IBM SPSS Decision Trees 26

注释

使用本信息及其支持的产品之前，请阅读 第 15 页的『通知』 中的信息。

产品信息

此版本适用于 IBM SPSS Statistics V26.0.0 及所有后续发行版和修订，除非在新版本中另有说明。

目录

决策树

SPSS® Statistics Professional Edition 或"决策树"选项中包含以下决策树功能。

创建决策树

"决策树"过程创建基于树的分类模型。它将个案分为若干组，或根据自变量（预测变量）的值预测因变量（目标变量）的值。此过程为探索性和证实性分类分析提供验证工具。

此过程可以用于：

分段。 确定可能成为特定组成员的人员。

分层。 将个案指定为几个类别之一，如高风险组、中等风险组和低风险组。

预测。 创建规则并使用它们预测将来的事件，如某人将拖欠贷款或者车辆或住宅潜在转售价值的可能性。

数据降维和变量过滤。 从大的变量集中选择有用的预测变量子集，以用于构建正式的参数模型。

交互确定。 确定仅与特定子组有关的关系，并在正式的参数模型中指定这些关系。

类别合并和连续变量分箱化。 以最小的损失信息对组预测类别和连续变量进行重新编码。

示例。一家银行希望根据贷款申请人是否表现出合理的信用风险来对申请人进行分类。根据各种因素（包括过去客户的已知信用等级），您可以构建模型以预测客户将来是否可能拖欠贷款。

基于树的分析提供了一些引人注意的功能：

- 通过分析功能，您可以确定具有高风险或低风险的同类组。
- 还可轻松构建用于预测个别个案的规则。

数据注意事项

数据。因变量和自变量可以是：

- 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。

频率权重如果加权有效，则将分数权重四舍五入为最接近的整数；所以，为权重值小于 0.5 的个案指定权重 0，因而会从分析中排除它们。

假设。此过程假定已经为所有分析变量指定适当的测量级别，一些功能假定分析中包括的因变量的所有值都定义了值标签。

- **测量级别。**测量级别影响树计算；因此，应该为所有变量指定适当的测量级别。缺省情况下，假定数值变量是刻度变量，而字符串变量假定为名义变量，这可能没有准确地反映真实的测量级别。变量列表中每个变量旁的图标标识变量类型。

可以暂时更改变量的测量级别，方法是在源变量列表中右键单击该变量，然后从弹出菜单中选择测量级别。

- **值标签**。此过程的对话框界面假设分类（名义、有序）因变量的所有非缺失值均已定义值标签或未定义值标签。除非分类因变量至少有两个非缺失值具有值标签，否则某些功能将不可用。如果至少两个非缺失值已经定义了值标签，则将从分析中排除带有其他没有值标签的值的所有个案。

获取决策树

1. 从菜单中选择：

   分析 > **分类** > **树...**
2. 选择一个因变量。
3. 选择一个或多个自变量。
4. 选择生长法。

根据需要，您可以：

- 更改源列表中所有变量的测量级别。
- 强制自变量列表中的第一个变量作为第一个拆分变量进入模型。
- 选择定义个案对树生长过程的影响程度的影响变量。影响值较低的个案影响较小；而影响值较高的个案影响较大。影响变量值必须为正。
- 验证树。
- 自定义树生长条件。
- 将终端节点编号、预测值和预测概率保存为变量。
- 以 XML (PMML) 格式保存模型。

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，就会显示测量级别警告。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。

手动指定
: 列出测量级别未知的所有字段。您可以对这些字段指定测量级别。此外，也可以在数据编辑器的"变量列表"窗格中指定测量级别。

由于测量级别对于此过程很重要，因此除非所有字段均定义有测量级别，否则您无法运行此过程。

更改测量级别

1. 右键单击源列表中的变量。
2. 从弹出菜单中选择测量级别。

这将暂时更改测量级别以用于"决策树"过程。

生长法

可用的生长法如下：

CHAID
: 卡方自动交互检测。在每一步，CHAID 选择与因变量有最强交互作用的自变量（预测变量）。如果每个预测变量的类别与因变量并非显著不同，那么合并这些类别。

穷举 *CHAID (Exhaustive CHAID)*
: 改进的 CHAID 方法，可检查每个预测变量的所有可能分裂。

CRT
: 分类和回归树。CRT 将数据拆分为若干尽可能与因变量同质的段。所有个案中因变量值都相同的终端节点是同质的"纯"节点。

QUEST
: 快速、无偏、有效的统计树。一种快速方法，它可避免其他方法对具有许多类别的预测变量的偏倚。只有在因变量是名义变量时才能指定 QUEST。

每种方法都有其各自的优点和限制，其中包括：

表 *1.* 生长法的功能.

功能	CHAID*	CRT	QUEST
基于卡方**	X		
替代自变量（预测变量）		X	X
树修剪		X	X
多阶节点拆分	X		
二元节点拆分		X	X
影响变量	X	X	
先验概率		X	X
误分类成本	X	X	X
快速计算	X		X

*包括穷举 CHAID。

**QUEST 也将卡方测量用于名义自变量。

选择类别

对于分类（名义、有序）因变量，可以：

- 控制将哪些类别包括在分析中。
- 确定目标类别。

包括/排除类别

可以将分析限制为因变量的特定类别。

- 因变量的值在"排除"列表中的个案不会包括在分析中。
- 对于名义因变量，您也可以在分析中包括用户缺失的类别。（缺省情况下，用户缺失的类别显示在"排除"列表中。）

目标类别

选定（选中）的类别被视为分析中主要对其感兴趣的类别。例如，如果主要对确定最可能拖欠贷款的那些人感兴趣，则可能选择"bad"信用等级类别作为目标类别。

- 没有缺省的目标类别。如果未选定任何类别，则某些分类规则选项和与增益相关的输出将不可用。

- 如果选定了多个类别，则为每个目标类别生成单独的增益表和图表。
- 将一个或多个类别指定为目标类别，对树模型、风险估计或误分类结果没有影响。

类别

此对话框要求因变量具有已定义的值标签。除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

包括/排除类别以及选择目标类别

1. 在"决策树"主对话框中，选择带有两个或两个以上已定义值标签的分类（名义或有序）因变量。
2. 单击**类别**。

验证

通过验证可以评估树结构广义化为更大总体的程度。可以使用两种验证方法：交叉验证和分割样本验证。

交叉验证(C)

交叉验证将样本分割为许多子样本（或**样本群**）。然后，生成树模型，并依次排除每个子样本中的数据。第一个树基于第一个样本群的个案之外的所有个案，第二个树基于第二个样本群的个案之外的所有个案，依此类推。对于每个树，估计其误分类风险的方法是将树应用于生成它时所排除的子样本。

要点：选择修剪时，交叉验证不可用于 CRT 和 Quest 方法。

- 最多可以指定 25 个样本群。该值越大，每个树模型中排除的个案数就越小。
- 交叉验证生成单个最终树模型。最终树经过交叉验证的风险估计计算为所有树的风险的平均值。

拆分样本验证

对于拆分样本验证，模型是使用训练样本生成的，并在延续样本上进行测试。

- 您可以指定训练样本大小（表示为样本总大小的百分比），或将样本拆分为训练样本和测试样本的变量。
- 如果使用变量定义训练样本和测试样本，则将变量值为 1 的个案指定给训练样本，并将所有其他个案指定给测试样本。该变量不能是因变量、权重变量、影响变量或强制的自变量。
- 您可以同时显示训练样本和测试样本的结果，或者仅显示测试样本的结果。
- 对于小的数据文件（个案数很少的数据文件），应该谨慎使用拆分样本验证。训练样本很小可能会导致很差的模型，因为在某些类别中，可能没有足够的个案使树充分生长。

验证决策树

1. 在"决策树"主对话框中，单击**验证**。
2. 选择**交叉验证**或**拆分样本验证**。

注：这两种验证方法均随机地将个案指定给样本组。如果希望能够在后续分析中再现完全相同的结果，则应该在第一次运行分析之前设置随机数种子（"转换"菜单，"随机数字生成器"），然后将种子重置为该值以用于后续分析。

树生长条件

可用的生长条件可能取决于生长法、因变量的测量级别或这两者的组合。

增长限制

使用"生长限制"对话框，可以限制树中的级别数，以及控制父节点和子节点的最小个案数。

最大树深度

控制根节点下的最大增长级别数。对于 CHAID 和穷举 CHAID 方法，**自动**设置将树限制为根节点下的三个级别；而对于 CRT 和 QUEST 方法，则限制为根节点下的五个级别。

最小个案数

控制节点的最小个案数。不会拆分不满足这些条件的节点。

- 增大最小值往往会生成具有更少节点的树。
- 而减小最小值则会生成具有更多节点的树。

对于个案数目很小的数据文件，父节点的缺省值（100 个个案）和子节点的缺省值（50 个个案）有时可能导致树在根节点下没有任何节点；在这种情况下，减小最小值可能产生更有用的结果。

指定生长限制

1. 在"决策树"主对话框中，单击**生长限制**。

CHAID 条件

对于 CHAID 和穷举 CHAID 方法，您可以控制：

显著性水平

您可以控制用于拆分节点和合并类别的显著性值。对于这两个条件，缺省的显著性水平都是 0.05。

拆分节点

值必须大于 0 且小于 1。较小的值往往会生成节点较少的树。

合并类别

值必须大于 0 且小于或等于 1。要阻止合并类别，请指定值 1。对于标度自变量，这意味着最终树中变量的类别数是指定的区间数（缺省值是 10）。请参阅主题第 6 页的『CHAID 分析的刻度区间』，了解更多信息。

卡方统计

对于有序因变量，用于确定节点拆分和类别合并的卡方是使用似然比方法计算的。对于名义因变量，可以选择以下方法：

Pearson

此方法提供更快的计算，但是对于小样本应该谨慎使用它。这是缺省方法。

似然比

此方法比 Pearson 方法更稳健，但是所用的计算时间更长。对于小样本，这是首选的方法。

模型估计

对于名义和有序因变量，可以指定：

最大迭代次数

缺省值为 100。如果树由于达到最大迭代次数而停止生长，您可能希望增大最大值，或更改控制树生长的一个或多个其他条件。

期望单元格频率的最小变化

该值必须大于 0 且小于 1。缺省值为 0.05。较小的值往往会产生具有较少节点的树。

使用 Bonferroni 方法调整显著性值

对于多个比较，使用 Bonferroni 方法调整用于合并和拆分条件的显著性值。这是缺省值。

在节点内允许重新拆分合并类别

除非明显阻止类别合并，否则该过程将尝试将自变量（预测变量）类别合并在一起，以产生描述模型的最简单的树。此选项允许该过程重新拆分合并的类别（如果这样可以提供更好的方案）。

指定 CHAID 条件

1. 在"决策树"主对话框中，选择 **CHAID** 或**穷举 CHAID** 作为生长法。
2. 单击 **CHAID**。

***CHAID* 分析的刻度区间：** 在 CHAID 分析中，刻度自变量（预测变量）在分析之前始终分段到离散组（例如，0–10、11–20、21–30 等）中。您可以控制初始/最大组数（尽管该过程可能在初始拆分后合并连续组）：

固定数目
所有的刻度自变量最初都分段到相同数量的组中。缺省值为 10。

定制 每个刻度自变量最初都分段到该变量所指定数量的组中。

指定标度自变量的区间

1. 在"决策树"主对话框中，选择一个或多个标度自变量。
2. 对于**生长法**，请选择 **CHAID** 或**穷举 CHAID**。
3. 单击**区间**。

在 CRT 和 QUEST 分析中，所有拆分均为二元的，而且刻度和有序自变量的处理方式是相同的；因此，无法为刻度自变量指定多个区间。

CRT 条件

CRT 生长法尝试最大化节点内的同质性。对不代表同质个案子集的节点，它的程度显示为**杂质**。例如，其中所有个案都具有相同的因变量值的终端节点是无需进一步拆分（因为它是"纯的"）的同质节点。

可以选择用于测量杂质的方法，以及拆分节点所需的杂质中的最小减少值。

杂质测量
对于刻度因变量，使用最小二乘偏差 (LSD) 测量杂质。它为节点内的方差，并根据任意频率权重或影响值进行调整。对于分类（名义、有序）因变量，可以选择杂质测量：

Gini 找到可以根据因变量的值最大化子节点同质性的拆分。对于因变量的每个类别，Gini 基于成员身份的平方概率。它在节点中的所有个案都属于单个类别时达到其最小值（零）。这是缺省测量。

两分法
因变量的类别分组为两个子类。找到最适合于分隔两个组的拆分。

顺序两分法
与两分法相似，但它只能对相邻类别进行分组。此度量仅可用于有序因变量。

改进中的最小更改
这是拆分节点所需的杂质中的最小减少值。缺省值为 0.0001。较大的值往往会产生节点较少的树。

指定 CRT 条件

1. 对于**生长法**，请选择 **CRT**。
2. 单击 **CRT**。

QUEST 条件

对于 QUEST 方法，可以指定用于拆分节点的显著性水平。除非显著性水平小于或等于指定的值，否则自变量不能用于拆分节点。该值必须大于 0 且小于 1。缺省值为 0.05。较小的值往往会从最终模型中排除较多的自变量。

指定 QUEST 条件

1. 在"决策树"主对话框中，选择一个名义因变量。
2. 对于**生长法**，请选择 **QUEST**。
3. 单击 **QUEST**。

修剪树

使用 CRT 和 QUEST 方法，可以通过**修剪**树来避免模型过度拟合：在满足停止生长的条件之前保持树处于生长状态，然后根据指定的最大风险差值，自动将其修剪到最小子树。以标准误差表示风险值。缺省值为 1。该值必须为非负数。要获取具有最小风险的子树，请指定 0。

要点：选择修剪时，交叉验证不可用于 CRT 和 Quest 方法。

修剪树

1. 在"决策树"主对话框中，对于**生长法**，选择 **CRT** 或 **QUEST**。
2. 单击**修剪**。

修剪与隐藏节点

创建修剪树时，从树中修剪的任何节点在最终树中都是不可用的。您可以以交互方式隐藏和显示最终树中的选定子节点，但是不能显示在创建树的过程中修剪的节点。

替代变量

CRT 和 QUEST 可以将**替代变量**用于自变量（预测变量）。对于缺失该变量的值的个案，将使用与原始变量高度相关的其他自变量进行分类。这些备用预测变量称为替代变量。可以指定要在模型中使用的最大替代变量数。

- 缺省情况下，最大替代变量数比自变量数小 1。换句话说，针对每个自变量，其他的所有自变量均可能被用作替代变量。
- 如果不希望模型使用替代变量，请指定 0 作为替代变量数。

指定替代变量

1. 在"决策树"主对话框中，对于**生长法**，选择 **CRT** 或 **QUEST**。
2. 单击**替代变量**。

选项(O)

可用选项可能取决于生长法、因变量的测量级别和/或为因变量的值定义的值标签是否存在。

误分类成本

对于分类（名义、有序）因变量，通过误分类成本，可以包括有关与错误分类关联的相对惩罚的信息。例如：

- 拒绝为信用良好的客户发放贷款的成本，可能与为之后拖欠贷款的客户提供贷款的成本不同。
- 将患有心脏病的高风险个人误分类为低风险的成本，可能比将低风险的个人误分类为高风险的成本要高得多。
- 向也许不会回复的个人发送大量邮件的成本可能非常低，但不向可能回复的个人发送邮件的成本却相对较高（在失去的收入方面）。

注：除非分类因变量至少有两个值已定义值标签，否则这个"误分类成本"对话框不可用。

指定误分类成本

1. 在"决策树"主对话框中，选择带有两个或两个以上已定义值标签的分类（名义或有序）因变量。
2. 单击**误分类成本**。
3. 单击**自定义**。
4. 在"预测类别"网格中输入一个或多个误分类成本。值必须为非负数。（在对角线上表示的正确分类始终为 0。）

填充矩阵

在许多情况下，可能希望成本是对称的，即，将 A 误分类为 B 的成本与将 B 误分类为 A 的成本是相同的。使用以下控件可以更轻松地指定对称的成本矩阵：

复制下三角形

将矩阵的下三角形中的值（在对角线之下）复制到对应的上三角形单元格中。

复制上三角形

将矩阵的上三角形中的值（在对角线之上）复制到对应的下三角形单元格中。

使用单元格平均值

对于矩阵的每一半中的每个单元格，对这两个值（上三角形和下三角形）进行平均，并用平均值替换这两个值。例如，如果将 A 误分类为 B 的成本为 1，而将 B 误分类为 A 的成本为 3，则此控件会将这两个值都替换为平均值 (1+3)/2 = 2。

利润

对于分类因变量，可以将收入值和支出值指定给因变量的水平。

- 利润是通过收入减去支出计算出来的。
- 利润值影响增益表中的平均利润值和 ROI（投资回报）值。但它们不影响树模型的基础结构。
- 收入值和支出值必须为数值型，且必须为网格中显示的因变量的所有类别指定它们。

注：此对话框要求因变量具有已定义的值标签。除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

指定利润

1. 在"决策树"主对话框中，选择带有两个或两个以上已定义值标签的分类（名义或有序）因变量。
2. 单击**利润**。
3. 单击**自定义**。
4. 输入网格中列出的所有因变量类别的收入值和支出值。

先验概率

对于具有分类因变量的 CRT 和 QUEST 树，可以指定组成员身份的先验概率。**先验概率**是在了解有关自变量（预测变量）值的任何信息之前，对因变量的每个类别的总体相对频率的评估。使用先验概率有助于更正由不代表整个总体的样本中的数据导致的树的任何生长。

从训练样本获取（先验）

如果数据文件中因变量值的分布代表总体分布，则使用此设置。如果使用的是拆分样本验证，则使用训练样本中的个案分布。

注：因为在拆分样本验证中个案会随机分配给训练样本，所以事先不知道训练样本中个案的实际分布。请参阅主题第 4 页的『验证』，了解更多信息。

各类别之间相等

如果因变量的类别在总体中是以相等方式表示的，则使用此设置。例如，如果有四个类别，则每个类别中的个案约为 25%。

定制 对于网格中列出的每个因变量类别，输入一个非负值。这些值可以是比例、百分比、频率计数或表示各类别之间值分布的任何其他值。

使用错误分类成本调整先验

如果定义自定义的误分类成本，则可以根据这些成本调整先验概率。请参阅主题第 7 页的『误分类成本』，了解更多信息。

注：此对话框要求因变量具有已定义的值标签。除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

指定先验概率

1. 在"决策树"主对话框中，选择带有两个或两个以上已定义值标签的分类（名义或有序）因变量。
2. 对于**生长法**，请选择 **CRT** 或 **QUEST**。
3. 单击**先验概率**。

得分

对于具有有序因变量的 CHAID 和穷举 CHAID，可以将自定义得分指定给每个因变量类别。得分定义因变量类别的顺序以及它们之间的距离。您可以使用得分来增大或减小有序值之间的相对距离或更改值的顺序。

为每个类别使用序数等级

指定给最低的因变量类别的得分是 1，为次最高类别指定的得分是 2，依此类推。这是缺省值。

定制 对于网格中列出的每个因变量类别，输入一个数值型得分值。

注：此对话框要求因变量具有已定义的值标签。除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

示例

表 2. 定制得分值

值标签	原始值	得分
Unskilled	1	1
Skilled manual	2	4
Clerical	3	4.5
Professional	4	7
Management	5	6

- 得分增大了 *Unskilled* 和 *Skilled* 之间的相对距离，而减小了 *Skilled manual* 和 *Clerical* 之间的相对距离。
- 得分交换了 *Management* 和 *Professional* 的顺序。

指定得分

1. 在"决策树"主对话框中，选择带有两个或两个以上已定义值标签的有序因变量。
2. 对于**生长法**，请选择 **CHAID** 或**穷举 CHAID**。
3. 单击**得分**。

缺失值(S)

"缺失值"对话框控制名义值、用户缺失值和自变量（预测变量）值的处理。

- 用户缺失的有序和刻度自变量值的处理随生长法的不同而不同。
- 名义因变量的处理在"类别"对话框中指定。请参阅 第 3 页的『选择类别』主题以获取更多信息。
- 对于有序和刻度因变量，始终排除具有系统缺失或用户缺失的因变量值的个案。

名义自变量的用户缺失值

视为缺失值
用户缺失值当作系统缺失值处理。系统缺失值的处理随生长法的不同而不同。

视为有效值
名义自变量的用户缺失值在树生长和分类中被视为普通值。

依赖于方法的规则

如果一些（而不是所有）自变量值是系统缺失或用户缺失的：

- 对于 CHAID 和穷举 CHAID，系统缺失和用户缺失的自变量值作为单个组合类别包括在分析中。对于刻度和有序自变量，算法首先使用有效值生成类别，然后确定是将缺失类别与其最类似的（有效的）类别合并，还是将其作为单独的类别保留。
- 对于 CRT 和 QUEST，从树生长过程中排除具有缺失自变量值的个案，但如果方法中包括替代变量，则使用替代变量对其进行分类。如果将名义用户缺失值视为缺失，同样也按此方式对其进行处理。请参阅主题 第 7 页的『替代变量』，了解更多信息。

指定名义自变量用户缺失处理

1. 在"决策树"主对话框中，至少选择一个名义自变量。
2. 单击**缺失值**。

保存模型信息

可以将模型中的信息另存为工作数据文件中的变量，也可以将整个模型以 XML (PMML) 格式保存到外部文件中。

保存变量

终端节点编号
为其指定每个个案的终端节点。该值是树节点编号。

预测值
模型所预测的因变量的分类（组）或值。

预测概率
与模型的预测关联的概率。为每个因变量类别保存一个变量。对刻度因变量不可用。

样本分配（训练/检验）
对于拆分样本验证，此变量指示在训练或检验样本中是否使用了某个案。对于训练样本，值为 1；对于检验样本，值为 0。只在选择了拆分样本验证时才可用。请参阅主题第 4 页的『验证』，了解更多信息。

将树模型导出为 XML
可以以 XML (PMML) 格式保存整个树模型。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。

培训样本

将模型写入指定的文件。对于拆分样本验证的树，这是训练样本的模型。

检验样本

将检验样本的模型写入指定文件。只在选择了拆分样本验证时才可用。

输出

可用的输出选项取决于生长法、因变量的测量级别和其他设置。

树显示

可以控制树的初始外观或完全取消树的显示。

树 缺省情况下，树形图包括在"输出"选项卡所显示的输出中。取消选择此选项可以从输出中排除树形图。

显示 这些选项控制"输出"选项卡中树形图的初始外观。也可以通过编辑生成的树修改所有这些属性。

方向 可以自上而下（根节点在顶部）、从左向右或从右向左地显示树。

节点内容

节点可以显示表、图表或这两者。对于分类因变量，表显示频率计数和百分比，而图表则是条形图。对于刻度因变量，表显示平均值、标准差、个案数和预测值，而图表则是直方图。

刻度 缺省情况下，大树会自动按比例缩小以适合页上的树。可以指定最大为 200% 的自定义缩放百分比。

自变量统计

对于 CHAID 和穷举 CHAID，统计包括 *F* 值（对于刻度因变量）或卡方值（对于分类因变量）以及显著性值和自由度。对于 CRT，显示改进值。对于 QUEST，显示 *F*、显著性值和自由度（对于刻度和有序自变量）；对于名义自变量，显示卡方、显著性值和自由度。

节点定义

节点定义显示在每个节点拆分中使用的自变量的值。

表格式树

树中每个节点的摘要信息，包括该节点的父节点编号、自变量统计、自变量值，刻度因变量的平均值和标准差，或者分类因变量的计数和百分比。

控制初始树显示

1. 在"决策树"主对话框中，单击**树**。

Statistics

可用的统计表取决于因变量的测量级别、生长法和其他设置。

模型

摘要 摘要包括所用的方法、模型中包括的变量以及已指定但未包括在模型中的变量。

风险 风险估计及其标准误差。对树的预测准确性的测量。

- 对于分类因变量，风险估计是在为先验概率和误分类成本调整后不正确分类的个案的比例。
- 对于刻度因变量，风险估计是节点中的方差。

分类表

对于分类（名义、有序）因变量，此表显示每个因变量类别的正确分类和不正确分类的个案数。对刻度因变量不可用。

成本、先验概率、得分和利润值

对于分类因变量，此表显示在分析中使用的成本、先验概率、得分和利润值。对刻度因变量不可用。

自变量(V)

对模型的重要性

对于 CRT 生长法，根据每个自变量（预测变量）对模型的重要性对其进行分类。对 QUEST 或 CHAID 方法不可用。

替代变量（按拆分）

对于 CRT 和 QUEST 生长法，如果模型包括替代变量，则在树中列出每个拆分的替代变量。对 CHAID 方法不可用。请参阅主题第 7 页的『替代变量』，了解更多信息。

节点性能

摘要 对于刻度因变量，该表包括因变量的节点编号、个案数和平均值。对于带有已定义利润的分类因变量，该表包括节点编号、个案数、平均利润和 ROI（投资回报）值。对不带已定义利润的分类因变量不可用。请参阅主题第 8 页的『利润』，了解更多信息。

按目标类别

对于带有已定义目标类别的分类因变量，该表包括按节点或百分位组显示的百分比增益、响应百分比和指标百分比（提升）。将对每个目标类别生成一个单独的表。对于不带已定义目标类别的刻度因变量或分类因变量不可用。请参阅主题第 3 页的『选择类别』，了解更多信息。

行 节点性能表可以按终端节点、百分位数或这两者显示结果。如果选择这两者，则为每个目标类别生成两个表。百分位数表根据排序顺序显示每个百分位数的累计值。

排序顺序

根据因变量的测量级别不同，值也会有所不同，并且增益汇总表与增益表的值互不相同。

百分位数增量

对于百分位数表，可以选择以下百分位数增量：1、2、5、10、20 或 25。

显示累积统计信息

对于终端节点表，在具有累积结果的每个表中显示附加列。

选择统计输出

1. 在"决策树"主对话框中，单击**统计**。

图表

可用的图表取决于因变量的测量级别、生长法和其他设置。

自变量对模型的重要性

按自变量（预测变量）显示的模型重要性的条形图。仅对 CRT 生长法可用。

节点性能

增益 增益是每个节点的目标类别中的总个案百分比，它的计算方法如下：(节点目标 *n* / 总计目标 *n*) x 100。收益图是累积百分位数增益的折线图，它的计算方法如下：(累积百分位数目标 *n* / 总计目标 *n*) x 100。将对每个目标类别生成一个单独的折线图。只对带有已定义目标类别的分类因变量可用。请参阅主题第 3 页的『选择类别』，了解更多信息。

增益图表绘制将在百分位数增益表（它还报告累计值）的增益百分比列中看到的相同值。

指标 指标是目标类别的节点响应百分比与整个样本的总体目标类别响应百分比的比率。 指标图表是累积百分位数指标值的折线图。仅对分类因变量可用。累积百分位数指标的计算方法如下：(累积百分位数响应百分比 / 总响应百分比) x 100。将为每个目标类别生成单独的图表，且必须定义目标类别。

指标图表绘制将在百分位数增益表的指标列中看到的相同值。

响应 节点中的个案在指定目标类别中的百分比。 响应图表是累积百分位数响应的折线图，它的计算方法如下：(累积百分位数目标 *n* / 累积百分位数总计 *n*) x 100。只对带有已定义目标类别的分类因变量可用。

响应图表绘制将在百分位数增益表的响应列中看到的相同值。

平均值

因变量的累积百分位数平均值的折线图。仅对刻度因变量可用。

平均利润

累积平均利润的折线图。只对带有已定义利润的分类因变量可用。请参阅主题第 8 页的『利润』，了解更多信息。

平均利润图表绘制将在百分位数增益摘要表的利润列中看到的相同值。

投资收益 (ROI)

累积 ROI（投资回报）的折线图。ROI 计算为利润与支出之比。只对带有已定义利润的分类因变量可用。

ROI 图表绘制将在百分位数增益摘要表的 *ROI* 列中看到的相同值。

百分位数增量

对于所有的百分位数图表，此设置控制在图表上显示的百分位数增量：1、2、5、10、20 或 25。

选择图表输出

1. 在"决策树"主对话框中，单击**图**。

选择规则和评分规则

使用"规则"对话框，可以生成命令语法、SQL 或简单（纯英文）文本形式的选择或分类/预测规则。您可以在"输出"选项卡中显示这些规则以及/或者将其保存到外部文件。

生成分类规则

选择此项即可启用选择和评分规则设置。

语法 控制选择规则的形式，它将应用于"输出"选项卡中显示的输出，以及保存到外部文件的选择规则。

SPSS Statistics

命令语法语言。规则表示为一组定义过滤条件以用于选择个案子集的命令，或表示为可用于对个案评分的 `COMPUTE` 语句。

SQL 生成标准的 SQL 规则，以便从数据库中选择或提取记录，或者将值指定给那些记录。生成的 SQL 规则不包含任何表名称或其他数据源信息。

简单文本

纯英文的伪代码。规则表示为一组"if...then"逻辑语句，而这些语句描述了模型的分类或每个节点的预测。此形式的规则可以使用已定义变量和值标签或者变量名称和数据值。

类型 对于 IBM® SPSS Statistics 和 SQL 规则，控制生成的规则类型：选择和评分规则

为个案指定值

此规则可用于为满足节点成员条件的个案指定模型的预测值。将为满足节点成员条件的每个节点生成单独的规则。

选择个案

此规则可用于选择满足节点成员条件的个案。对于 IBM SPSS Statistics 和 SQL 规则，将生成单个规则用于选择满足选择条件的所有个案。

将替代变量包括在 SPSS Statistics 和 SQL 规则中

对于 CRT 和 QUEST，可以在规则中包含来自模型的替代预测变量。包含替代变量的规则可能非常复杂。一般来说，如果只想获得有关树的概念信息，请排除替代变量。如果某些个案有不完整的自变量（预测变量）数据并且您需要规则来模拟树，请包含替代变量。请参阅主题第 7 页的『替代变量』，了解更多信息。

节点 控制已生成规则的范围。为范围中包含的每个节点生成单独的规则。

所有终端节点

为每个终端节点生成规则。

最佳终端节点

基于指标值为排在前面的 *n* 个终端节点生成规则。如果该数超过树中的终端节点数，则为所有终端节点生成规则。

最佳终端节点（以指定的个案百分比为限）。

基于指标值为排在前面的 *n* 个个案百分比的终端节点生成规则。

索引值满足或超过分界值的终端节点。

为指标值大于或等于指定值的所有终端节点生成规则。大于 100 的指标值表示，该节点中目标类别的个案百分比超过根节点中的百分比。

所有节点

为所有节点生成规则。

附注：

- 基于指标值的节点选择仅适用于具有已定义的目标类别的分类因变量。如果已指定多个目标类别，则为每个目标类别生成一组单独的规则。
- 实际上，对于用于选择个案的 IBM SPSS Statistics 和 SQL 规则（而不是用于赋值的规则），**所有节点**和**所有终端节点**会生成一条选择分析中使用的所有个案的规则。

将规则导出到文件

在外部文本文件中保存规则。

也可以基于最终树模型中的选定节点，以交互方式生成和保存选择规则或评分规则。

注：如果将命令语法形式的规则应用于另一个数据文件，那么该数据文件中包含的变量必须与最终模型中的自变量同名，以相同的单位测量，并具有相同的用户定义缺失值（如果有）。

指定选择或评分规则

1. 在"决策树"主对话框中，单击**规则**。

通知

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：

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

S

Printed in China