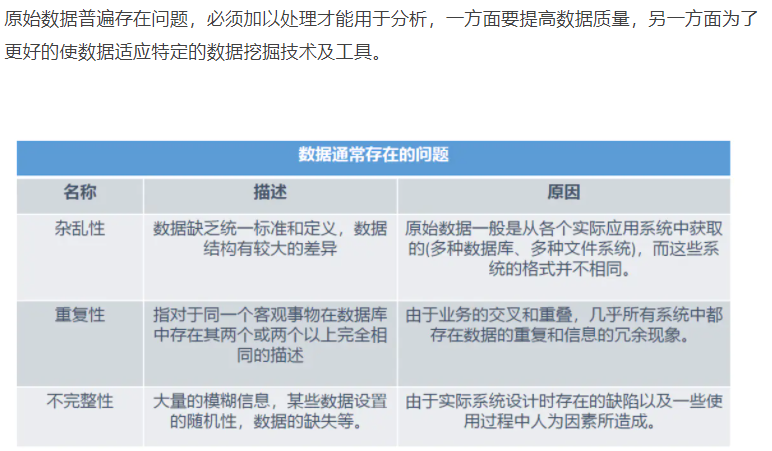
# 原理阶段

1. **为什么要做数据预处理？**



1. **常用的数据预处理包括什么?**

https://zhuanlan.zhihu.com/p/100442371

1) 去除唯一属性

2) 缺失值处理

3) 离群值处理

4) 特征编码

5) 数据标准化

6) 特征选择

1. **缺失值处理包括哪些？**
2. **缺失值分类**
   1. 完全随机缺失：指的是数据的缺失是完全随机的；
   2. 随机缺失：指的是数据的缺失不是完全随机的，和完全变量有关；
   3. 完全不随机缺失：指的是数据的缺失与不完全变量自身的取值相关； 缺失值会使得系统丢失了大量的有用信息，系统所表现出来的不确定性更加显著，系统中蕴含的确定性成分更难把握，包含空值的不完全变量会使得挖掘过程陷入混乱。
3. **缺失值处理的三种方法**

直接使用含有缺失值的特征；删除含有缺失值的特征（该方法在包含缺失值的属性含有大量缺失值而仅仅包含极少量有效值时是有效的）；缺失值补全。

删除含有缺失值的特征：若变量的缺失率较高（大于80%），覆盖率较低，且重要性较低，可以直接将变量删除。

**(1)均值插补**。数据的属性分为定距型和非定距型。如果缺失值是定距型的，就以该属性存在值的平均值来插补缺失的值；如果缺失值是非定距型的，就根据统计学中的众数原理，用该属性的众数(即出现频率最高的值)来补齐缺失的值。

Ps : 定距型数据--数据的中间级，用数字表示个体在某个有序状态中所处的位置，不能做四则运算。例如，“受教育程度”，文盲半文盲=1，小学=2，初中=3，高中=4，大学=5，硕士研究生=6，博士及其以上=7。

**(2)利用同类均值插补**。同均值插补的方法都属于单值插补，不同的是，它用层次聚类模型预测缺失变量的类型，再以该类型的均值插补。假设X=(X1,X2…Xp)为信息完全的变量，Y为存在缺失值的变量，那么首先对X或其子集行聚类，然后按缺失个案所属类来插补不同类的均值。如果在以后统计分析中还需以引入的解释变量和Y做分析，那么这种插补方法将在模型中引入自相关，给分析造成障碍。

**(3)极大似然估计**（Max Likelihood ,ML）。在缺失类型为随机缺失的条件下，假设模型对于完整的样本是正确的，那么通过观测数据的边际分布可以对未知参数进行极大似然估计（Little and Rubin）。这种方法也被称为忽略缺失值的极大似然估计，对于极大似然的参数估计实际中常采用的计算方法是期望值最大化(Expectation Maximization，EM）。该方法比删除个案和单值插补更有吸引力，它一个重要前提：适用于大样本。有效样本的数量足够以保证ML估计值是渐近无偏的并服从正态分布。但是这种方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。

**(4)多重插补**（Multiple Imputation，MI）。多值插补的思想来源于贝叶斯估计，认为待插补的值是随机的，它的值来自于已观测到的值。具体实践上通常是估计出待插补的值，然后再加上不同的噪声，形成多组可选插补值。根据某种选择依据，选取最合适的插补值。

多重插补方法分为三个步骤：①为每个空值产生一套可能的插补值，这些值反映了无响应模型的不确定性；每个值都可以被用来插补数据集中的缺失值，产生若干个完整数据集合。②每个插补数据集合都用针对完整数据集的统计方法进行统计分析。③对来自各个插补数据集的结果，根据评分函数进行选择，产生最终的插补值。

**(5)插值法填充**。包括随机插值，多重差补法，热平台插补，拉格朗日插值，牛顿插值等

**(6)模型填充**。使用回归、贝叶斯、随机森林、决策树等模型对缺失数据进行预测。

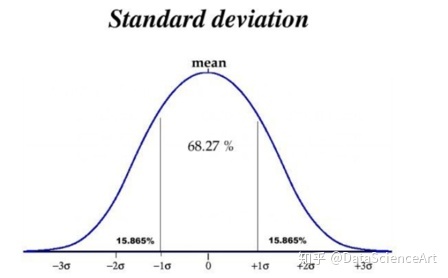
1. **离群值的判断方法vs处理方法？**

因为过大或过小的数据可能会影响到分析结果，尤其是在做回归的时候，我们需要对那些离群值进行处理。实际上离群值和极值是有区别的，因为极值不代表异常，但实际处理中这两个所用方法差不多，所以这里也不强行区分了。

处理方法是调整因子值中的离群值至上下限（Winsorzation处理），其中上下限由离群值判断的标准给出，从而减小离群值的影响力。离群值的判断标准有三种，分别为 MAD、 3σ、百分位法。

**（1）MAD法**。MAD又称为绝对值差中位数法（Median Absolute Deviation）。MAD 是一种先需计算所有因子与平均值之间的距离总和来检测离群值的方法。处理的逻辑：第一步，找出所有因子的中位数 Xmedian；第二步，得到每个因子与中位数的绝对偏差值 Xi−Xmedian；第三步，得到绝对偏差值的中位数 MAD；最后，确定参数 n，从而确定合理的范围为 [Xmedian−nMAD,Xmedian+nMAD]，并针对超出合理范围的因子值做调整

**(2) 3σ法**。又称为标准差法。标准差本身可以体现因子的离散程度，是基于因子的平均值 Xmean而定的。在离群值处理过程中，可通过用 Xmean±nσ来衡量因子与平均值的距离。标准差法处理的逻辑与MAD法类似，首先计算出因子的平均值与标准差，其次确认参数 n，从而确认因子值的合理范围为 [Xmean−nσ,Xmean+nσ]，并对因子值作的调整。



**(3) 百分位法**。计算的逻辑是将因子值进行升序的排序，对排位百分位高于97.5%或排位百分位低于2.5%的因子值，进行类似于 MAD 、 3σ 的方法进行调整。

**(4) 基于距离**。通过定义对象之间的临近性度量，根据距离判断异常对象是否远离其他对象，缺点是计算复杂度较高，不适用于大数据集和存在不同密度区域的数据集。

**（5）基于密度**。离群点的局部密度显著低于大部分近邻点，适用于非均匀的数据集。

**（6）基于聚类**。利用聚类算法，丢弃远离其他簇的小簇。

总结来看，在数据处理阶段将离群点作为影响数据质量的异常点考虑，而不是作为通常所说的异常检测目标点，因而楼主一般采用较为简单直观的方法，结合箱线图和MAD的统计方法判断变量的离群点。

**异常值的处理方法**

（1）根据异常点的数量和影响，考虑是否将该条记录删除，信息损失多

（2）若对数据做了log-scale 对数变换后消除了异常值，则此方法生效，且不损失信息

（3）平均值或中位数替代异常点，简单高效，信息的损失较少

（4）在训练树模型时，树模型对离群点的鲁棒性较高，无信息损失，不影响模型训练效果

1. **特征编码**

**（1）分箱—特征二元化/多元化**

特征二元化的过程是将数值型的属性转换为布尔值的属性，设定一个阈值作为划分属性值为0和1的分隔点。

**（2）独热编码（One-HotEncoding）**

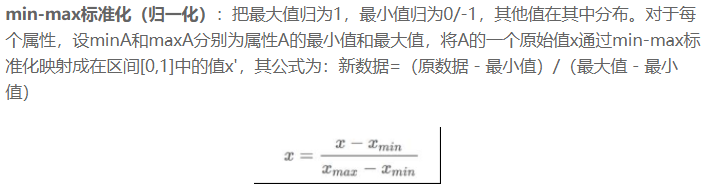
独热编码采用N位状态寄存器来对N个可能的取值进行编码，每个状态都由独立的寄存器来表示，并且在任意时刻只有其中一位有效。

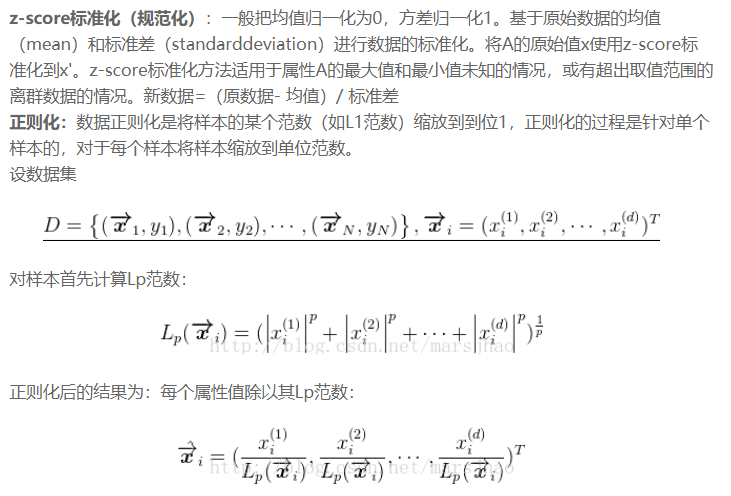
独热编码的优点：能够处理非数值属性；在一定程度上扩充了特征；编码后的属性是稀疏的，存在大量的零元分量。

1. **数据标准化**
2. **数据标准化的原因**

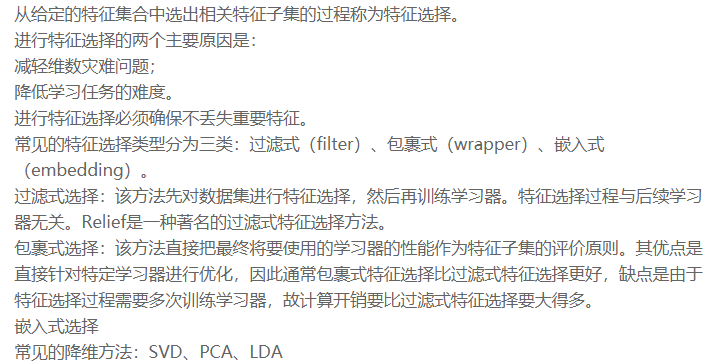
某些算法要求样本具有零均值和单位方差；需要消除样本不同属性具有不同量级时的影响：①数量级的差异将导致量级较大的属性占据主导地位；②数量级的差异将导致迭代收敛速度减慢；③依赖于样本距离的算法对于数据的数量级非常敏感。归一化后求优过程范围变小，寻优过程变得平缓，更容易正确收敛到最优解

1. **数据标准化的方法**

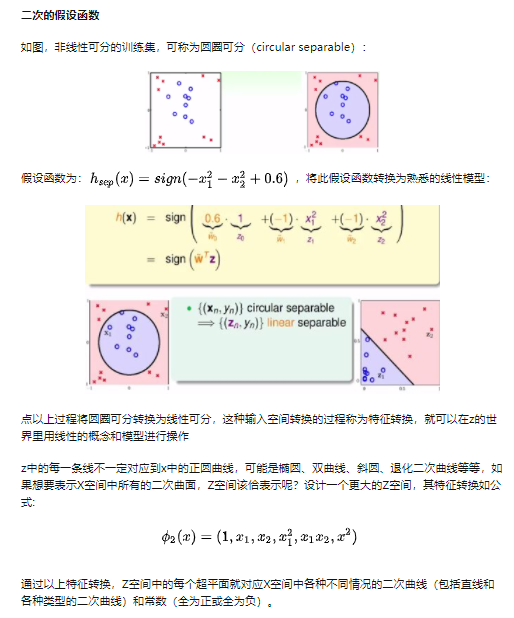




1. **特征选择**



1. **非线性变换**



1. **归一化的作用**

简而言之，是使得没有可比性的数据变得具有可比性，同时又保持相比较的两个数据之间的相对关系，如大小关系；或是为了作图，原来很难在一张图上作出来，归一化后就可以很方便的给出图上的相对位置等。

1. **归一化和标准化的区别**

