3-3-cn
1
本节课我们将讨论数据清洗技术。
2
数据清洗技术可以帮助解决三种问题，重复记录清洗、缺失值清洗和消除噪声数据。首先让我们看一下重复记录清理
3
丢失值的原因可能是
设备异常，或在输入时，有些数据不被重视而不输入。
缺失的值必须被推断和添加。我们可以忽略这个记录;使用默认;使用属性平均;
使用相似样本的平均值，并预测最有可能的值
4
然后让我们看看缺失值数据清洗
5
重复记录会导致错误的分析结果，所以要去除数据集中的重复记录以提高分析的精度和速度。
1 引起重复值的原因
1.1 整合多个数据源的数据
1.2 在输入时，有些数据重复输入
2 重复值经过推断进行合并
2.1 删除完全重复的记录
2.2 合并不同的表时，增加部分冗余属性（例如时间）
6
判断两条记录是否重复：
需要比较两条记录的相关属性，根据每个属性的相似度和属性的权重，加权平均后得到记录的相似度，相似度超过某一阈值，则认为是重复记录。
7
第三是剔除噪声数据
8
噪声数据是一组测量数据中由随机错误或者偏差引起的孤立数据，噪声数据往往使得数据超出了规定的数据域，对后续的数据分析结果噪声不良影响
数据噪声可以通过
分箱/分箱算法
聚类算法
回归算法。
首先让我们看看 Binning 分箱算法
9
噪声数据分箱处理
分箱:
将需要处理的数据按照一定的规则放入一些盒子中，检查每个盒子里的数据，用一定的方法分别处理每个盒子里的数据。
箱子:
一个子区间按照属性值分割。 如果某个属性值在某个子区间范围内，则称将该属性值放入该子区间所代表的“盒子”中。
主要问题：
如何划分盒子
数据平滑法：即如何对每个箱子内的数据进行平滑处理，即用什么方法来决定每个箱子用什么值来代表。

在这个例子中，有 4 个不同的 bin 箱子，灰色、黄绿色和紫色，每个年龄值都放在一定的范围内，比如年龄 10 放在 10 到 16 的范围内，灰色框。 等等

10
噪声数据分箱处理
分箱方法：分箱前根据目标属性值的大小对记录集进行排序， 然后采用
等深分箱
等宽分箱
用户自定义间隔分箱

11
等深分箱法（相同权重）：
箱子根据记录的行数进行划分，每个 箱子中的记录数相同，每个箱子中的记录数称为 bin 的权重，也称为 bin 的深度。
在这个例子中，我们有 16 条记录，我们决定箱子深度为 4 ，分箱后，我们得到如图所示 4 个箱子结果每个箱子包含 4 条记录
12
等宽分箱法，
均匀分布在整个属性值区间上，即每个箱子的属性取值区间范围是一个常数，称为箱子宽度。
例子中，设置区间范围（箱子的宽度）为 1000 元，那么结果也是 4 个箱子，但是箱子内的记录与等深分箱方法不同。
13
分箱后，我们应该选择一个值来表示箱子内的所有值，这称为平滑。
平滑主要有 3 种方式：
Smooth by average：Average the data in the same box value and replace all the data in the box with the average value.
Smooth according to the boundary value：Replace each data in the box with a boundary value with a smaller distance.
Smooth according to the median：Take the median value of the box and use it to replace all the data in the box.
中位数（Median）又称中值，统计学中的专有名词，是按顺序排列的一组数据中居于中间位置的数，代表一个样本、种群或概率分布中的一个数值，其可将数值集合划分为相等的上下两部分。对于有限的数集，可以通过把所有观察值高低排序后找出正中间的一个作为中位数。如果观察值有偶数个，通常取最中间的两个数值的平均数作为中位数。
14 **剔除噪声数据**

① **分箱/分箱算法**
② **聚类算法**
③ 回归算法

我们在看一下聚类算法
15
簇/类：数据对象的集合。 同一个聚类中的所有对象都是相似的，而不同簇中的对象则非常不同的。
聚类：将物理或抽象对象的集合分组到不同的类中，查找并清除那些不在类范围内的值（异常值）。 这些孤立点被视为噪声。
通过聚类分析发现异常数据：
相似或相邻的数据聚合形成簇，这些簇之外的数据对象自然被认为是异常数据。
聚类方法特点：无需任何先验知识，直接形成簇并描述簇。

16

剔除噪声数据

① **分箱/分箱算法**

② **聚类算法**

③ 回归算法

我们最后看一下回归算法

17

回归:

找出图中 x 和 y 这两个相关变量之间的**变化规律**，通过将数据拟合到一个**函数**，即使用拟合函数对数据进行平滑处理。

方法:

线性回归(简单回归):

使用直线建模将一个变量视为另一个变量的线性函数。例如: Y=aX+b，其中 a 和 b 称为回归系数，系数 a 和 b 可以通过最小二乘法得到。

18

本节课我们概要的介绍了数据预处理，今天的课就到这里，谢谢大家