数据探索

通过检验数据集的数据质量、绘制图表、计算某些特征量等手段，对样本数据集的结构和规律进行分析的过程就是数据探索。数据探索有助于选择合适的数据预处理和建模方法，甚至完成一些由数据挖掘解决的问题。

数据质量分析

检查原始数据中是否存在脏数据，脏数据一般是指不符合要求，以及不能直接进行相应分析的数据：缺失值、异常值、不一致的值和重复数据及含有特殊符号的数据。

缺失值分析

数据的缺失主要包括记录的缺失和记录中某个字段信息的缺失，两者都会造成分析结果的不准确，以下从缺失值产生的原因及影响等方面展开分析。

（1）缺失值产生的原因

1）有些信息暂时无法获取，或者获取信息的代价太大。

2）有些信息是被遗漏的。可能是因为输入时认为不重要、忘记填写或对数据理解错误等一些人为因素而遗漏，也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障等非人为原因而丢失。

3）属性值不存在。在某些情况下，缺失值并不意味着数据有错误。对一些对象来说某些属性值是不存在的，如一个未婚者的配偶姓名、一个儿童的固定收入等。

（2）缺失值的影响

1）数据挖掘建模将丢失大量的有用信息。

2）数据挖掘模型所表现出的不确定性更加显著，模型中蕴涵的规律更难把握。

3）包含空值的数据会使建模过程陷入混乱，导致不可靠的输出。

（3）缺失值的分析

使用简单的统计分析，可以得到含有缺失值的属性的个数，以及每个属性的未缺失数、缺失数与缺失率等。

处理可以分为删除存在缺失值的记录、对可能值进行插补和不处理。

异常值分析

异常值分析是检验数据是否有录入错误以及含有不合常理的数据。忽视异常值的存在是十分危险的，不加剔除地把异常值包括进数据的计算分析过程中，对结果会产生不良影响；重视异常值的出现，分析其产生的原因，常常成为发现问题进而改进决策的契机。

异常值是指样本中的个别值，其数值明显偏离其余的观测值。异常值也称为离群点，异常值的分析也称为离群点分析。

（1）简单统计量分析

可以先对变量做一个描述性统计，进而查看哪些数据是不合理的。最常用的统计量是最大值和最小值，用来判断这个变量的取值是否超出了合理的范围。如客户年龄的最大值为199岁，则该变量的取值存在异常。

（2）3σ 原则

如果数据服从正态分布，在 3σ 原则下，异常值被定义为一组测定值中与平均值的偏差超过 3 倍标准差的值。在正态分布的假设下，距离平均值 3σ 之外的值出现的概率为 $P(|x-\mu|>3\sigma) \leqslant 0.003$，属于极个别的小概率事件。

如果数据不服从正态分布，也可以用远离平均值的多少倍标准差来描述。

（3）箱型图分析

箱型图提供了识别异常值的一个标准：异常值通常被定义为小于 $Q_L - 1.5IQR$ 或大于 $Q_U + 1.5IQR$ 的值。Q_L 称为下四分位数，表示全部观察值中有四分之一的数据取值比它小；Q_U称为上四分位数，表示全部观察值中有四分之一的数据取值比它大；IQR 称为四分位数间距，是上四分位数Q_U与下四分位数Q_L之差，其间包含了全部观察值的一半。

一致性分析

数据不一致性是指数据的矛盾性、不相容性。直接对不一致的数据进行挖掘，可能会产生与实际相违背的挖掘结果。

不一致数据的产生主要发生再数据集成的过程中，这可能四由于被挖掘数据来自于不同的数据源、对于重复存放的数据未能进行一致性更新造成的。

数据特征分析

分布分析

分布分析能揭示数据的分布特征和分布类型。对于定量数据，欲了解其分布形式是对称的还是非对称的，发现某些特大或特小的可疑值，可通过绘制频率分布表、绘制频率分布直方图、绘制茎叶图进行直观地分析；对于定性分类数据，可用饼图和条形图直观地显示分布情况。

1. 定量数据的分布分析

对于定量变量而言，选择“组数”和“组宽”是做频率分布分析时最主要的问题，一般按照以下步骤进行。

1）求极差。

2）决定组距与组数。

3）决定分点。

4）列出频率分布表。

5）绘制频率分布直方图。

遵循的主要原则如下。

1）各组之间必须是相互排斥的。

2）各组必须将所有的数据包含在内。

3）各组的组宽最好相等。

定性数据的分布分析

对于定性变量，常常根据变量的分类类型来分组，可以采用饼图和条形图来描述定性变量的分布。

饼图的每一个扇形部分代表每一类型的百分比或频数，根据定性变量的类型数目将饼图

分成几个部分，每一部分的大小与每一类型的频数成正比；条形图的高度代表每一类型的百分比或频数，条形图的宽度没有意义。

对比分析

对比分析是指把两个相互联系的指标进行比较，从数量上展示和说明研究对象规模的大小，水平的高低，速度的快慢，以及各种关系是否协调。特别适用于指标间的横纵向比较、时间序列的比较分析。在对比分析中，选择合适的对比标准是十分关键的步骤，只有选择合适，才能做出客观的评价，选择不合适，评价可能得出错误的结论。

对比分析主要有以下两种形式。

（1）绝对数比较

绝对数比较是利用绝对数进行对比，从而寻找差异的一种方法。

（2）相对数比较

相对数比较是由两个有联系的指标对比计算的，用以反映客观现象之间数量联系程度的综合指标，其数值表现为相对数。由于研究目的和对比基础不同，相对数可以分为以下几种。

1）结构相对数：将同一总体内的部分数值与全部数值对比求得比重，用以说明事物的性质、结构或质量。如居民食品支出额占消费支出总额比重、产品合格率等。

2）比例相对数：将同一总体内不同部分的数值进行对比，表明总体内各部分的比例关系。如人口性别比例、投资与消费比例等。

3）比较相对数：将同一时期两个性质相同的指标数值进行对比，说明同类现象在不同空间条件下的数量对比关系。如不同地区商品价格对比，不同行业、不同企业间某项指标对比等。

4）强度相对数：将两个性质不同但有一定联系的总量指标进行对比，用以说明现象的强度、密度和普遍程度。如人均国内生产总值用“元 / 人”表示，人口密度用“人 / 平方公里”表示，也有用百分数或千分数表示的，如人口出生率用‰表示。

5）计划完成程度相对数：是某一时期实际完成数与计划数的对比，用以说明计划完成程度。

6）动态相对数：将同一现象在不同时期的指标数值进行对比，用以说明发展方向和变化的速度。如发展速度、增长速度等。

统计量分析

用统计指标对定量数据进行统计描述，常从集中趋势和离中趋势两个方面进行分析。

平均水平的指标是对个体集中趋势的度量，使用最广泛的是均值和中位数；反映变异程度的指标则是对个体离开平均水平的度量，使用较广泛的是标准差（方差）、四分位间距。

1. 集中趋势度量

（1）均值

均值是所有数据的平均值。

如果求 n 个原始观察数据的平均数，计算公式为：

$$\text{mean}(x) = \bar{x} = \frac{\sum x_i}{n} \tag{3-1}$$

有时，为了反映在均值中不同成分所占的不同重要程度，为数据集中的每一个 x_i 赋予 w_i，这就得到了加权均值的计算公式：

$$\text{mean}(x) = \bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \tag{3-2}$$

类似地，频率分布表（见表 3-4）的平均数可以使用下式计算：

$$\text{mean}(x) = \bar{x} = \Sigma f_i x_i = f_1 x_1 + f_2 x_2 + \cdots f_k x_k \tag{3-3}$$

式中，$x_1, x_2, \cdots, x_k$ 分别为 k 个组段的组中值；$f_1, f_2, \cdots, f_k$ 分别为 k 个组段的频率。这里的 f_i 起了权重的作用。

作为一个统计量，均值的主要问题是对极端值很敏感。如果数据中存在极端值或者数据是偏态分布的，那么均值就不能很好地度量数据的集中趋势。为了消除少数极端值的影响，可以使用截断均值或者中位数来度量数据的集中趋势。截断均值是去掉高、低极端值之后的平均数。

（2）中位数

中位数是将一组观察值按从小到大的顺序排列，位于中间的那个数。即在全部数据中，小于和大于中位数的数据个数相等。

将某一数据集 x:$\{x_1, x_2, \cdots, x_n\}$ 按从小到大排序：$\{x_{(1)}, x_{(2)}, \cdots, x_{(n)}\}$。

当 n 为奇数时

$$M = x_{(\frac{n+1}{2})} \tag{3-4}$$

当 n 为偶数时

$$M = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}) \tag{3-5}$$

（3）众数

众数是指数据集中出现最频繁的值。众数并不经常用来度量定性变量的中心位置，更适用于定性变量。众数不具有唯一性。当然，众数一般用于离散型变量而非连续型变量。

2. 离中趋势度量

（1）极差

极差＝最大值－最小值

极差对数据集的极端值非常敏感，并且忽略了位于最大值与最小值之间的数据的分布情况。

（2）标准差

标准差度量数据偏离均值的程度，计算公式为：

$$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n}} \tag{3-6}$$

（3）变异系数

变异系数度量标准差相对于均值的离中趋势，计算公式为：

$$\text{CV} = \frac{s}{\bar{x}} \times 100\% \tag{3-7}$$

变异系数主要用来比较两个或多个具有不同单位或不同波动幅度的数据集的离中趋势。

（4）四分位数间距

四分位数包括上四分位数和下四分位数。将所有数值由小到大排列并分成四等份，处于第一个分割点位置的数值是下四分位数，处于第二个分割点位置（中间位置）的数值是中位数，处于第三个分割点位置的数值是上四分位数。

四分位数间距，是上四分位数 Q_U 与下四分位数 Q_L 之差，其间包含了全部观察值的一半。其值越大，说明数据的变异程度越大；反之，说明变异程度越小。

周期性分析

周期性分析是探索某个变量是否随着时间变化而呈现出某种周期变化趋势。时间尺度相对较长的周期性趋势有年度周期性趋势、季节性周期趋势，相对较短的有月度周期性趋势、周度周期性趋势，甚至更短的天、小时周期性趋势。

贡献度分析

贡献度分析又称帕累托分析，它的原理是帕累托法则，又称 2-8 定律。同样的投入放在不同的地方会产生不同的效益。

相关性分析

分析连续变量之间线性相关程度的强弱，并用适当的统计指标表示出来的过程称为相关分析。

1. 直接绘制散点图

判断两个变量是否具有线性相关关系的最直观的方法是直接绘制散点图，如图 3-11 所示。

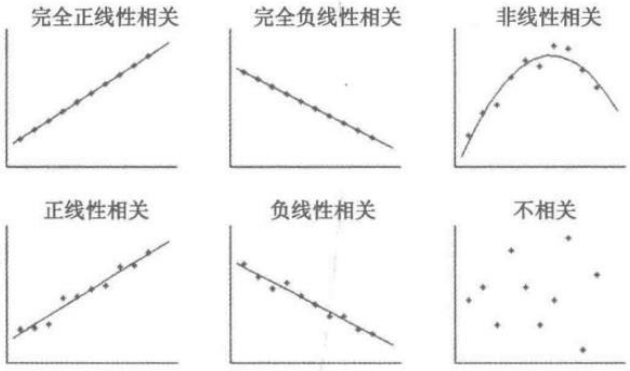

图 3-11　相关关系的图示

2. 绘制散点图矩阵

需要同时考察多个变量间的相关关系时，一一绘制它们间的简单散点图是十分麻烦的。此时可利用散点图矩阵同时绘制各变量间的散点图，从而快速发现多个变量间的主要相关性，这在进行多元线性回归时显得尤为重要。

3. 计算相关系数

为了更加准确地描述变量之间的线性相关程度，可以通过计算相关系数来进行相关分析。在二元变量的相关分析过程中比较常用的有 Pearson 相关系数、Spearman 秩相关系数和判定系数。

（1）Pearson 相关系数

一般用于分析两个连续性变量之间的关系，其计算公式如下。

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3-8}$$

相关系数 r 的取值范围：$-1 \leqslant r \leqslant 1$

$$\begin{cases} r > 0 \text{ 为正相关}, r < 0 \text{ 为负相关} \\ |r| = 0 \text{ 表示不存在线性关系} \\ |r| = 1 \text{ 表示完全线性相关} \end{cases}$$

$0<|r|<1$ 表示存在不同程度线性相关：

$$\begin{cases} |r| \leqslant 0.3 \text{ 为不存在线性相关} \\ 0.3 < |r| \leqslant 0.5 \text{ 为低度线性相关} \\ 0.5 < |r| \leqslant 0.8 \text{ 为显著线性相关} \\ |r| > 0.8 \text{ 为高度线性相关} \end{cases}$$

（2）Spearman 秩相关系数

Pearson 线性相关系数要求连续变量的取值服从正态分布。不服从正态分布的变量、分类或等级变量之间的关联性可采用 Spearman 秩相关系数，也称等级相关系数来描述。

其计算公式如下。

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n(n^2-1)} \tag{3-9}$$

对两个变量成对的取值分别按照从小到大（或者从大到大小）顺序编秩，R_i 代表 x_i 的秩次，Q_i 代表 y_i 的秩次，R_i–Q_i 为 x_i、y_i 的秩次之差。

只要两个变量具有严格单调的函数关系，那么它们就是完全 Spearman 相关的，这与 Pearson 相关不同，Pearson 相关只有在变量具有线性关系时才是完全相关的。

在实际应用计算中，上述两种相关系数都要对其进行假设检验，使用 t 检验方法检验其显著性水平以确定其相关程度。研究表明，在正态分布假定下，Spearman 秩相关系数与 Pearson 相关系数在效率上是等价的，而对于连续测量数据，更适合用 Pearson 相关系数来进行分析。

（3）判定系数

判定系数是相关系数的平方，用 r^2 表示；用来衡量回归方程对 y 的解释程度。判定系数取值范围：$0 \leqslant r^2 \leqslant 1$。$r^2$ 越接近于 1，表明 x 与 y 之间的相关性越强；r^2 越接近于 0，表明两个变量之间几乎没有直线相关关系。

方 法 名	函数功能	所 属 库
sum()	计算数据样本的总和（按列计算）	Pandas
mean()	计算数据样本的算术平均数	Pandas
var()	计算数据样本的方差	Pandas
std()	计算数据样本的标准差	Pandas
corr()	计算数据样本的 Spearman（Pearson）相关系数矩阵	Pandas
cov()	计算数据样本的协方差矩阵	Pandas
skew()	样本值的偏度（三阶矩）	Pandas
kurt()	样本值的峰度（四阶矩）	Pandas
describe()	给出样本的基本描述（基本统计量如均值、标准差等）	Pandas

作图函数名	作图函数功能	所属工具箱
plot()	绘制线性二维图，折线图	Matplotlib/Pandas
pie()	绘制饼型图	Matplotlib/Pandas
hist()	绘制二维条形直方图，可显示数据的分配情形	Matplotlib/Pandas
boxplot()	绘制样本数据的箱形图	Pandas
plot(logy = True)	绘制 y 轴的对数图形	Pandas
plot(yerr = error)	绘制误差条形图	Pandas