 2-7
多层感知器：“保存”选项卡

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。

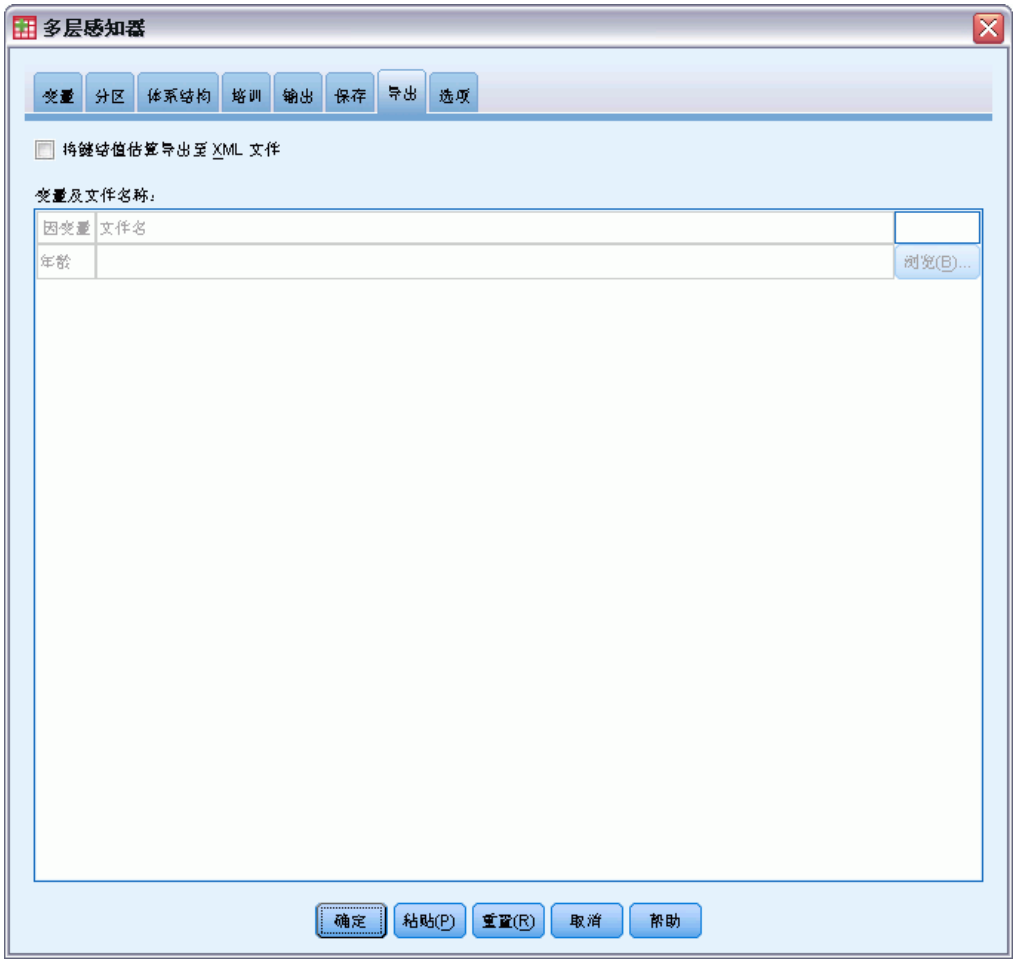
具有平方和错误的分类因变量将拥有每个类别的预测值，但预测值不能理解为概率。该过程保存这些预测拟概率，即使某些预测拟概率小于 0 或大于 1，或给定因变量的和不为 1。

基于拟概率创建 ROC、累积增益图和增益图（请参见 输出 第 12 页码）。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、.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

图片 2-8
多层感知器：“导出”选项卡



导出选项卡用于将每个因变量的键结值估算保存到 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

图片 2-9
多层感知器：“选项”选项卡

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。

注意：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。
- **培训错误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。

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

- 在自动体系结构选择中，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。

章 3

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。

因变量。 因变量可以是：

- **标定**. 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**. 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**. 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上下文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：

	数值	字符串	日期	时间
尺度（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 矢量，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）。

此编码设计增加键结值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（请参见 分区 第 22 页码）。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。 在默认情况下，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（请参见 分区 第 22 页码）。也就是说，根据重标度的类型，仅使用培训数据计算均值、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。有关此问题的详情，请参见以下内容：

- **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。请参见 准备数据以进行分析第 64 页码 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。

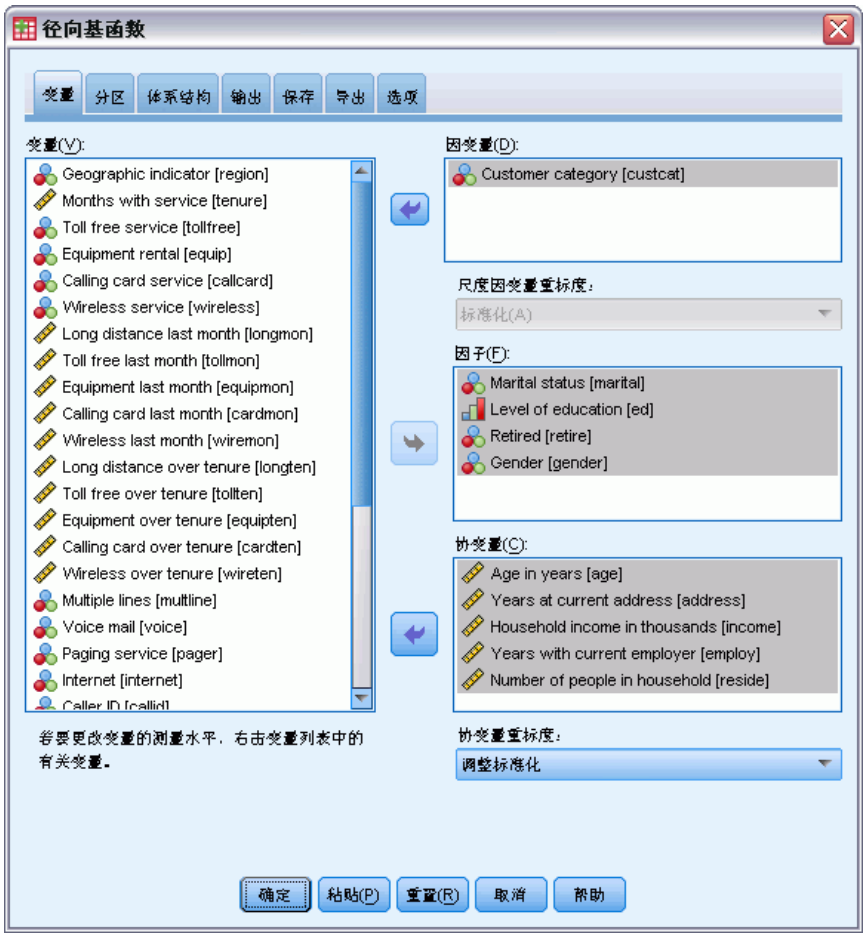
  要使顺序的影响降至最低程度，可随机排列个案的顺序。想要验证给定解的稳定性，您可能想要通过以不同随机顺序排序的案例来得到多个不同的解。在文件非常大的情况，可使用以不同随机顺序排序的个案样本运行多次。

创建一个径向基函数网络

从菜单中选择：

分析 > **神经网络** > **径向基函数...**

图片 3-1
径向基函数：“变量”选项卡



- 选择至少一个因变量。
- 至少选择一个因子或协变量。

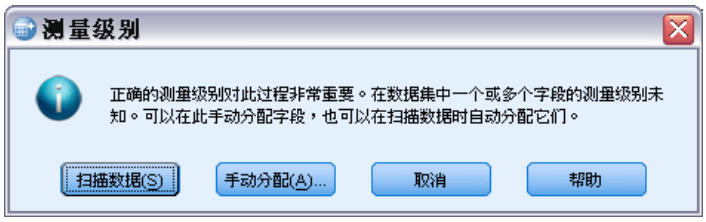
根据需要，在变量选项卡上您可以更改重标度协变量的方法。选项为：

- **标准化。** 减去均值并除以标准差，(x–均值)/s。
- **标准化。** 减去均值并除以范围，(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。

图片 3-2
测量级别警报



- **扫描数据。** 读取活动数据集中的数据，并分配默认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

图片 3-3
径向基函数：“分区”选项卡

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

图片 3-4
径向基函数：体系结构选项卡

“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。

如果定义了一个检验样本，则该过程使用检验数据标准：隐藏单位的最佳数量为检验数据中产生最小错误的单位。如果未定义检验样本，则该过程使用 BIC 准则：隐藏单位的最佳数量为基于培训数据产生最小 BIC 的单位。

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加权和。

- **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+0.1d，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

图片 3-5
径向基函数：“输出”选项卡

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。
- **键结值。** 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。键结值以培训样本为基础，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，键结值数目会变得非常大，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。注意：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。
- **增益图。** 显示每个分类因变量的增益图。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分群变量。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

图片 3-6
径向基函数：“保存”选项卡

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、累积增益图和增益图（请参见 输出 第 25 页码）。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、.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

图片 3-7
径向基函数：“导出”选项卡

导出选项卡用于将每个因变量的键结值估算保存到 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

图片 3-8
径向基函数：“选项”选项卡

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。

部分 II:
示例

章 4

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。

使用多层感知器评估信用风险

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。

假设 850 名以往客户和潜在客户的信息包含在 bankloan.sav 中。有关详细信息，请参阅第 76 页码附录 A 中的样本文件。前 700 个个案是以前曾获得贷款的客户。请使用这 700 名客户的随机样本创建多层感知器，而留出其余客户用于验证分析。然后使用该模型将 150 名潜在客户按高或低信用风险分类。

此外，信贷员先前使用 Logistic 回归（在“回归”选项中）分析数据并考虑多层感知器如何作为分类工具。

准备数据以进行分析

通过设置随机数种子您可以精确复制此分析。

- 要设置随机数种子，请从菜单中选择：
  转换 > **随机数字生成器...**

图片 4-1
“随机数生成器”对话框

- 选择**设置起点**。
- 选择**固定值**并键入 9191972 作为值。
- 单击**确定**。

在之前 Logistic 回归分析中，大约 70% 以往客户被分配至训练样本，30% 被分配至坚持样本。将需要分区变量精确地重新创建用于那些分析的样本。

▶ 要创建分区变量，请从菜单中选择：
转换 > 计算变量...

图片 4-2
“计算变量”对话框

▶ 在“目标变量”文本框中键入**分区**。

▶ 在“数值表达式”文本框中键入 `2*rv.bernoulli(0.7)-1`。

此操作将分区值设置为随机生成的概率参数为 0.7 的 Bernoulli 变量，修改之后取值 1 或 −1，而不是 1 或 0。调用将分区变量上为正值的个案分配给训练样本，将负值个案分配给坚持样本，将值为 0 的个案分配给检验样本。现在我们不指定检验样本。

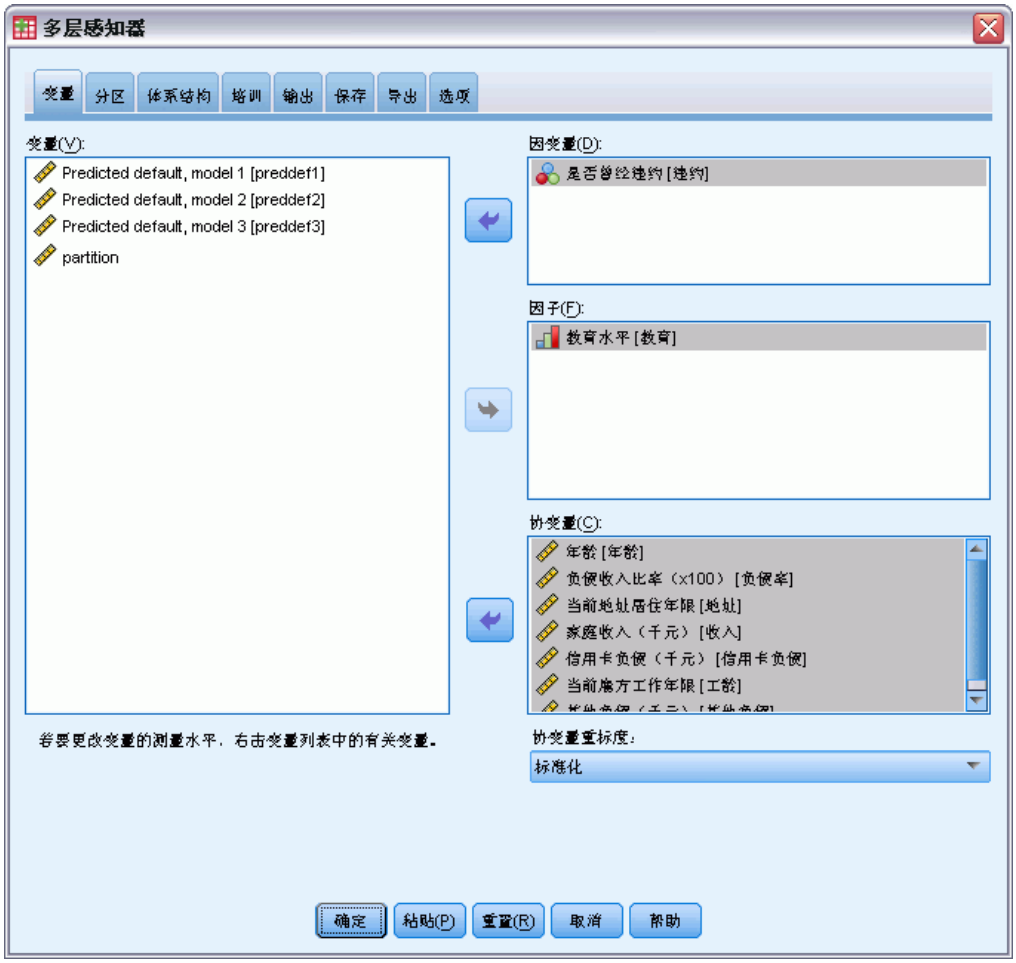
▶ 在“计算变量”对话框中单击**确定**。

约 70% 以前曾获得贷款的客户的分区值将为 1。这些客户将用于创建模型。以前曾获得贷款的其他客户的分区值将为−1，并将用于验证模型结果。

运行分析

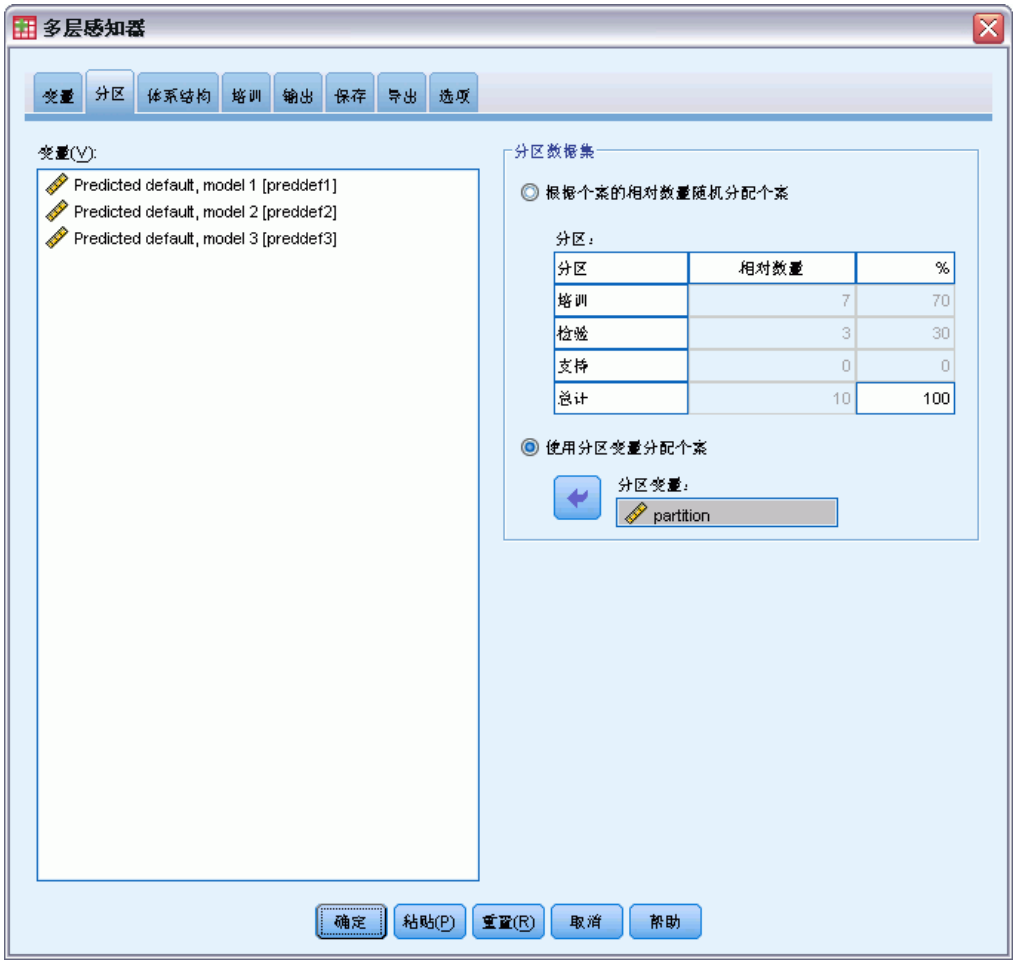
- 要运行“多层感知器”分析，请从菜单中选择：
  分析 > 神经网络 > 多层感知器...

图片 4-3
多层感知器：“变量”选项卡



- 选择 Previously defaulted 作为因变量。
- 选择 Level of education [ed] 作为因子。
- 选择 Age in years [age] 到 Other debt in thousands [othdebt] 作为协变量。
- 单击分区选项卡。

图片 4-4
多层感知器：“分区”选项卡



- 选择**使用分区变量分配个案**。
- 选择分区作为分区变量。
- 单击**输出**选项卡。

图片 4-5
多层感知器：“输出”选项卡

- 在“网络结构”组选项中取消选择**图表**。
- 在“网络性能”组选项中选择 **ROC 曲线**、**累积增益图**、**增益图**和**观察预测图**。因为因变量非刻度变量，残差分析图不可用。
- 选择**自变量重要性分析**。
- 单击**确定**。

个案处理摘要

图片 4-6
个案处理摘要

		N	百分比
样本	培训	499	71.3%
	保留	201	28.7%
有效		700	100.0%
排除		150	
合计		850	

个案处理摘要显示 499 个案已分配给训练样本，201 个案分配给坚持样本。不包括在分析中的 150 个个案为潜在客户。

网络信息

图片 4-7
网络信息

输入层	因子	1	教育水平
	协变量	1	年龄
		2	当前雇方工作年限
		3	当前地址居住年限
		4	家庭收入（千元）
		5	负债收入比率（x100）
		6	信用卡负债（千元）
		7	其他负债（千元）
	单位数		12
	协变量的重标度方法		标准化
隐藏层	隐藏层数		1
	隐藏层 1 中的单位数		4
	激活函数		双曲正切
输出层	因变量	1	是否曾经违约
	单位数		2
	激活函数		Softmax
	错误函数		交叉熵

网络信息表显示有关神经网络的信息，它对于确保指定正确很有用。此处特别要注意的是：

- 输入层的单元数为协变量数加上因子水平总数；为 Level of education 的每个类别创建单独单元，并且没有任何类别被视为在许多建模过程中典型的“冗余”单元。
- 同样，为 Previously defaulted 的每个类别创建一个单独输出单元，输出层共有两个单元。
- 自动体系结构选择选择了隐藏层中的四个单元。
- 所有其他网络信息都是过程的缺省值。

模型摘要

图片 4-8
模型摘要

训练	交叉熵错误	156.606
	百分比错误预测	15.6%
	中止使用的规则	已超过的最大时程数 (100)
	培训时间	0:00:00.765
保持	百分比错误预测	25.4%

因变量:是否曾经违约

模型摘要显示与训练结果及将最终网络应用于坚持样本相关的信息。

- 因为输出层使用 softmax 激活函数，将显示交叉熵错误。这是网络试图在训练中最小化的错误函数。
- 错误预测值的百分比取自分类表，并将在该主题中作进一步讨论。
- 因为达到最大时程数，所以估计算法停止。理想情况下，因为错误收敛，所以训练应停止。这提出了关于训练中是否出现错误的问题，并且成为在进一步检查输出时需谨记的事项。

Classification

图片 4-9
Classification

		已预测		
样本	已观测	否	是	正确百分比
训练	否	347	28	92.5%
	是	50	74	59.7%
	总计百分比	79.6%	20.4%	84.4%
保持	否	123	19	86.6%
	是	32	27	45.8%
	总计百分比	77.1%	22.9%	74.6%

因变量:是否曾经违约

分类表显示使用网络的实际结果。对于每个个案，如果该个案的预测拟概率大于 0.5，则预测响应为是。对于每个样本：

- 个案交叉分类对角线上的单元格是正确的预测值。
- 个案交叉分类偏离对角线的单元格是不正确的预测值。

在用于创建模型的个案中，以前拖欠贷款的 124 人中有 74 人分类正确。375 名未欠贷者中有 347 人分类正确。整体上，84.4% 训练个案分类正确，与模型摘要表中 15.6% 显示不正确项相对应。更好的模型应正确识别出更高百分比的个案。

基于创建模型所用个案的分类从其分类率有所夸大的意义上来说，倾向于过度“乐观”。保持样本帮助验证模型；这些个案中，有 74.6% 是由模型正确分类的。这意味着，总体来说，您的模型实际上有七五成是正确的。

矫正超额训练

回顾之前执行的 logistic 回归分析，信贷员调用正确预测个案相似百分比约为 80% 的训练和支持样本。相比之下，神经网络拥有更高百分比的正确训练样本个案，而支持样本在预测实际拖欠贷款的客户时表现相对较差（支持样本 45.8% 正确对比训练样本 59.7% 正确 ）。与模型摘要表中报告的中止规则结合，这让您怀疑网络可能**超额训练**；即，其根据随机变化搜寻显示在培训数据中的虚假模式。

不过解决方案相对简单：指定一个检验样本来帮助保持网络“正常运行”。我们创建分区变量以便其精确地重新创建用于 Logistic 回归分析的训练和支持样本；但是，Logistic 回归没有“检验”样本的概念。让我们应用部分训练样本并将其重新指派给检验样本。

创建检验样本

图片 4-10
“计算变量”对话框

- 调用“计算变量”对话框。
- 在“数值表达式”文本框中键入 `partition - rv.bernoulli(0.2)`。
- 单击如果。

图片 4-11
计算变量：“If 个案”对话框

- 选中**如果个案满足条件则包含**。
- 在文本框中键入**分区>0**。
- 单击**继续**。
- 在“计算变量”对话框中单击**确定**。

这将重置大于 0 的分区值，可以使约 20% 取值 0，并且 80% 保持值 1。整体上，大约 100*(0.7*0.8)=56% 以前曾获得贷款的客户将为训练样本，14% 为检验样本。原先分配到坚持样本的客户仍然在那里。

运行分析

- 调用多层感知器对话框并单击**保存**选项卡。
- 为各因变量选择**保存预测拟概率**。
- 单击**确定**。

个案处理摘要

图片 4-12
检验样本模型的个案处理摘要

		N	百分比
样本	培训	398	56.9%
	检验	101	14.4%
	保留	201	28.7%
有效		700	100.0%
排除		150	
合计		850	

在原先指定到训练样本的 499 个案中，101 个案已重新指定到检验样本。

网络信息

图片 4-13
网络信息

输入层	因子	1	教育水平
	协变量	1	年龄
		2	当前雇方工作年限
		3	当前地址居住年限
		4	家庭收入（千元）
		5	负债收入比率（x100）
		6	信用卡负债（千元）
		7	其他负债（千元）
	单位数		12
	协变量的重标度方法		标准化
隐藏层	隐藏层数		1
	隐藏层 1 中的单位数		7
	激活函数		双曲正切
输出层	因变量	1	是否曾经违约
	单位数		2
	激活函数		Softmax
	错误函数		交叉熵

网络信息表的唯一变化是自动体系结构选择已选择隐藏层中的七个单元。

模型摘要

图片 4-14
模型摘要

训练	交叉熵错误	159.870
	百分比错误预测	20.1%
	中止使用的规则	错误未减少的 1 连续步骤[a]
	培训时间	0:00:00.750
测试	交叉熵错误	40.068
	百分比错误预测	17.8%
保持	百分比错误预测	20.4%

因变量: 是否曾经违约

a. 基于检验样本的错误计算。

模型摘要显示几个正值符号：

- 训练、检验和坚持样本的错误预测百分比大致相同。
- 因为算法进行一步之后错误为减少，所以估计算法停止。

这进一步表明原始模型实际上可能超额训练并且通过添加一个检验样本可以解决该问题。当然，样本大小相对较小，同时我们也许不应该为百分点变化添加过多涵义。

Classification

图片 4-15
Classification

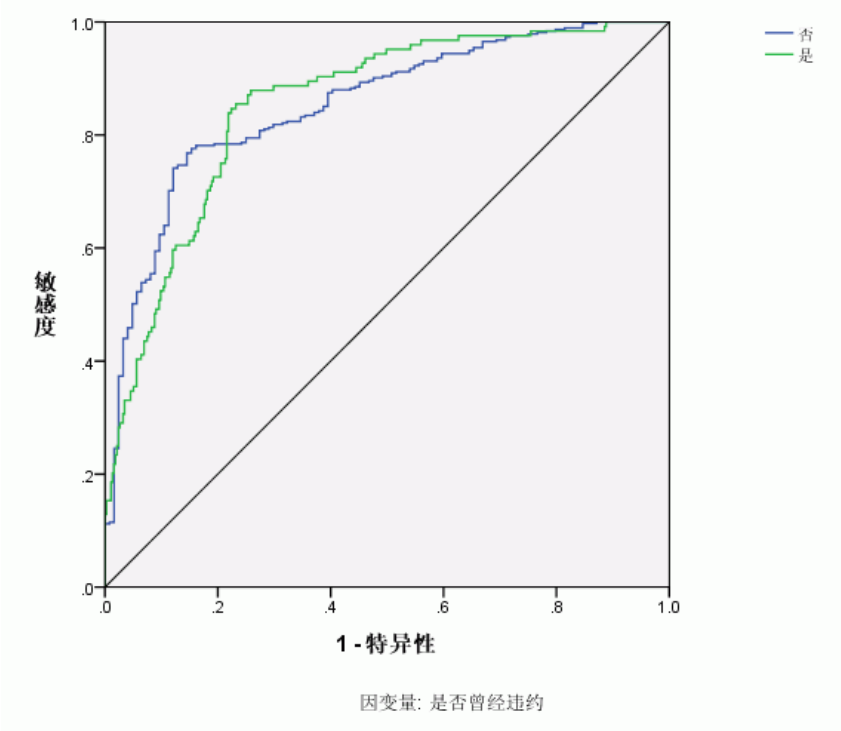
		已预测		
样本	已观测	否	是	正确百分比
训练	否	263	34	88.6%
	是	46	55	54.5%
	总计百分比	77.6%	22.4%	79.9%
测试	否	73	5	93.6%
	是	13	10	43.5%
	总计百分比	85.1%	14.9%	82.2%
保持	否	124	18	87.3%
	是	23	36	61.0%
	总计百分比	73.1%	26.9%	79.6%

因变量: 是否曾经违约

分类表显示使用 0.5 作为分类拟概率分界，网络预测未欠贷者比预测欠贷者相对更接近。不过单一分界值提供非常有限的网络预测能力视图，所以对于比较同类网络并不十分有用。而是查看 ROC 曲线。

ROC 曲线

图片 4-16
ROC 曲线



ROC 曲线可以使您以视图显示单个图中所有界限的**敏感度**和**特异性**，比系列表格更清晰且更有效。此处显示的图表显示两条曲线，一条指类别 No，一条指类别 Yes。因为只有两个类别，所以曲线自图表左上角向右下角围绕 45 度线（不显示）对称。

注意，该图表以组合的训练和测试样本为基础。为坚持样本生成一个 ROC 图表，在分区变量拆分文件并在保存预测拟概率运行“ROC 曲线”过程。

图片 4-17
曲线范围

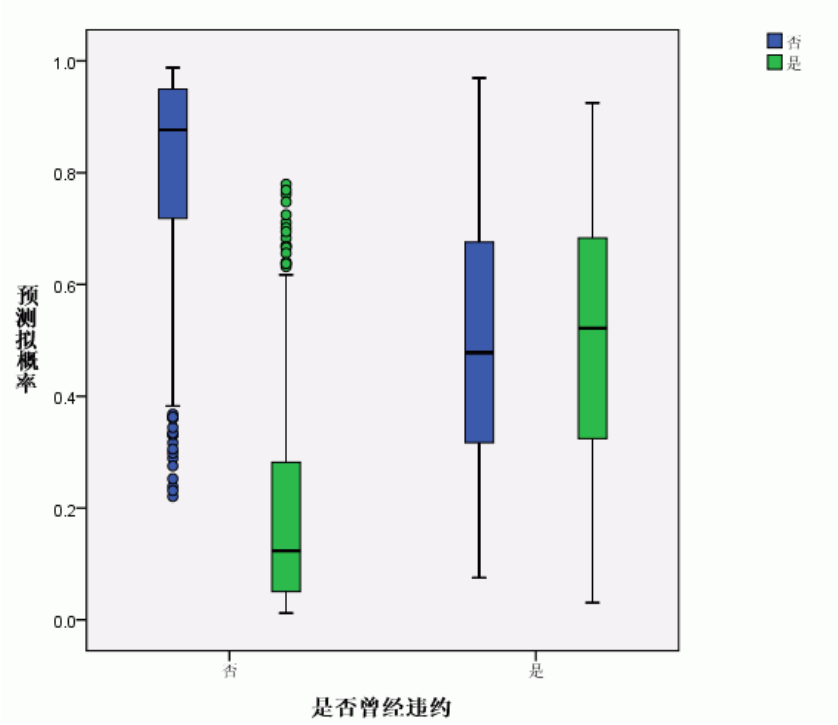
		区域
Previously defaulted	No	.853
	Yes	.853

曲线范围是 ROC 曲线的数字摘要，对于每一个类别，表中的值代表了对于该类别中的预测拟概率，该类别中一个随机选择的个案要高于非该类别中一个随机选择的个案的概率。例如，对于随机选择的拖欠贷款者与随机选择的未拖欠贷款者，就缺省模型预测拟概率而言，前者高于后者的概率为 0.853 。

曲线范围为网络精确性的一个有用的统计量摘要，您需要能够选择一个特定客户分类标准。观察预测图表使此过程以视图开始。

观察预测图

图片 4-18
观察预测图



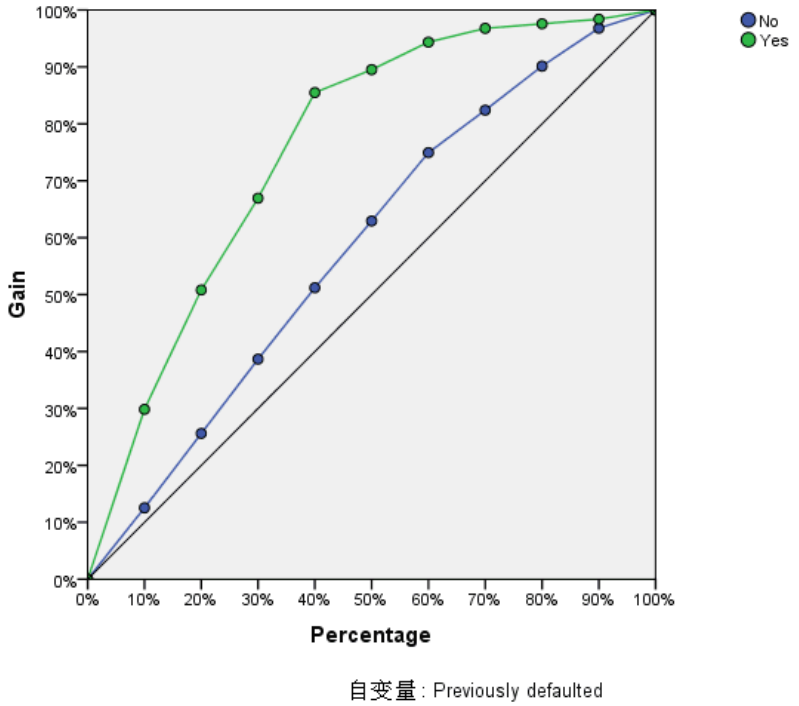
对于分类因变量，观察预测图显示组合的培训和测试样本的预测拟概率的聚类箱图。x 轴对应观察响应类别，而图注对应预测类别。

- 最左侧的箱图显示，对于观察类别 No 的个案，类别 No 的预测拟概率。在 y 轴 0.5 标记之上的箱图部分代表分类表中显示的正确预测值。0.5 标记以下部分代表不正确的预测值。请记住，在分类表中网络善于使用 0.5 界限预测 No 类别的个案，所以只有部分较低细线和一些偏离的个案分类错误。
- 下一个箱图显示，对于观察类别 No 的个案，类别 Yes 的预测拟概率。由于目标变量中只有两个类别，因此前两个箱图在水平线 0.5 对称。
- 第三个箱图显示，对于观察类别 Yes 的个案，类别 No 的预测拟概率。它和最后一个箱图在水平线 0.5 对称。
- 最后一个箱图显示，对于观察类别 Yes 的个案，类别 Yes 的预测拟概率。在 y 轴 0.5 标记之上的箱图部分代表分类表中显示的正确预测值。0.5 标记以下部分代表不正确的预测值。请记住，在分类表中网络使用 0.5 界限预测具有 Yes 类别的稍过半的个案，所以箱图大部分被错误分类。

请看图，其显示通过把将个案分类为 Yes 的界限从 0.5 降至大约 0.3，此大致为第二个箱图的顶端与第四个箱图的底端的值，您可以在不损失大量潜在优质客户的前提下增加准确找出潜在欠贷者的几率。也就是说，沿第二个箱图从 0.5 移至 0.3会将细线处相对较少的未欠贷客户错误地重新分类为预测欠贷者；而对于第四个箱图，此移动会将此箱图中大量欠贷客户正确地重新分类为预测欠贷者。

累积增益和增益图

图片 4-19
累积增益图

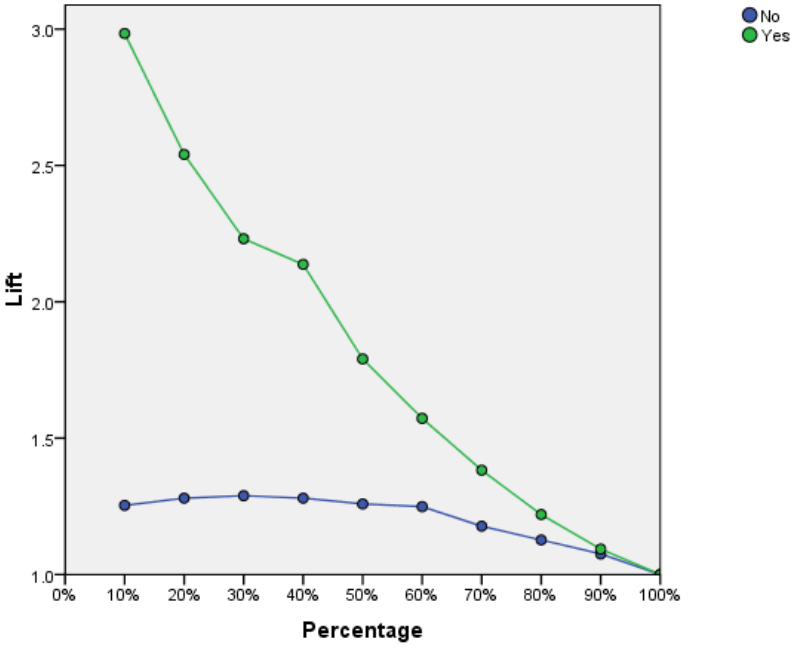


累积增益图会在给定的类别中显示通过把个案总数的百分比作为目标而“增益”的个案总数的百分比。例如，Yes 类别曲线上的第一点在（10%，30%），即如果您使用网络对数据集进行评分并通过 Yes 预测拟概率对所有个案进行排序，您将会期望前 10% 包括实际上为类别 Yes（欠贷者）的所有个案中的大约 30%。同样，前 20% 包括约 50% 欠贷者，前 30% 个案包括 70% 欠贷者，依此类推。如果选择已打分数据集的 100%，您会获得数据集中的所有欠贷者。

对角线为“基线”曲线；如果您从评分数据集随机选择 10% 个案，您期望“获取”实际上为 Yes 的所有个案中的大约 10%。曲线离基线的上方越远，增益越大。您可以使用累积增益图帮助通过选择对应于大量收益的百分比选择分类标准值， 然后将百分比与适当分界值映射。

生成“大量”收益取决于类型 I 与类型 II 错误的成本。也就是说，将拖欠贷款者归类为未拖欠贷款者会造成什么损失（I 类）？将未拖欠贷款者归类为拖欠贷款者会造成什么损失（II 类）？如果主要问题是坏账，您想降低类型 I 错误；在累积增益图，这可能对应于拒绝贷款给 Yes 预测拟概率前 40% 的申请者，这将获取几乎 90% 潜在欠贷者，但几乎移除一半申请者池。 如果首要任务是扩展客户群，您会希望减少 II 类错误。在图表中这可能对应于拒绝贷款给前 10%，这将获取 30% 欠贷者并使大多数申请者池保持完整。通常情况下，两者都是重要问题，因此需要选择客户分类的决策规则，使敏感度和特异性达到最佳配合。

图片 4-20
增益图



自变量: Previously defaulted

增益图源自累积增益图；y 轴上的值对应每条曲线与基线的累积增益比率。这样，类别 Yes 10% 的收益为 30%/10% = 3.0。它提供了另一种在累积增益图中查看信息的方法。

注意：累积增益图和增益图都是以组合的培训和测试样本为基础的。

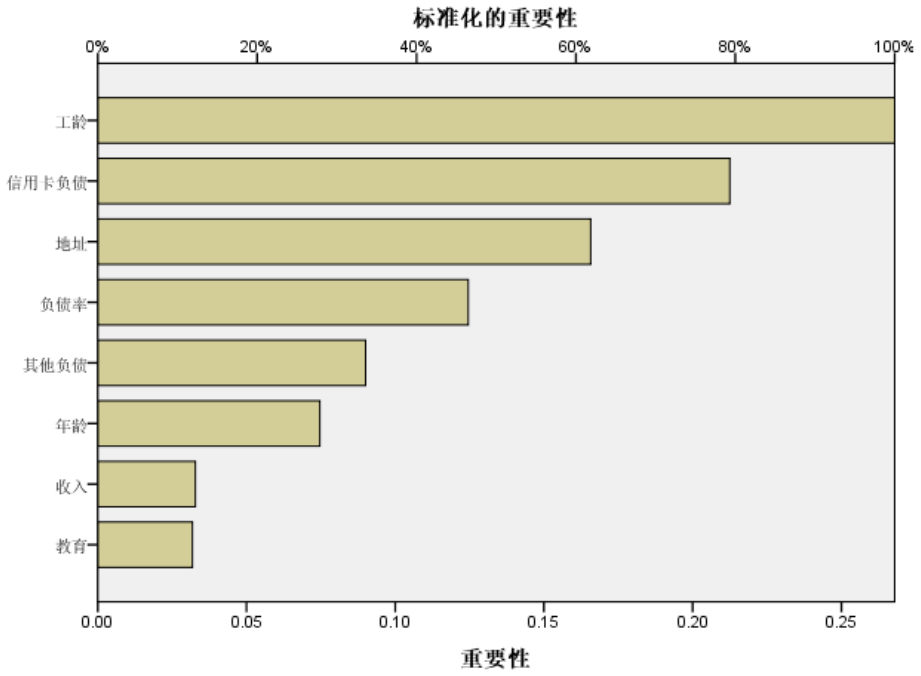
自变量重要性

图片 4-21
自变量重要性

	Importance	Normalized Importance
Level of education	.032	11.9%
Age in years	.075	27.9%
Years with current employer	.268	100.0%
Years at current address	.166	61.8%
Household income in thousands	.033	12.2%
Debt to income ratio (x100)	.125	46.5%
Credit card debt in thousands	.213	79.3%
Other debt in thousands	.090	33.6%

自变量重要性是针对不同自变量值测量网络模型预测值变化量。标准化重要性是由重要性最大值划分的重要性并表示为百分比。

图片 4-22
自变量重要性图表



重要性图为重要性表格中值的条形图，以重要性值降序排序。其显示与客户稳定性（employ、address）和负债（creddebt、debtinc）相关的变量对于网络如何对客户进行分类有重大影响；您不能判断这些变量和拖欠预测概率之间关系的“方向”。您将猜测负债量越大表明拖欠可能性越大，但请确定您需要使用带有更易于解释的参数的模型。

摘要

通过使用多层感知器过程，您构造了一个用于预测给定客户欠贷概率的网络。模型结果与那些使用 Logistic 回归或判别分析获取的模型结果相当，所以您可以相当确信数据不包括不能由那些模型获取的关系，因此您可以使用它们来进一步探索自变量和因变量之间关系的性质。

使用多层感知器估计保健成本与住院时间

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。

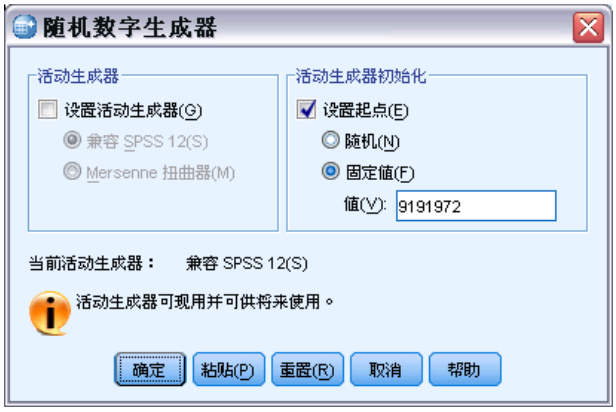
数据文件 patient_los.sav 包含接受 MI 治疗的病人样本的治疗记录。有关详细信息，请参阅第 76 页码附录 A 中的样本文件。使用多层感知器过程创建预测成本与住院时间的网络。

准备数据以进行分析

通过设置随机数种子您可以精确复制此分析。

- 要设置随机数种子，请从菜单中选择：
  转换 > 随机数字生成器...

图片 4-23
“随机数生成器”对话框

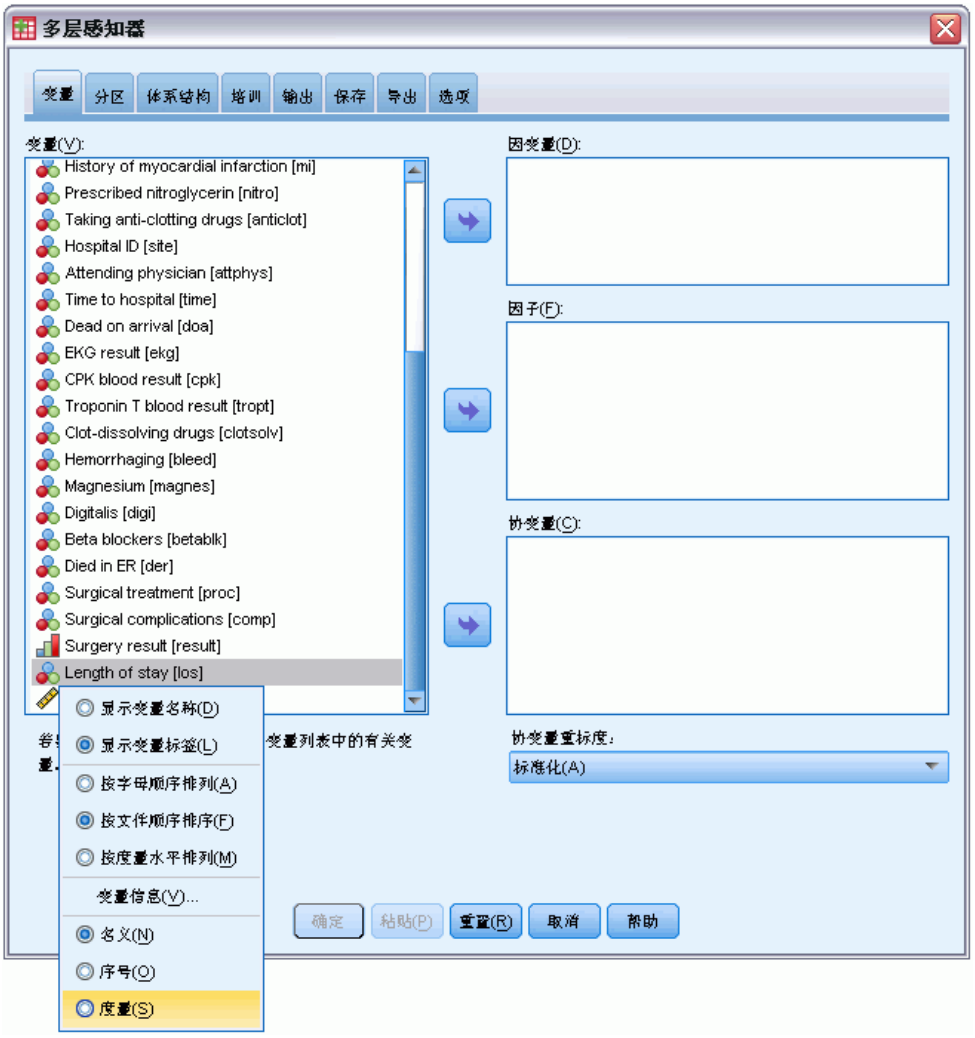


- 选择设置起点。
- 选择固定值并键入 9191972 作为值。
- 单击确定。

运行分析

- 要运行“多层感知器”分析，请从菜单中选择：
  分析 > 神经网络 > 多层感知器...

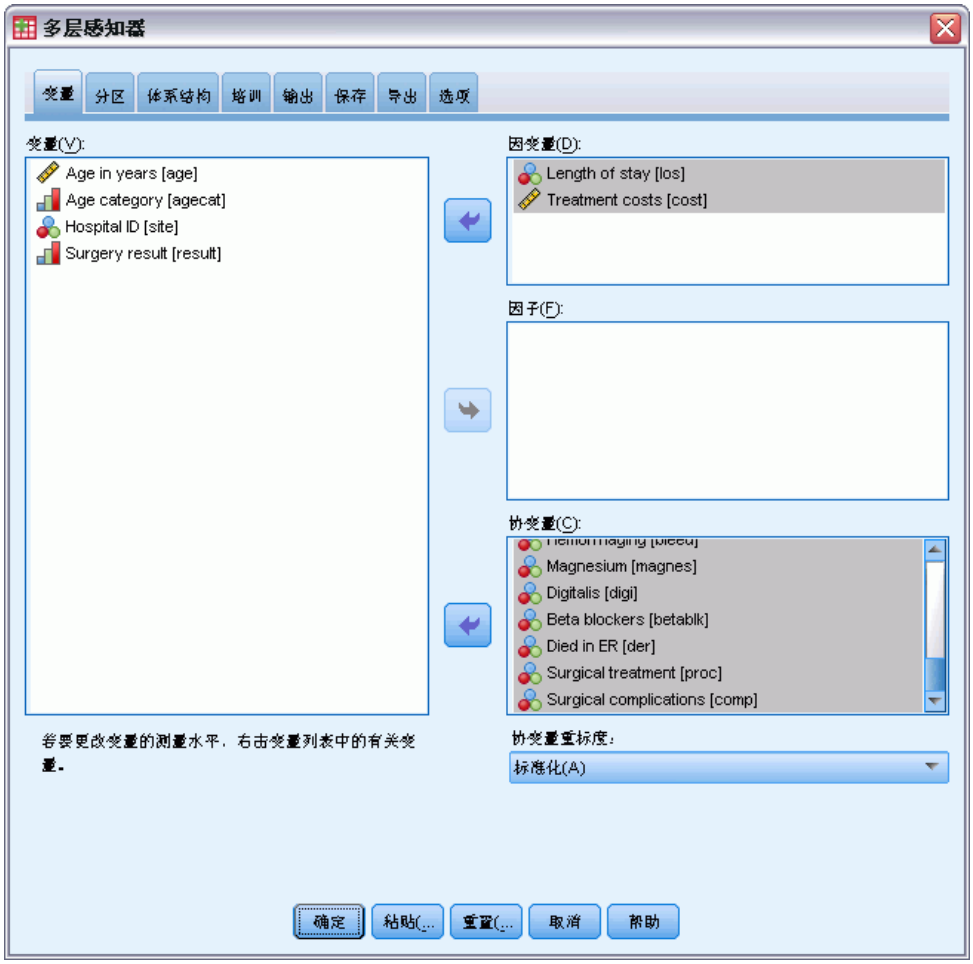
图片 4-24
多层感知器：“变量”选项卡，“住院时间”上下文菜单



Length of stay [los] 存在名义测量级别，但您想让网络将其视为刻度。

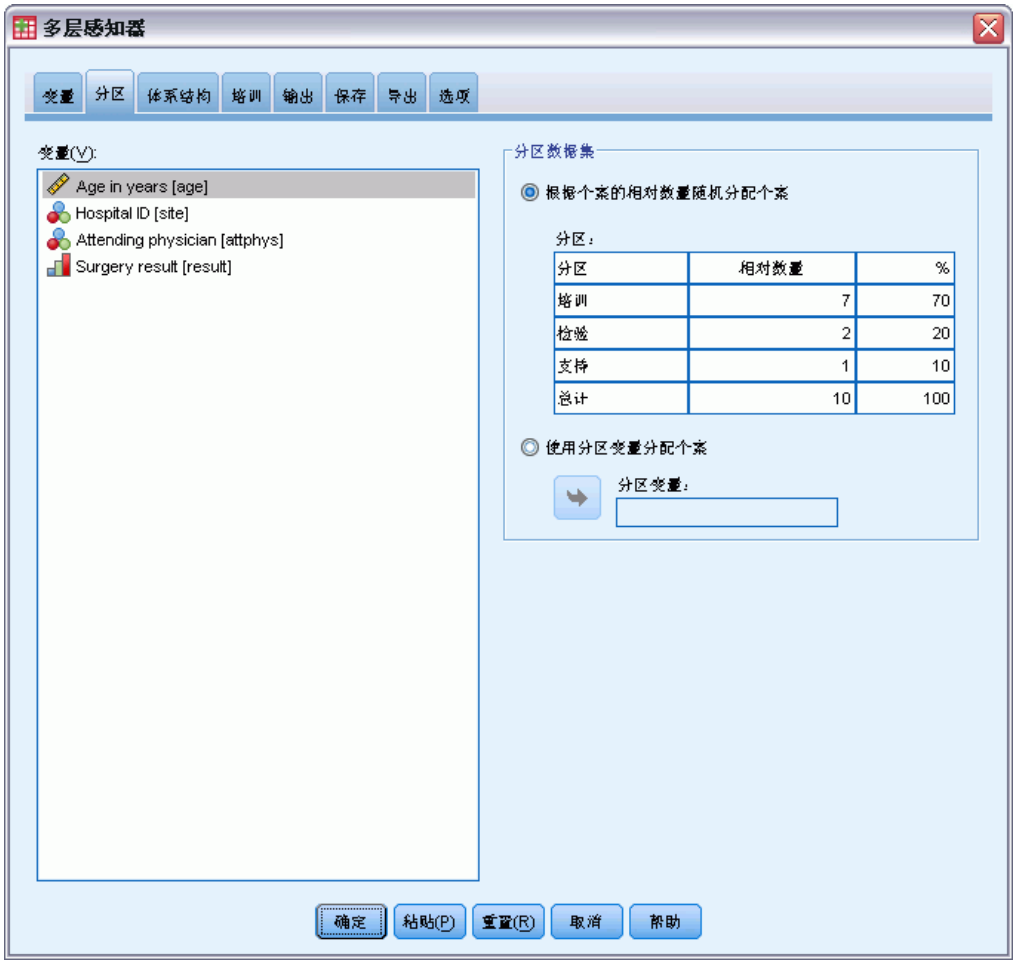
▶ 右键单击 Length of stay [los] 并在上下文菜单选择**刻度**。

图片 4-25
多层感知器：选择了因变量和因子的“变量”选项卡



- 选择 Length of stay [los] 和 Treatment costs [cost] 作为因变量。
- 选择 Age category [agecat] 到 Taking anti-clotting drugs [anticlot] 和 Time to hospital [time] 到 Surgical complications [comp] 作为因素。为确保精确复制以下模型结果，请确保维持因子列表中变量的顺序。至此，您可能发现其有助于选择每组预测变量并使用按钮将其移至因子列表，而不是使用拖放。另外，更改变量顺序将有助于您评估解决方案的稳定性。
- 单击分区选项卡。

图片 4-26
多层感知器：“分区”选项卡



- 键入 2 作为个案的相对数量以分配至检验样本。
- 键入 1 作为个案的相对数量以分配至坚持样本。
- 单击体系结构选项卡。

图片 4-27
多层感知器：体系结构选项卡

- 选择**定制体系结构**。
- 选择 2 作为隐藏层数量。
- 选择**双曲正切**作为输出层激活函数。请注意，这将自动为因变量设置重标度方法以**调整标准化**。
- 单击**培训**选项卡。

图片 4-28
多层感知器：“培训”选项卡

- 选择 Online 作为培训类型。在线培训在“大型”相关预测变量数据集应该效果很好。请注意，这将自动设置梯度下降为相应缺省选项优化算法.
- 单击输出选项卡。

图片 4-29
多层感知器：“输出”选项卡

- 取消选择**图表**；存在大量输入，结果图表将难以使用。
- 选择网络性能组中的**观察预测图**和**残差分析图**。因为无因变量被视为分类因变量（标定或名义），所以分类结果、ROC 曲线、累积增益图和增益图不可用。
- 选择**自变量重要性分析**。
- 单击**选项**选项卡。

图片 4-30
“选项”选项卡

- 选择**包括**用户缺失变量。未经历手术过程的病人的手术并发症变量存在用户缺失值。这确保将那些病人包括在分析中。
- 单击**确定**。

警告

图片 4-31
警告

警告

以下自变量在训练样本中是常数，且不在分析范围之内: der, doa.

警告表提出需注意变量 doa 和 der 在训练样本中为常数。到达时已死亡或在急救室死亡的患者在 Length of stay 存在用户缺失值。既然我们将 Length of stay 视为此分析的刻度变量并且已排除刻度变量存在用户缺失值的个案，只包括在急救室急救之后存活的患者。

个案处理摘要

图片 4-32
个案处理摘要

		N	百分比
样本	培训	5647	70.6%
	检验	1570	19.6%
	保留	781	9.8%
有效		7998	100.0%
排除		2002	
合计		10000	

个案处理摘要显示，有 5647 个个案被分配到培训样本、1570 个被分配到测试样本以及 781 个被分配到了保持样本。从分析中排除的 2002 个案为在前往医院途中或在急救室死亡的患者。

网络信息

图片 4-33
网络信息

输入层	因子	1	Obesity
		2	History of angina
		3	History of myocardial infarction
		4	Prescribed nitroglycerin
		5	Taking anti-clotting drugs
		6	Time to hospital
		7	EKG result
		8	CPK blood result
		9	Troponin T blood result
		10	Clot-dissolving drugs
		11	Hemorrhaging
		12	Magnesium
		13	Digitalis
		14	Beta blockers
		15	Surgical treatment
		16	Surgical complications
		17	Gender
		18	History of diabetes
		19	Blood pressure
		20	Smoker
		21	Cholesterol
		22	Physically active
		23	Age category
	单位数		63
隐藏层	隐藏层数		2
	隐藏层 1 中的单位数		12
	Number of Units in Hidden Layer 2		9
	激活函数		双曲正切
输出层	因变量	1	Length of stay
		2	Treatment costs
	单位数		2
	Rescaling Method for Scale Dependents		Adjusted Normalized
	激活函数		双曲正切
	错误函数		平方和

网络信息表显示有关神经网络的信息，它对于确保指定正确很有用。此处特别要注意的是：

- 输入层的单元数为因子水平总数（无协变量）。
- 请求两个隐藏层，过程已在第一个隐藏层选择 12 个单元，在第二个单元选择 9 个单元。

- 为各刻度因变量创建一个单独输出单元。通过调整标准化方法将其重新调整，这要求输出层使用双曲正切激活函数。
- 因为因变量为刻度变量，所以报告平方和错误。

模型摘要

图片 4-34
模型摘要

训练	平方和错误		93.990
	平均整体相对错误		.085
	尺度因变量的相对错误	Length of stay	.136
		Treatment costs	.032
	中止使用的规则		错误未减少的 1 连续步骤[a]
	培训时间		0:00:11.110
测试	平方和错误		26.855
	平均整体相对错误		.088
	尺度因变量的相对错误	Length of stay	.141
		Treatment costs	.034
保持	平均整体相对错误		.097
	尺度因变量的相对错误	Length of stay	.153
		Treatment costs	.039

a. 基于检验样本的错误计算。

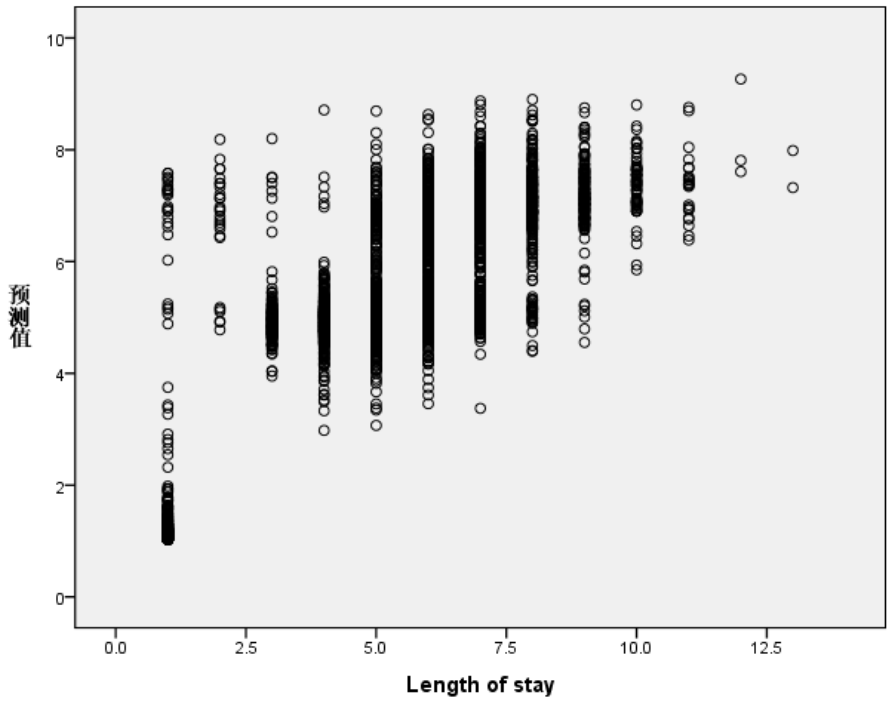
模型摘要显示与训练结果及将最终网络应用于坚持样本相关的信息。

- 因为输出层存在刻度因变量，所以显示平方和错误。这是网络试图在训练中最小化的错误函数。请注意，因变量重新调整值计算平方和与所有以下错误值。
- 各刻度因变量的相对错误为因变量平方和错误与“零”模型平方和错误的比率，在“零”模型中因变量均值被用作每个个案的预测值。Length of stay 的预测值似乎比Treatment costs的预测值存在更多错误。
- 平均整体错误为所有因变量的平方和错误与“零”模型平方和错误的比率，在“零”模型中因变量均值被用作每个个案的预测值。在本例中，平均整体错误恰好接近于相对错误平均值，但并不总是如此。

  平均整体相对错误与相对错误在训练、检验和坚持样本中都是常数，这将使您在某种程度上相信模型未超额训练并且在由网络评分的将来个案中的错误将接近于此表格中报告的错误。
- 因为算法进行一步之后错误为减少，所以估计算法停止。

观察预测图

图片 4-35
Length of stay 观察预测图

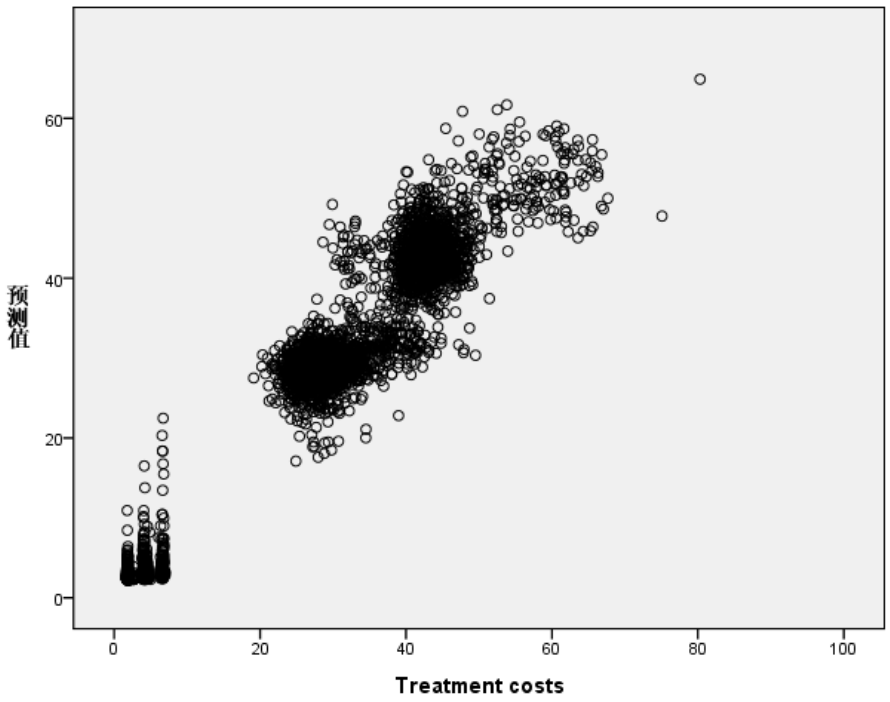


对于刻度因变量，观察预测图针对组合的训练与检验样本显示 y 轴上的预测值与 x 轴上的观察值的散点图。理想的情况下，值应大致位于从原点起始的 45 度线。在此图表垂直线上的点代表各观察 Length of stay 天数。

请看图，其显示网络是预测 Length of stay 的好办法。图的常规趋势是理想 45 度线外，这意味着五天内观察的住院时间预测值往往过高估计了住院时间，而六天以上观察的住院时间预测值则往往低估了住院时间。

在图左下方部分的一组患者很可能还未做手术。在图左上方部分也有一组患者，观察住院时间为一至三天，因此预测值过大。很可能这些个案为在医院手术后死亡的患者。

图片 4-36
治疗成本的观察预测图



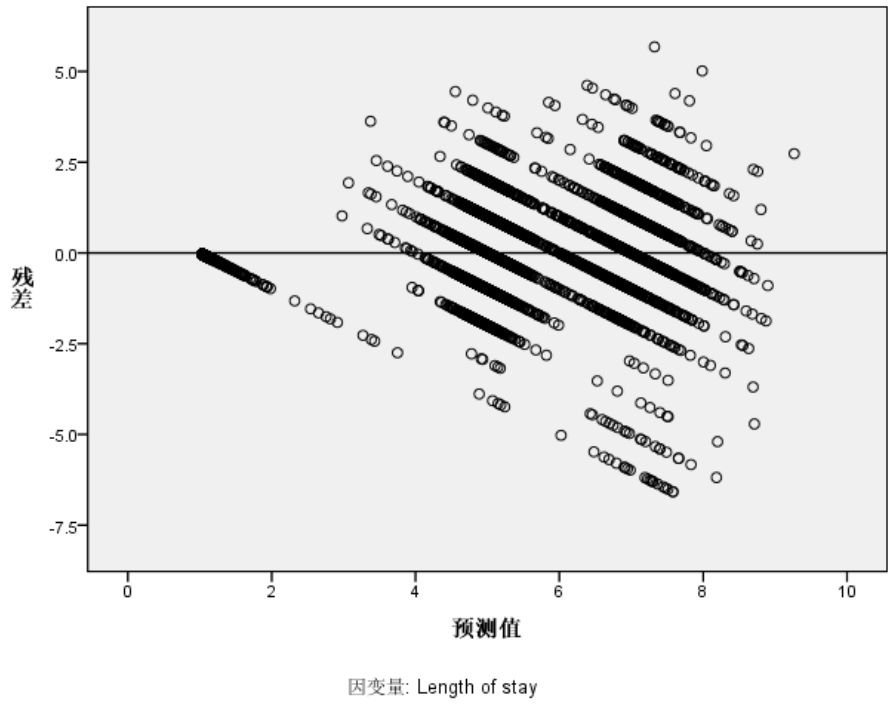
网络预测Treatment costs效果似乎很好。似乎有三组主要患者。

- 左下方主要为未做手术的患者。其成本相对较低，可根据在急救室中控制的 Clot-dissolving drugs [clotsolv] 的类型来区分。
- 下一组患者的治疗成本大约为 $30,000。存在做过皮肤冠状血管修复术（PTCA）的患者。
- 最后一组的治疗成本超过 $40,000。这些为做过冠状动脉绕道手术（CABG）的患者。此手术费用比 PTCA 高些，并且患者住院康复时间较长，这也进一步增加了成本。

还有许多成本超过 $50,000 的个案，但网络预测效果不佳。这些是在手术时经历并发症的患者，这会增加手术成本与住院时间。

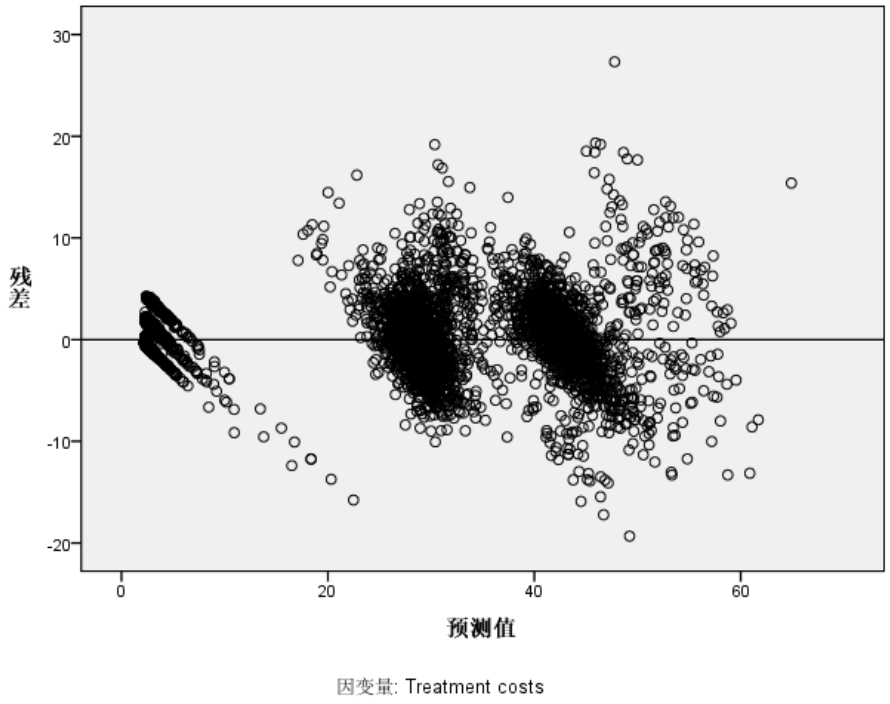
残差分析图

图片 4-37
住院时间的残差分析图



残差分析图显示 y 轴上的残差（观察值减去预测值）与 x 轴上的预测值的散点图。此图中的各对角线对应于观察预测图中的垂直线，当观察住院时间增加时，您能更清晰地看到住院时间从高预测值到低预测值的变化过程。

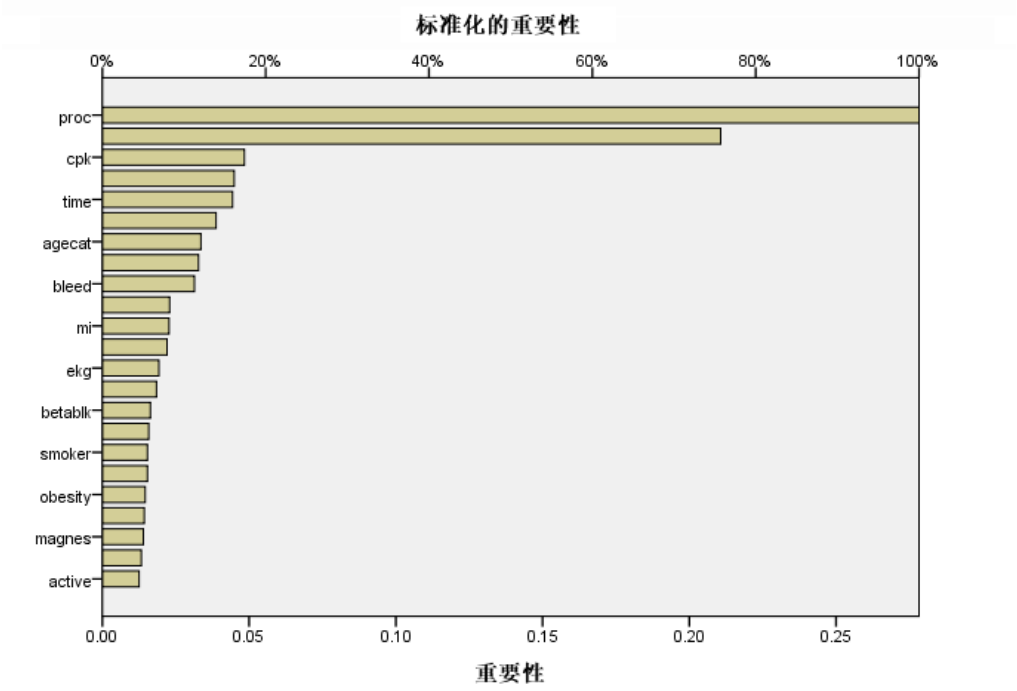
图片 4-38
治疗成本的残差分析图



同样，对于观察预测图中的三组被观察患者的Treatment costs，当观察成本增加时，残差分析图显示成本从高预测值到低预测值的变化过程。在 CABG 期间有并发症的患者仍然清晰可见，但也可以更方便地看到在 PTCA 期间有并发症的患者；其将显示为在 x 轴 $30,000 标记周围的主要 PTCA 患者组右上方的子组。

自变量重要性

图片 4-39
自变量重要性图表



重要性图显示结果由完成的手术过程控制，随后是否出现并发症，再随后就是其他预测值。手术过程的重要性在Treatment costs图中清晰可见，而在Length of stay中却不太清晰，尽管Length of stay中的并发症影响在最大观察住院时间的患者中仍可见。

摘要

网络在为“典型”患者预测值时似乎效果很好，但不获取手术后死亡的患者。处理该情况的一种可能的方法就是创建多个网络。一个网络预测患者结果，可能只是预测患者是否存活，然后单独的网络预测 Treatment costs 和 Length of stay，条件是患者是否存活。然后您可以结合网络结果，以可能获取更好的预测值。您可以采用类似方法应用于过低预测在手术期间经历并发症的患者的成本与住院时间的问题。

推荐参考

有关神经网络和多层感知器的更多信息，请参见以下内容：

Bishop, C. M. 1995. Neural Networks for Pattern Recognition, 3rd ed. Oxford: Oxford University Press.

细, T. L. 1999. Feedforward Neural Network Methodology, 3rd ed. New York: Springer-Verlag.

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

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

章 5

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。

使用径向基函数分类电信客户

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。如果人口统计学数据可用于预测组成员资格，则可以为各个潜在客户定制服务。

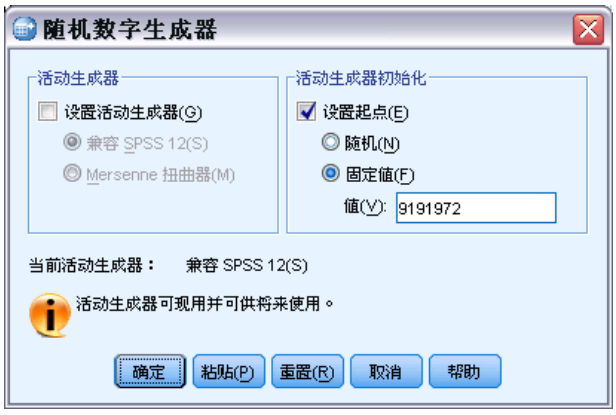
假设当前客户的信息包含在 telco.sav 中。有关详细信息，请参阅第 76 页码附录 A 中的样本文件。使用径向基函数过程分类客户。

准备数据以进行分析

通过设置随机数种子您可以精确复制此分析。

- 要设置随机数种子，请从菜单中选择：
  转换 > **随机数字生成器...**

图片 5-1
“随机数生成器”对话框

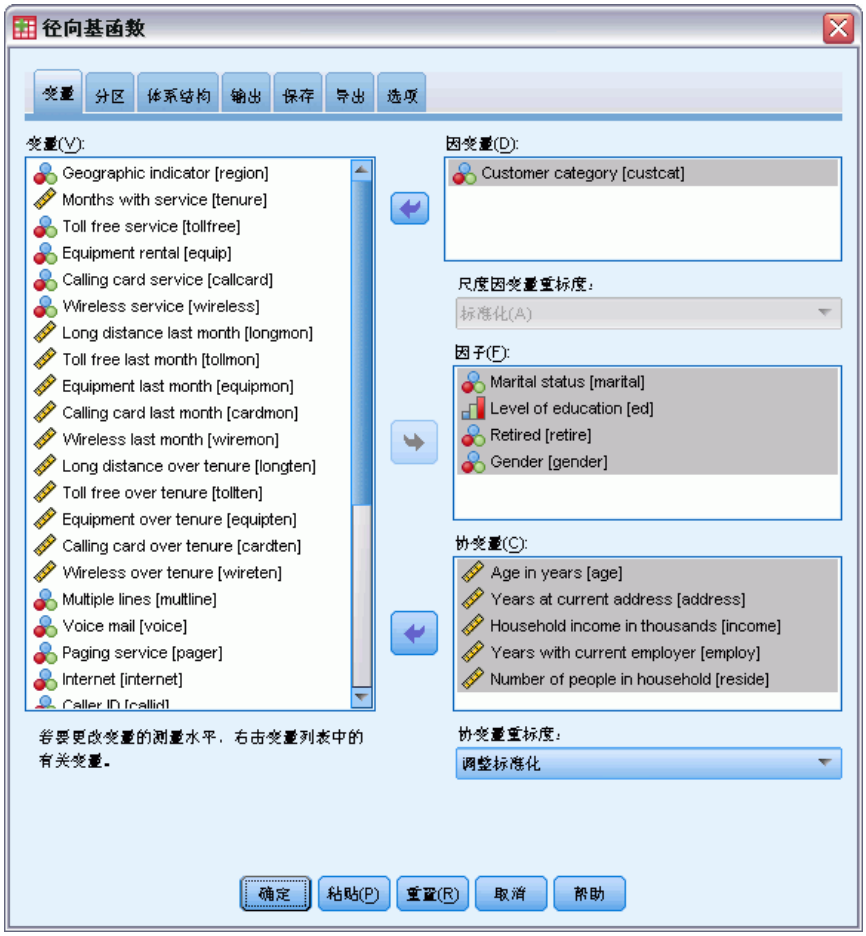


- 选择**设置起点**。
- 选择**固定值**并键入 9191972 作为值。
- 单击**确定**。

运行分析

- 要运行“径向基函数”分析，请从菜单中选择：
  分析 > **神经网络** > **径向基函数...**

图片 5-2
径向基函数：“变量”选项卡



- 选择客户类别 [custcat] 作为因变量。
- 选择 婚姻状况 [marital]、教育程度 [ed]、退休 [retire]和性别 [gender]作为因子。
- 选择 年龄 [age] 到 家庭成员人数 [reside]作为协变量。
- 选择 **调整标准化** 作为重标度协变量的方式。
- 单击**分区**选项卡。

图片 5-3
径向基函数：“分区”选项卡

通过指定个案的相对数量，可以很方便地创建难以指定百分比的小数分区。比如您想将数据集的 2/3 分配给培训样本，将剩余个案的 2/3 分配给测试样本。

- 键入 6 作为培训样本的相对数量。
- 键入 2 作为测试样本的相对数量。
- 键入 1 作为保持样本的相对数量。

总共指定了 9 个相对个案。6/9 = 2/3，即大约 66.67% 被分配给培训样本；2/9，即大约 22.22% 被分配给测试样本；1/9，即大约 11.11% 被分配给保持样本。

- 单击**输出**选项卡。

图片 5-4
径向基函数：“输出”选项卡

- 在“网络结构”组选项中取消选择图表。
- 在“网络性能”组选项中选择 ROC 曲线、累积增益图、增益图和观察预测图。
- 单击保存选项卡。

图片 5-5
径向基函数：“保存”选项卡

- 选择 保存每个因变量的预测值或类别和保存每个因变量的预测拟概率。
- 单击确定。

个案处理摘要

图片 5-6
个案处理摘要

		N	百分比
样本	培训	665	66.5%
	检验	224	22.4%
	保留	111	11.1%
有效		1000	100.0%
排除		0	
合计		1000	

个案处理摘要显示，有 665 个个案被分配到培训样本、224 个被分配到测试样本以及 111 个被分配到了保持样本。没有个案从分析中排除。

网络信息

图片 5-7
网络信息

输入层	因子	1	Marital status
		2	Level of education
		3	Retired
		4	Gender
	协变量	1	Age in years
		2	Years at current address
		3	Household income in thousands
		4	Years with current employer
		5	Number of people in household
	单位数		16
	协变量的重标度方法		Adjusted Normalized
隐藏层	单位数		9[a]
	激活函数		Softmax
输出层	因变量	1	Customer category
	单位数		4
	激活函数		恒等
	错误函数		平方和

a. 由检验数据标准确定：隐藏单位的"最佳"数量为检验数据中产生最小错误的单位。

网络信息表显示有关神经网络的信息，它对于确保指定正确很有用。此处特别要注意的是：

- 输入层的单位数是协变量数与因子级别总数的和；为每个婚姻状况、教育程度、退休和性别类别创建一个单独的单位，而且没有一个类别被认为是“冗余”单位，这是许多建模过程中的典型。
- 类似地，为每个客户类别类别创建一个单独的输出单位，所以在输出层总共有 4 个单位。
- 使用调整标准化方式来重标度协变量。
- 自动体系结构选项选择了隐藏层中的 9 个单位。
- 所有其他网络信息都是过程的缺省值。

模型摘要

图片 5-8
模型摘要

训练	平方和错误	235.969
	百分比错误预测	61.8%
	培训时间	0:00:03.110
测试	平方和错误	80.851[a]
	百分比错误预测	62.9%
保持	百分比错误预测	59.5%

因变量: Customer category

a. 隐藏单位的数量由检验数据标准确定：隐藏单位的"最佳"数量为检验数据中产生最小错误的单位。

模型摘要显示有关培训、测试以及将最终网络应用到保持样本的结果的信息。

- 因为平方和错误经常用于 RBF 网络，所以它也被显示出来。这是网络在培训和测试过程中试图最小化的错误函数。
- 错误预测值的百分比取自分类表，并将在该主题中作进一步讨论。

Classification

图片 5-9
Classification

		已预测				
样本	已观测	Basic service	E-service	Plus service	Total service	正确百分比
训练	Basic service	64	0	66	45	36.6%
	E-service	22	1	57	61	.7%
	Plus service	47	0	104	34	56.2%
	Total service	29	1	49	85	51.8%
	总计百分比	24.4%	.3%	41.5%	33.8%	38.2%
测试	Basic service	18	0	26	15	30.5%
	E-service	15	0	16	22	.0%
	Plus service	11	0	39	15	60.0%
	Total service	4	0	17	26	55.3%
	总计百分比	21.4%	.0%	43.8%	34.8%	37.1%
保持	Basic service	11	0	11	10	34.4%
	E-service	4	0	9	10	.0%
	Plus service	10	0	19	2	61.3%
	Total service	5	0	5	15	60.0%
	总计百分比	27.0%	.0%	39.6%	33.3%	40.5%

因变量: Customer category

分类表显示使用网络的实际结果。对于每个个案，预测响应都是预测拟概率最高的类别。

- 对角线上的单元格是正确的预测值。
- 偏离对角线的单元格是不正确的预测值。

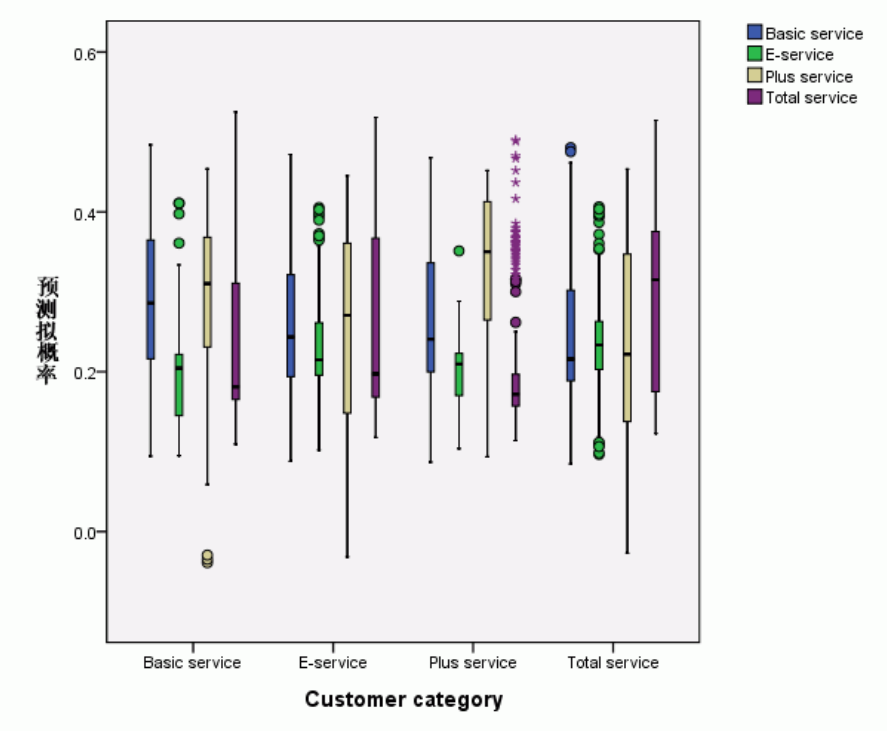
如果给定观察数据，“零”模型（即没有预测变量的模型）将把所有客户分类到模态组的附加服务。因此，零模型有 281/1000 = 28.1% 的可能性是正确的。RBF 网络获得了 10.1% 以上，即 38.2% 的客户。实际上，您的模型对于识别附加服务和总体服务客户效

果最好。但是，对于分类电子服务客户，其效果很差。您可能需要找到另一个预测变量以便分离这些客户；或者，如果这些客户经常被误分类为附加服务和总体服务客户，那么公司可以尝试对那些通常会落到电子服务类别的潜在客户进行直销。

基于创建模型所用个案的分类从其分类率有所夸大的意义上来说，倾向于过度“乐观”。保持样本帮助验证模型；这些个案中，有 40.2% 是由模型正确分类的。尽管保持样本有点小，但这意味着您的模型实际上有四成是正确的。

观察预测图

图片 5-10
观察预测图



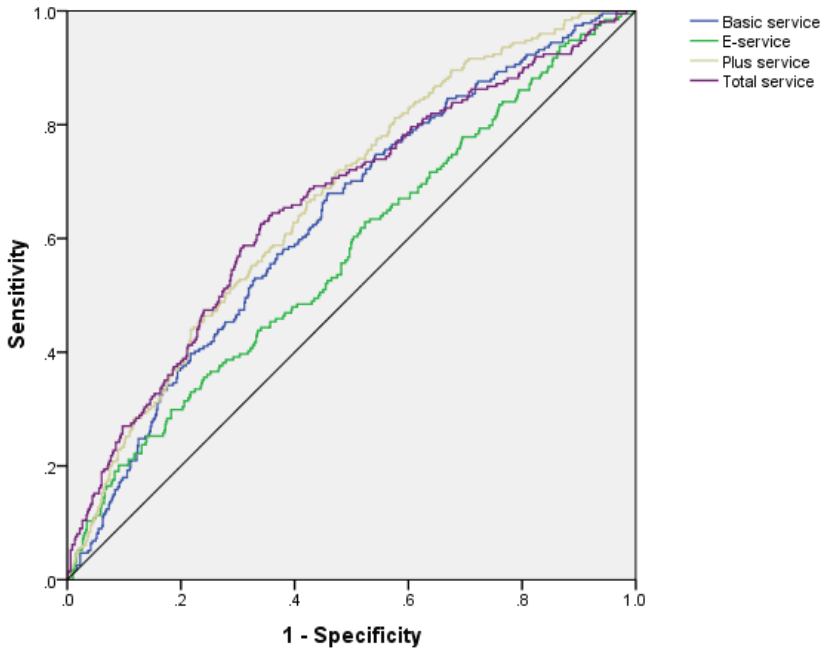
对于分类因变量，观察预测图显示组合的培训和测试样本的预测拟概率的聚类箱图。x 轴对应观察响应类别，而图注对应预测类别。因此：

- 对于具有观察类别基本服务的个案，最左边的箱图会显示类别基本服务的预测拟概率。
- 对于具有观察类别基本服务的个案，右侧第二张箱图会显示类别电子服务的预测拟概率。
- 对于具有观察类别基本服务的个案，第三张箱图会显示类别附加服务的预测拟概率。从分类表中大致可以看出，基本服务客户中，被误分类为附加服务客户的人数与被正确分类为基本服务客户的人数一样多；因此，该箱图大致等同于最左边的箱图。
- 对于具有观察类别基本服务的个案，第四张箱图会显示类别总体服务的预测拟概率。

由于目标变量中有两个以上的类别，所以前四张箱图与 0.5 的水平线始终无法对称。结果，使用两个以上的类别来为目标变量解释该图会很困难，因为仅凭在一个箱图中查看部分个案是不可能确定那些个案在另一个箱图中的相应位置的。

ROC 曲线

图片 5-11
ROC 曲线



目变量：Customer category

ROC 曲线通过所有可能的分类界限的**敏感度**为您提供**敏感度**的可视显示。这里显示的图表显示四条曲线，每一条代表一个目标变量类别。

注意，该图表以组合的训练和测试样本为基础。要为坚持样本生成一个 ROC 图表，请在分区变量拆分文件并在预测拟概率运行“ROC 曲线”过程。

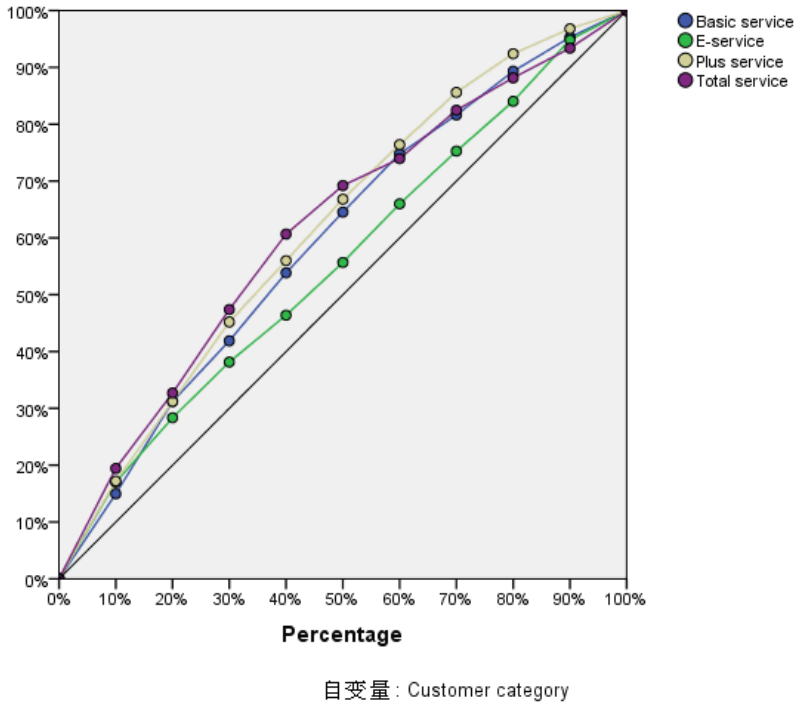
图片 5-12
曲线范围

		区域
Customer category	Basic service	.635
	E-service	.573
	Plus service	.668
	Total service	.659

曲线范围是 ROC 曲线的数字摘要，对于每一个类别，表中的值代表了对于该类别中的预测拟概率，该类别中一个随机选择的个案要高于非该类别中一个随机选择的个案的概率。例如，对于在附加服务中随机选择的客户和在基本服务、电子服务或总体服务中随机选择的客户，附加服务中客户的缺省模型预测拟概率将偏高的概率为 0.668。

累积增益和增益图

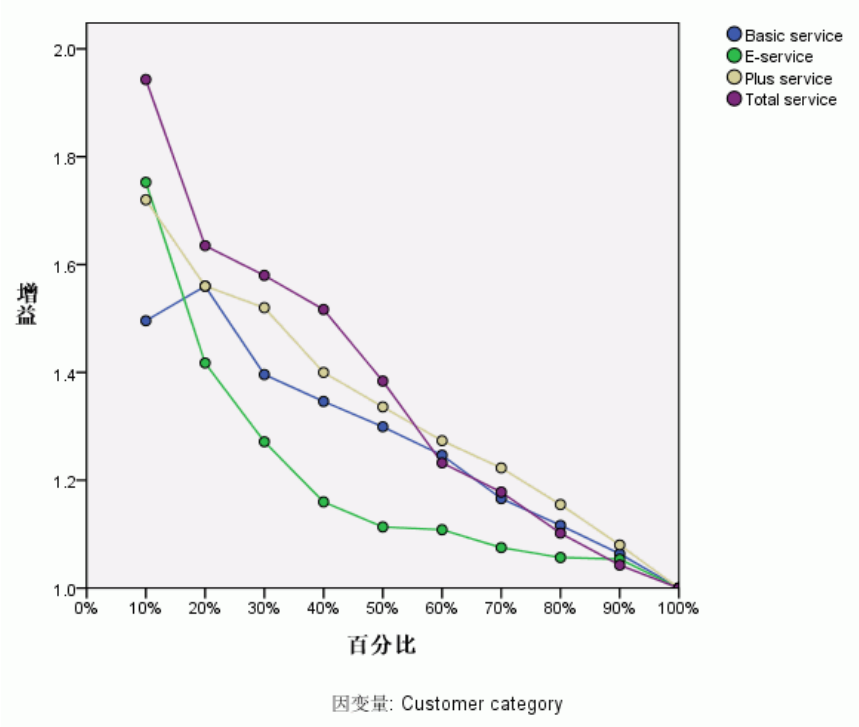
图片 5-13
累积增益图



累积增益图会在给定的类别中显示通过把个案总数的百分比作为目标而“增益”的个案总数的百分比。例如，总体服务类别曲线上的第一个点大致位于 (10%，20%)，这就意味着如果您使用网络来为数据集打分或通过总体服务的预测拟概率来对所有个案进行排序，那么前 10% 将可以包含所有实际分配到总体服务类别的个案的大约 20%。类似地，前 20% 将包含大约 30% 的欠贷者，前 30% 的个案将包含 50% 的欠贷者，依此类推。如果选择已打分数据集的 100%，您会获得数据集中的所有欠贷者。

对角线是“基线”曲线；如果您从已打分数据集中随机选择 10% 的个案，那么您将“增益”实际分配到任何给定类别的所有个案的大约 10%。曲线离基线的上方越远，增益越大。

图片 5-14
增益图



增益图源自累积增益图；y 轴上的值对应每条曲线与基线的累积增益比率。因此，总体服务类别 10% 的增益约为 20%/10% = 2.0。它提供了另一种在累积增益图中查看信息的方法。

注意：累积增益图和增益图都是以组合的培训和测试样本为基础的。

推荐参考

有关径向基函数的更多信息，请参见以下内容：

Bishop, C. M. 1995. Neural Networks for Pattern Recognition, 3rd ed. Oxford: Oxford University Press.

细, T. L. 1999. Feedforward Neural Network Methodology, 3rd ed. New York: Springer-Verlag.

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

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

Tao, K. K. 1993. A closer look at the radial basis function (RBF) networks. 输入: Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers, A. Singh，ed. Los Alamitos, Calif.: IEEE Comput. Soc. Press.

Uykan, Z., C. Guzelis, M. E. Celebi, 和 H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN. IEEE Transactions on Neural Networks, 11, .

附录
A

样本文件

随产品一起安装的样本文件可以在安装目录的 Samples 子目录中找到。对于以下每种语言在“样本”子目录中有单独的文件夹：英语、法语、德语、意大利语、日语、韩语、波兰语、俄语、简体中文、西班牙语和繁体中文。

并非所有样本文件均提供此处的全部语言版本。如果样本文件未提供某种语言的版本，则相应语言文件夹中包含该样本文件的英语版本。

描述

以下是对在整个文档的各种示例中使用的样本文件的简要描述。

- **accidents.sav。** 该假设数据文件涉及某保险公司，该公司正在研究给定区域内汽车事故的年龄和性别风险因子。每个个案对应一个年龄类别和性别类别的交叉分类。
- **adl.sav。**该假设数据文件涉及在确定针对脑卒中患者的建议治疗类型的优点方面的举措。医师将女性脑卒中患者随机分配到两组中的一组。第一组患者接受标准的物理治疗，而第二组患者则接受附加的情绪治疗。在进行治疗的三个月时间里，将为每个患者进行一般日常生活行为的能力评分并作为原始变量。
- **advert.sav。**该假设数据文件涉及某零售商在检查广告支出与销售业绩之间的关系方面的举措。为此，他们收集了过去的销售数据以及相关的广告成本..
- **aflatoxin.sav。**该假设数据文件涉及对谷物的黄曲霉毒素的检测，该毒素的浓度会因谷物产量的不同（不同谷物之间及同种谷物之间）而有较大变化。谷物加工机从 8 个谷物产量的每一个中收到 16 个样本并以十亿分之几 (PPB) 为单位来测量黄曲霉毒素的水平。
- **anorectic.sav。**在研究厌食/暴食行为的标准症状参照时，研究人员(Van der Ham, Meulman, Van Strien, 和 Van Engeland, 1997)对 55 名已知存在进食障碍的青少年进行了调查。其中每名患者每年都将进行四次检查，因此总观测数为 220。在每次观测期间，将对这些患者按 16 种症状逐项评分。但 71 号和 76 号患者的症状得分均在时间点 2 缺失，47 号患者的症状得分在时间点 3 缺失，因此有效观测数为 217。
- **bankloan.sav。**该假设数据文件涉及某银行在降低贷款拖欠率方面的举措。该文件包含 850 位过去和潜在客户的财务和人口统计信息。前 700 个个案是以前曾获得贷款的客户。剩下的 150 个个案是潜在客户，银行需要按高或低信用风险对他们进行分类。
- **bankloan_binning.sav。**该假设数据文件包含 5,000 位过去客户的财务和人口统计信息。
- **behavior.sav。**在一个经典示例中(Price 和 Bouffard, 1974)，52 名学生被要求以 10 分的标度对 15 种情况和 15 种行为的组合进行评价，该 10 分的标度介于 0 = 平均值在个人值之上，值被视为相异性。
- **behavior_ini.sav。**该数据文件包含 behavior.sav 的二维解的初始配置。

- **brakes.sav。**该假设数据文件涉及某生产高性能汽车盘式制动器的工厂的质量控制。该数据文件包含对 8 台专用机床中每一台的 16 个盘式制动器的直径测量。盘式制动器的目标直径为 322 毫米。
- **breakfast.sav。**在一项经典研究中(Green 和 Rao, 1972)，21 名 Wharton School MBA 学生及其配偶被要求按照喜好程度顺序对 15 种早餐食品进行评价，从 1 =他们的喜好根据六种不同的情况加以记录，从“全部喜欢”到“只带饮料的快餐”。
- **breakfast-overall.sav。**该数据文件只包含早餐食品喜好的第一种情况，即“全部喜欢”。
- **broadband_1.sav。**该假设数据文件包含各地区订制了全国宽带服务的客户的数量。该数据文件包含 4 年期间 85 个地区每月的订户数量。
- **broadband_2.sav。**该数据文件和 broadband_1.sav 一样，但包含另外三个月的数据。
- **car_insurance_claims.sav。**在别处被提出和分析的(McCullagh 和 Nelder, 1989)关于汽车损坏赔偿的数据集。平均理赔金额可以当作其具有 gamma 分布来建模，通过使用逆联接函数将因变量的均值与投保者年龄、车辆类型和车龄的线性组合关联。提出理赔的数量可以作为尺度权重。
- **car_sales.sav。**该数据文件包含假设销售估计值、订价以及各种品牌和型号的车辆的物理规格。订价和物理规格可以从 edmunds.com 和制造商处获得。
- **car_sales_uprepared.sav。**这是 car_sales.sav 的修改版本，不包含字段的任何已转换版本。
- **carpet.sav。**在一个常用示例中(Green 和 Wind, 1973)，一家公司非常重视一种新型地毯清洁用品的市场营销，希望检验以下五种因素对消费者偏好的影响—包装设计、品牌名称、价格、优秀家用品 标志和退货保证。包装设计有三个因子水平，每个因子水平因刷体位置而不同；有三个品牌名称（K2R、Glory 和 Bissell）；有三个价格水平；最后两个因素各有两个级别（有或无）。十名消费者对这些因素所定义的 22 个特征进行了排序。变量优选包含对每个特征的平均等级的排序。低排序与高偏好相对应。此变量反映了对每个特征的偏好的总体度量。
- **carpet_prefs.sav。**该数据文件所基于的示例和在 carpet.sav 中所描述的一样，但它还包含从 10 位消费者的每一位中收集到的实际排列顺序。消费者被要求按照从最喜欢到最不喜欢的顺序对 22 个产品特征进行排序。carpet_plan.sav 中定义了变量 PREF1 到 PREF22 包含相关特征的标识符。
- **catalog.sav。**该数据文件包含某编目公司出售的三种产品的假设每月销售数据。同时还包括 5 个可能的预测变量的数据。
- **catalog_seasfac.sav。**除添加了一组从“季节性分解”过程中计算出来的季节性因子和附带的日期变量外，该数据文件和 catalog.sav 是相同的。
- **cellular.sav。**该假设数据文件涉及某便携式电话公司在减少客户流失方面的举措。客户流失倾向分被应用到帐户，分数范围从 0 到 100。得到 50 分或更高分数的帐户可能会更换提供商。
- **ceramics.sav。**该假设数据文件涉及某制造商在确定新型优质合金是否比标准合金具有更高的耐热性方面的举措。每个个案代表对一种合金的单独检验；个案中会记录合金的耐热极限。
- **cereal.sav。**该假设数据文件涉及一份 880 人参于的关于早餐喜好的民意调查，该调查记录了参与者的年龄、性别、婚姻状况以及生活方式是否积极（根据他们是否每周至少做两次运动）。每个个案代表一个单独的调查对象。

- **clothing_defects.sav。**这是关于某服装厂的质量控制过程的假设数据文件。检验员要对工厂中每次大批量生产的服装进行抽样检测并清点不合格的服装的数量。
- **coffee.sav。**这是关于六种冰咖啡的认知品牌形象(Kennedy, Riquier, 和 Sharp, 1996)的数据文件。对于 23 种冰咖啡特征属性中的每种属性，人们选择了由该属性所描述的所有品牌。为保密起见，六种品牌用 AA、BB、CC、DD、EE 和 FF 来表示。
- **contacts.sav。**该假设数据文件涉及一组公司计算机销售代表的联系方式列表。根据这些销售代表所在的公司部门及其公司的秩来对每个联系方式进行分类。同时还记录了最近一次的销售量、最近一次销售距今的时间和所联系公司的规模。
- **creditpromo.sav。**该假设数据文件涉及某百货公司在评价最新信用卡促销的效果方面的举措。为此，随机选择了 500 位持卡人。其中一半收到了宣传关于在接下来的三个月内降低消费利率的广告。另一半收到了标准的季节性广告。
- **customer_dbase.sav。** 该假设数据文件涉及某公司在使用数据仓库中的信息来为最有可能回应的客户提供特惠商品方面的举措。随机选择客户群的子集并为其提供特惠商品，同时记录下他们的回应。
- **customer_information.sav。**该假设数据文件包含客户邮寄信息，如姓名和地址。
- **customer_subset.sav。**来自 customer_dbase.sav 的拥有 80 个个案的子集。
- **debate.sav。**该假设数据文件涉及在某政治辩论前后对该辩论的参与者所做的调查的成对回答。每个个案对应一个单独的调查对象。
- **debate_aggregate.sav。**该假设数据文件分类汇总了 debate.sav 中的回答。每个个案对应一个辩论前后的偏好的交叉分类。
- **demo.sav。**这是关于购物客户数据库的假设数据文件，用于寄出每月的商品。将记录客户对商品是否有回应以及各种人口统计信息。
- **demo_cs_1.sav。**该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第一步。每个个案对应不同的城市，并记录地区、省、区和城市标识。
- **demo_cs_2.sav。**该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第二步。每个个案对应来自第一步中所选城市的不同的家庭单元，并记录地区、省、区、市、子区和单元标识。还包括设计前两个阶段的抽样信息。
- **demo_cs.sav。**该假设数据文件包含用复杂抽样设计收集的调查信息。每个个案对应不同的家庭单元，并记录各种人口统计和抽样信息。
- **dmdata.sav。**该假设数据文件包含直销公司的人口统计学和购买信息。dmdata2.sav 包含收到测试邮件的联系人子集的信息，dmdata3.sav 包含未收到测试邮件的其余联系人的信息。
- **dietstudy.sav。**该假设数据文件包含对 “Stillman diet” (Rickman, Mitchell, Dingman, 和 Dalen, 1974) 的研究结果。每个个案对应一个单独的主体，并记录其在实行饮食方案前后的体重（磅）以及甘油三酸酯的水平（毫克/100 毫升）。
- **dvdplayer.sav。**这是关于开发新的 DVD 播放器的假设数据文件。营销团队用原型收集了焦点小组数据。每个个案对应一个单独的被调查用户，并记录他们的人口统计信息及其对原型问题的回答。
- **german_credit.sav。**该数据文件取自加州大学欧文分校的 Repository of Machine Learning Databases (Blake 和 Merz, 1998)中的 “German credit” 数据集。
- **grocery_1month.sav。**该假设数据文件是在数据文件 grocery_coupons.sav 的基础上加上了每周购物“累计”，所以每个个案对应一个单独的客户。所以，一些每周更改的变量消失了，而且现在记录的消费金额是为期四周的研究过程中的消费金额之和。

- **grocery_coupons.sav。**该假设数据文件包含由重视顾客购物习惯的杂货连锁店收集的调查数据。对每位顾客调查四周，每个个案对应一个单独的顾客周，并记录有关顾客购物地点和方式的信息（包括那一周里顾客在杂货上的消费金额）。
- **guttman.sav。** Bell (Bell, 1961)创建了一个表，用来阐释可能的社会群体。Guttman (Guttman, 1968)引用了该表的一部分，其中包括五个变量，用于描述以下七个理论社会群体的社会交往、对群体的归属感、成员的物理亲近度以及关系正式性：观众（比如在足球比赛现场的人们）、听众（比如在剧院或听课堂讲座的人们）、公众（比如报纸或电视观众）、组织群体（与观众类似但具有紧密的关系）、初级群体（关系密切）、次级群体（自发组织）及现代社区（因在物理上亲近而导致关系松散并需要专业化服务）。
- **health_funding.sav。**该假设数据文件包含关于保健基金（每 100 人的金额）、发病率（每 10,000 人的比率）以及保健提供商拜访率（每 10,000 的比率）的数据。每个个案代表不同的城市。
- **hivassay.sav。**该假设数据文件涉及某药物实验室在开发用于检测 HIV 感染的快速化验方面的举措。化验结果为八个加深的红色阴影，如果有更深的阴影则表示感染的可能性很大。用 2,000 份血液样本来进行实验室试验，其中一半受到 HIV 感染而另一半没有受到感染。
- **hourlywagedata.sav。**该假设数据文件涉及在政府机关和医院工作的具有不同经验水平的护士的时薪。
- **insurance_claims.sav。**该假设数据文件涉及某保险公司，该公司希望构建一个模型用于标记可疑的、具有潜在欺骗性的理赔。每个个案代表一次单独的理赔。
- **insure.sav。**该假设数据文件涉及某保险公司，该公司正在研究指示客户是否会根据 10 年的人寿保险合同提出理赔的风险因子。数据文件中的每个个案代表一副根据年龄和性别进行匹配的合同，其中一份记录了一次理赔而另一份则没有。
- **judges.sav。**该假设数据文件涉及经过训练的裁判（加上一个体操爱好者）对 300 次体操表演给出的分数。每行代表一次单独的表演；裁判们观看相同的表演。
- **kinship_dat.sav。**Rosenberg 和 Kim (Rosenberg 和 Kim, 1975) 开始分析 15 个亲属关系项（伯母、兄弟、表兄妹、女儿、父亲、孙女、祖父、祖母、孙子、母亲、侄子或外甥、侄女或外甥女、姐妹、儿子和叔叔）。他们让四组大学生（两组女同学，两组男同学）根据相似程度将各项排序。他们让其中的两组同学（一组女同学，一组男同学）进行了两次排序，第二次排序和第一次排序采取的标准不同。这样，一共得到六组“源”。每个源对应一个 15×15 的相似性矩阵，其单元格中的值等于源中的人数减去此源中对象被划分的次数。
- **kinship_ini.sav。**该数据文件包含 kinship_dat.sav 的三维解的初始配置。
- **kinship_var.sav。**该数据文件包含自变量 gender、gener(ation) 和 degree (of separation)，这些变量可用于解释 kinship_dat.sav 的解的维数。具体而言，它们可用来将解的空间限制为这些变量的线性组合。
- **marketvalues.sav。**该数据文件涉及 1999 - 2000 年间 Algonquin，Ill. 地区新的房屋开发中的住房销售。这些销售仅仅来自公众记录。
- **nhis2000_subset.sav。**美国健康访问调查 (NHIS) 是针对美国全体公民的大型人口调查。该调查对美国的具有全国代表性的家庭样本进行了面对面的访问，并获取了每个家庭的成员的健康行为和健康状态的人口统计信息和观察数据。该数据文件包含取自 2000 年调查信息的子集。国家健康统计中心。2000 年美国健康访问调查。公用数据文件和文档。

ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/。2003 年发布。

- **ozone.sav。**这些数据包含了用来根据其余变量预测臭氧浓度的六个气象变量的 330 个观察值。在以前的研究人员中，(Breiman 和 Friedman(F), 1985) 和 (Hastie 和 Tibshirani, 1990) 发现了这些变量之间的非线性，这妨碍了标准回归方法。
- **pain_medication.sav。**该假设数据文件包含用于治疗慢性关节炎疼痛的抗炎药的临床试验结果。我们感兴趣的是该药见效的时间以及它和现有药物的比较。
- **patient_los.sav。**该假设数据文件包含被医院确诊为疑似心肌梗塞（即 MI 或“心脏病发作”）的患者的治疗记录。每个个案对应一位单独的患者，并记录与其住院期有关的一些变量。
- **patlos_sample.sav。**该假设数据文件包含在治疗心肌梗塞（即 MI 或“心脏病发作”）期间收到溶解血栓剂的患者的样本治疗记录。每个个案对应一位单独的患者，并记录与其住院期有关的一些变量。
- **poll_cs.sav。**该假设数据文件涉及民意测验专家在确定正式立法前公众对法案的支持水平方面的举措。个案对应注册的选民。每个个案记录选民居住的县、镇、区。
- **poll_cs_sample.sav。**该假设数据文件包含在 poll_cs.sav 中列出的选民的样本。该样本是根据 poll.csplan 中指定的设计来选取的，而且该数据文件记录包含概率和样本权重。请注意，由于该抽样计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (poll_jointprob.sav) 包含联合选择概率。在选取了样本之后，对应于选民人群统计信息及其对提交法案的意见的附加变量将被收集并添加到数据文件。
- **property_assess.sav。**该假设数据文件涉及某县资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应过去一年中县里所出售的资产。数据文件中的每个个案记录资产所在的镇、最后评估资产的评估员、该次评估距今的时间、当时的估价以及资产的出售价格。
- **property_assess_cs.sav。**该假设数据文件涉及某州资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应该州的资产。数据文件中的每个个案记录资产所在的县、镇和区，最后一次评估距今的时间以及当时的估价。
- **property_assess_cs_sample.sav。**该假设数据文件包含在 property_assess_cs.sav 中列出的资产的样本。该样本是根据 property_assess.csplan 中指定的设计来选取的，而且该数据文件记录包含概率和样本权重。在选取了样本之后，附加变量 Current value 将被收集并添加到数据文件。
- **recidivism.sav。**该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应先前的一名罪犯，并记录其人口统计信息和第一次犯罪的详细资料；如果在第一次被捕后两年内又第二次被捕，则还将记录两次被捕间隔的时间。
- **recidivism_cs_sample.sav。**该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应在 2003 年 6 月期间第一次被捕释放的先前的一名罪犯，并记录其人口统计信息和第一次犯罪的详细资料，及其第二次被捕的数据（如果发生在 2006 年 6 月底之前）。根据 recidivism_cs.csplan 中指定的抽样计划从抽样部门选择罪犯；该计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (recidivism_cs_jointprob.sav) 包含联合选择概率。
- **rfm_transactions.sav。** 此假设数据文件包含购买交易数据，即每笔交易的购买日期、购买商品和消费金额。

- **salesperformance.sav。**这是关于评估两个新的销售培训课程的假设数据文件。60 名员工被分成 3 组且都接受标准的培训。另外，组 2 接受技术培训；组 3 接受实践教程。在培训课程结束时，对每名员工进行测验并记录他们的分数。数据文件中的每个个案代表一名单独的受训者，并记录其被分配到的组以及测验的分数。
- **satisf.sav。**该假设数据文件涉及某零售公司在 4 个商店位置所进行的满意度调查。总共对 582 位客户进行了调查，每个个案代表一位单独客户的回答。
- **screws.sav。**该数据文件包含关于螺钉、螺栓、螺母和图钉的特征的信息(Hartigan, 1975)。
- **shampoo_ph.sav**这是关于某发制品厂的质量控制的假设数据文件。在规定的时间间隔对六批独立输出的产品进行检测并记录它们的 pH 值。目标范围是 4.5 - 5.5。
- **ships.sav。**在别处被提出和分析的(McCullagh 等。, 1989)关于波浪对货船造成的损坏的数据集。在给定了船的类型、建造工期和服务期后，可以根据以泊松比率发生来为事件计数建模。在因子交叉分类构成的表格中，每个单元格的分类汇总服务月数提供遇到风险的值。
- **site.sav。**该假设数据文件涉及某公司在为扩展业务而选择新址方面的举措。该公司聘请了两名顾问分别对选址进行评估，除了提供长期报告外，他们还要以“前景颇佳”、“前景良好”或“前景不佳”来对每个选址进行总结。
- **smokers.sav。**该数据文件摘自 1998 年全国家庭药物滥用调查并且是美国家庭的概率样本。(http://dx.doi.org/10.3886/ICPSR02934) 因此，分析该数据文件的第一步应该是对数据进行加权以反映总体趋势。
- **stocks.sav** 该假设数据文件包含某一年的股票价格和成交量。
- **stroke_clean.sav。**该假设数据文件包含某医学数据库在经过“数据准备”选项中的过程清理后的状态。
- **stroke_invalid.sav。**该假设数据文件包含某医学数据库的初始状态及一些数据输入错误。
- **stroke_survival。**此假设数据文件涉及正在研究结束缺血性中风后复元计划的患者存活时间的研究人员面临着很多挑战。中风后，记录心肌梗塞、缺血性中风或出血性中风的发生及其时间。样本为左侧截短，因为只包含在中风后管理的复元计划结束后存活的患者。
- **stroke_valid.sav。**该假设数据文件包含在使用“验证数据”过程检查值后，某医学数据库的状态。它仍包含潜在异常个案。
- **survey_sample.sav。**此数据文件包含调查数据，包括人口统计学数据和各种态度测量。它基于 1998 NORC 综合社会调查的变量子集，但某些数据值已经过修改，并添加了其他虚拟变量以供演示用途。
- **telco.sav。**该假设数据文件涉及某电信公司在减少客户群中的客户流失方面的举措。每个个案对应一个单独的客户，并记录各类人口统计和服务用途信息。
- **telco_extra.sav。**该数据文件与 telco.sav 数据文件类似，但删除了 “tenure” 和经对数转换的客户消费变量，代替它们的是标准化的对数转换客户消费变量。
- **telco_missing.sav。** 该数据文件是 telco.sav 数据文件的子集，但某些人口统计数据值已被缺失值替换。

- **testmarket.sav。**该假设数据文件涉及某快餐连锁店为其菜单添加新项目的计划。有三种可能的促销新产品的活动，所以会在多个随机选择的市场中的地点引入新的项目。在每个地点采用不同的促销方式，并记录新项目前四周的每周销售情况。每个个案对应单独地点的一周。
- **testmarket_1month.sav。**该假设数据文件是在数据文件 testmarket.sav 的基础上加上了每周销售“累计”，所以每个个案对应一个单独的地点。所以，一些每周更改的变量消失了，而且现在记录的销售是为期四周的研究过程中的销售之和。
- **tree_car.sav。**该假设数据文件包含人口统计和车辆购买价格数据。
- **tree_credit.sav。**该假设数据文件包含人口统计和银行贷款历史数据。
- **tree_missing_data.sav** 该假设数据文件包含具有大量缺失值的人口统计和银行贷款历史数据。
- **tree_score_car.sav。**该假设数据文件包含人口统计和车辆购买价格数据。
- **tree_textdata.sav。**这是一个只有两个变量的样本数据文件，主要打算在指定测量级别和值标签之前显示变量的默认状态。
- **tv-survey.sav。**该假设数据文件涉及由某电视演播室进行的一项关于是否要继续制作一档成功的节目的调查。906 位调查对象被问及他们在各种情况下是否会收看该节目。每行代表一位单独的调查对象；每列代表一种单独的情况。
- **ulcer_recurrence.sav。**此文件包含某项研究的部分信息，该研究旨在比较两种用来防止溃疡复发的治疗的功效。它提供了区间数据的优秀示例并且已在别处被提出和分析(Collett, 2003)。
- **ulcer_recurrence_recoded.sav。**该文件重新组织 ulcer_recurrence.sav 中的信息以允许为研究的每个区间的事件概率建模而不是简单地为研究结束事件概率建模。它已在别处被提出和分析(Collett 等。, 2003)。
- **verd1985.sav。**该数据文件涉及某项调查(Verdegaal, 1985)。该调查记录了 15 个主体对 8 个变量的响应。需要处理的变量被分成 3 个集。数据集 1 包含 年龄 和 婚姻；数据集 2 包含 宠物 和 新闻；数据集 3 包含 音乐 和 居住。宠物被尺度化为多名义而年龄被尺度化为有序；所有其他变量都被尺度化为单名义。
- **virus.sav。**该假设数据文件涉及某因特网服务提供商 (ISP) 在确定病毒对其网络的影响方面的举措。他们从发现病毒到威胁得以遏制这段时间内跟踪其网络上受感染的电子邮件的流量的（近似）百分比。
- **wheeze_steubenville.sav。**这是关于空气污染对儿童健康影响的纵向研究的一个子集(Ware, Dockery, Spiro III, Speizer, 和 Ferris Jr., 1984)。这些数据包含儿童的气喘状况的重复二分类测量（这些儿童来自 Steubenville，Ohio，年龄为 7 到 10 岁），以及母亲在研究的第一年中是否为吸烟者的固定记录。
- **workprog.sav。**该假设数据文件涉及一份尝试为弱势群体提供较好的工作的政府工作计划。文件后还有一个潜在计划参与者的样本，其中一些参与者是被随机选择来参加该计划的，而其他参与者则不是。每个个案代表一位单独的计划参与者。
- **worldsales.sav** 该假设数据文件包含按不同大洲和产品列出的销售收入。

附录
B

注意事项

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家/地区中不提供在本文档中讨论的产品、服务或功能。请咨询您当地的 IBM 代表以了解有关您所在地区当前可用产品和服务的信息。任何对 IBM 产品、程序或服务的引用，并不意味着仅可使用这些 IBM 产品、程序或服务。作为替代，可以使用任何功能相当的产品、程序或服务，前提是不侵犯任何 IBM 知识产权。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。

在本文档中介绍的主题可能涉及 IBM 的专利或申请中的专利。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, Armonk, NY 10504-1785, U.S.A.

要查询双字节字符集 (DBCS) 相关许可证信息，请联系所在国家/地区中的 IBM 知识产权部门，或者以书面形式将查询函件发送至：

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

以下段落不适用于英国或此类条款与当地法律不符的其他国家/地区： INTERNATIONAL BUSINESS MACHINES 公司“按原样”提供本出版物，不保证任何明示或暗示，包括但不限于对非侵权性、适销性或对特定用途适用性的暗示担保。有些州/省不允许针对某些交易的明示或暗示免责条款，因此本声明可能不适用于您。

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 可能随时对本出版物中所述的产品和/或程序进行改进和/或更改，恕不另行通知。

在本信息材料中对任何非 IBM 网站的引用仅为了方便用户，并不以任何方式表明对这些网站的认可。这些网站上的材料并非本 IBM 产品材料的一部分，您对这些网站的使用需自担风险。

IBM 可以自认为适当并且不会对您构成任何约束的任何方式使用或分发您提供的任何信息。

如果本程序的受许可方试图了解有关程序的信息以启用：(i) 在独立创建的程序和其他程序（包括本程序）之间交换信息；(ii) 相互使用交换的信息，则应联系：

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

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。

在本文档中介绍的受许可保护程序，及其所有受许可保护材料由 IBM 在双方签署的“IBM 客户协议”、“IBM 国际程序许可证协议”或任何其他等同协议下提供。

非 IBM 产品的相关信息来自这些产品的供应商，及其发布的公告或其他公开来源。IBM 并未测试过这些产品，无法确认有关非 IBM 产品的性能准确性、兼容性或任何其他声明。有关非 IBM 产品功能的问题应由这些产品的供应商负责。

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。所有这些名称均为虚构，与真实商业企业使用的名称和地址的任何雷同纯属巧合。

如果您正在查阅此信息的软拷贝，照片和彩色插图可能不会显示。

商标

IBM、IBM 徽标、ibm.com 和 SPSS 是 IBM Corporation 的商标，在全球许多司法辖区注册。有关最新的 IBM 商标列表，请访问网页 http://www.ibm.com/legal/copytrade.shtml。

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。

Java 和所有基于 Java 的商标和徽标是 Sun Microsystems, Inc. 在美国、其他国家或地区或两者的商标。

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。

此产品使用 Polar 工程咨询公司的 WinWrap Basic，版权 1993 年-2007 年，http://www.winwrap.com。

其他产品和服务名称可能是 IBM 或其他公司的商标。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。

参考书目

Bell, E. H. 1961. Social foundations of human behavior:Introduction to the study of sociology. New York: Harper & Row.

Bishop, C. M. 1995. Neural Networks for Pattern Recognition, 3rd ed. Oxford: Oxford University Press.

Blake, C. L., 和 C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., 和 J. H. Friedman(F). 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, .

Collett, D. 2003. Modelling survival data in medical research, 2 ed. Boca Raton: Chapman & Hall/CRC.

Green, P. E., 和 V. Rao. 1972. Applied multidimensional scaling. Hinsdale, Ill.: Dryden Press.

Green, P. E., 和 Y. Wind. 1973. Multiattribute decisions in marketing:A measurement approach. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. Psychometrika, 33, .

Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley and Sons.

Hastie, T., 和 R. Tibshirani. 1990. Generalized additive models. London: Chapman and Hall.

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

Kennedy, R., C. Riquier, 和 B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. Journal of Targeting, Measurement, and Analysis for Marketing, 5, .

McCullagh, P., 和 J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.

Price, R. H., 和 D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. Journal of Personality and Social Psychology, 30, .

Rickman, R., N. Mitchell, J. Dingman, 和 J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. Journal of the American Medical Association, 228, .

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

Rosenberg, S., 和 M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. Multivariate Behavioral Research, 10, .

Tao, K. K. 1993. A closer look at the radial basis function (RBF) networks. 输入: Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers, A. Singh, ed. Los Alamitos, Calif.: IEEE Comput. Soc. Press.

Uykan, Z., C. Guzelis, M. E. Celebi, 和 H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN. IEEE Transactions on Neural Networks, 11, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, 和 H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents:A longitudinal perspective. British Journal of Psychiatry, 170, .

Verdegaal, R. 1985. Meer sets analyse voor kwalitatieve gegevens (in Dutch). Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, 和 B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. American Review of Respiratory Diseases, 129, .

细, T. L. 1999. Feedforward Neural Network Methodology, 3rd ed. New York: Springer-Verlag.

索引