2013 教育部-IBM 产学合作专业综合改革项目
厦门大学《数据挖掘原理及实践》课程习题

第 3 章 数据预处理

1. 数据质量可以用精确性. 完整性和一致性来评估。提出数据质量的其他两种尺度度量方式。

2. 请举出至少 3 种数据降维的算法。

3. 请描述主成分分析方法的基本思路。

4. 假定用于分析的数据包含属性 age。数据元组的 age 值 (以递增排序)为: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70。

(1) 该数据的均值是什么? 中位数是什么?

(2) 粗略的找出数据的第一个四分位数 (Q1)和第三个四分位数 (Q3)。

(3) 给出数据的五数概括。

(4) 画出数据的盒图。

5. 假设医院检测随机选择的 18 个成年人年龄和身体脂肪数据, 得到如下结果:

年龄	23	23	27	27	39	41	47	49	50
脂肪 (%)	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

年龄	52	54	54	56	57	58	58	60	61
脂肪 (%)	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(1) 计算年龄和脂肪百分比的均值. 中位数和标准差。

(2) 绘制年龄和脂肪百分比的盒图。

(3) 根据这两个属性，绘制散布图和 q-q 图。

(4) 根据 z-score 规范化来规范化这两个属性。

(5) 计算相关系数 (皮尔逊系数)。这两个变量是正相关还是负相关。

6. 使用如下两种方法规范化如下数据组:

200, 300, 400, 600, 1000

(1) 令 min = 0, max = 1, min-max 规范化。

(2) Z-score 规范化。

7. 假设 12 个销售价格记录组已经排序如下:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

使用如下每种方法将它们划成三个箱 (bin)。

(1) 等频 (等深) 划分

(2) 等宽划分

(3) 聚类

8. 给定 m 个对象的集合，这些对象划成 K 组，其中第 i 组的大小为 m_i。如果目标是是得到容量为 $n<m$ 的样本，下面两种抽样方案有什么区别? (假定使用有放回抽样)

(1) 从每组随机地选择 $n*m_i / m$ 个元素。

(2) 从数据集中随机地选择 n 个元素，而不管对象属于哪个组。

9. 为什么在数据处理时要进行数据降维?

10. 最大最小规范化和 z 分数规范化的值域分别是什么?