**数据分箱的常用方法**

假设要将某个自变量的观测值分为k个分箱，一些常用的分箱方法有：1.无监督分箱（1）等宽分箱：将变量的取值范围分为k个等宽的区间，每个区间当作一个分箱。

（2）等频分箱：把观测值按照从小到大的顺序排列，根据观测的个数等分为k部分，每部分当作一个分箱，例如，数值最小的1/k比例的观测形成第一个分箱，等等。

（3）基于k均值聚类的分箱：使用第五章将介绍的k均值聚类法将观测值聚为k类，但在聚类过程中需要保证分箱的有序性：第一个分箱中所有观测值都要小于第二个分箱中的观测值，第二个分箱中所有观测值都要小于第三个分箱中的观测值，等等。

2.有监督分箱

在分箱时考虑因变量的取值，使得分箱后达到最小熵（minimumentropy）或最小描述长度（minimumdescriptionlength）。这里仅介绍最小熵。

（1）假设因变量为分类变量，可取值1，…，J。令pl（j）表示第l个分箱内因变量取值为j的观测的比例，l=1，…，k，j=1，…，J；那么 第l个分箱的熵值为Jj=1［-pl（j）×log（pl（j））］。如果第l个分箱内因变量各类别的比例相等，即 pl（1）=…=pl（J）=1/J，那么第l个分箱的熵值达到最大值；如果第l个分箱内因变量只有一种取值，即某个pl（j）等于1而其他类别的比例等 于0，那么第l个分箱的熵值达到最小值。

（2）令rl表示第l个分箱的观测数占所有观测数的比例；那么总熵值为kl= {1rl×Jj=1［-pl（j）×log（pl（j }））］。需要使总熵值达到最小，也就是使分箱能够最大限度地区分因变量的各类别。

**分箱原因：**

**1、算法要求。**某些特定算法（如Naive Bayes、Logistic 回归）要求分类输入。

**2、性能。**如果减少输入字段的不同值数量，算法（如多项Logistic）的性能可能会提高。例如，对每个分级使用中位数或均值，而不使用原始值。

**3、数据隐私。**敏感类个人信息（如工资）可采用范围的报告形式，而不使用实际工资数字，以保护个人隐私。

**分箱技术：**

**1、固定宽度分级（观测值）**

指定用于计算分级“宽度” 的值（整数或实数）。例如，可以使用默认值10 对字段Age 进行分级。由于Age 的范围为18–65，因此生成的分级如下：

[](http://photo.blog.sina.com.cn/showpic.html#blogid=6ce00d7b0100w9e8&url=http://s15.sinaimg.cn/orignal/6ce00d7bga739df671f8e)

分级间隔起点的计算方法为：扫描到的最低值减去分级宽度的一半（指定值）。例如，在上面显示的分级中，使用13 作为间隔的起点，依据的计算方法如下：18 [最低数据值] – 5 [0.5 × (分级宽度10)] = 13。

**2、分位数（相等计数或总和）**

分位数分级方法用于创建集合变量，这些变量可用于将扫描到的记录分割为百分位数（或四分位数、十分位数等）组，使每个组包含相同数量的记录，或使每个组中值的总和相等。记录根据指定的分级字段值按升序排列，因此所选分级变量的值最低的记录将获得等级1，下一组记录等级为2，依此类推。每个分级的阈值将根据所用的数据和分位方法自动生成。

a、四分位数。生成4 个分级，每个包含25% 的观测值；

b、五分位数。生成5 个分级，每个包含20% 的观测值；

c、十分位数。生成10 个分级，每个包含10% 的观测值；

d、二十分位数。生成20 个分级，每个包含5% 的观测值；

e、百分位数。生成100 个分级，每个包含1% 的观测值；

**3、均值和标准差（观测值）**

此方法可根据指定字段分布的均值和标准差的值生成具有划分类别的一个或多个新字段。

A、 +/– 1 标准差。选择此选项将生成三个分级。

B、+/– 2 标准差。选择此选项将生成五个分级。

C、+/– 3 标准差。选择此选项将生成七个分级。

例如，选择+/–1 标准差将产生三个分级，计算方法如下：

[](http://photo.blog.sina.com.cn/showpic.html#blogid=6ce00d7b0100w9e8&url=http://s9.sinaimg.cn/orignal/6ce00d7bga739e0af2708)

在正态分布中，68% 的观测值落入与均值相距不到一个标准差的范围内，95% 落入两个标准差的范围内，99% 落入三个标准差的范围内。

**4、排序观测值**

排序可创建包含数字字段的排序值、分数排序值和百分位数值的新字段。

A、排列顺序：选择升序（将最低值标记为1）或降序（将最高值标记为1）。

B、排序：选择此选项将按上面指定的升序或降序对观测值进行排序。新字段中的值的范围将是1–N，其中N 是原始字段中离散值的数量。结值将获得其排序值的平均值。

C、分数排序值：选择此选项将对观测值进行排序，其中新字段的值等于排序值除以非缺失观测值的权重和。分数排序值介于0–1 之间。

D、百分比分数排序值：每个排序值除以具有有效值的记录数然后乘以100。百分比分数排序值介于1–100 之间。

**5、相对于分类“主管”字段的最优化**

如果要分箱的字段与另一个分类字段强关联，则可选择分类字段作为“主管”字段以便

以类似于保留两个字段间的原始关联强度的方式创建分箱。