浙江工业大学

数据挖掘实验

2022/2023(2)

计算机科学与技术学院

数据预处理、降维

一、实验目的

- 熟悉基本的数据清洗，包括缺失值处理、异常值识别。
- 掌握基本规范化方法，包括 min-max 规范化和零均值规范化。
- 主成分分析(PCA)降维的实现、可视化。

二、实验内容

1、数据清洗

1.1 用 pandas 读入数据“credit.csv”，查看数据集的摘要。提示：.info()。

1.2 打印出缺失率最高的前 6 个特征以及对应的缺失率；

1.3 对有缺失值的连续属性 Couple_Year_Income，先识别出异常值并删除，然后用不包含异常值的那些观测值的中位数来填充缺失值（参考 ppt 例子）；

1.4 对类别属性'Marriage_State', 'Unit_Kind', 'Title'用众数填充。提示：可以调用 .value_counts()来查看离散特征的每个取值以及对应每个取值的样本数目，或者 .mode()直接得到众数。

1.5 确认已经没有缺失值，并把数据和标签列分离分别得到 X 和 y。

2、数据规范化

2.1 从 sklearn.datasets 导入 wine 数据集，转成 DataFrame，查看输出样本数、属性数，以及每个属性是否有缺失值。

2.2 统计并输出每个属性的均值、标准差。

2.3 对 df 所有属性进行 0-1 标准化，记为 df1。

2.4 对 df 的每个属性进行零均值标准化（均值为 0，标准差为 1），记为 df2。

2.5 计算并输出 df1 和 df2 每个属性的最大、最小值、均值、标准差。

2.6 比对上面结果并讨论 0-1 标准化和零均值标准化的作用及不同之处。

3、PCA 降维及可视化

说明：以下实验可调用 linalg.eig 或 linalg.svd 求特征向量，但不能直接调用其他 PCA 库

函数。

任务 1：降维前后分布对比

3.1 对 df（原始 wine 数据）实现 PCA 降维，降成两维后得到 df_re。

3.2 对 df2(零均值标准化后的 wine 数据)实现 PCA 降维,降成两维得到 df2_re。

3.3 按照实验 2 的方式将 df_re 和 df2_re 分别绘制成散点图。

3.4 观察 df_re 和 df2_re 的散点图，分析并讨论结果。

任务 2：重构误差与维度的关系

3.5 将 df2 降成 k 维，k 分别取 2,4,6,8,10。

3.6 以重构误差为纵坐标，k 为横坐标画折线图。

3.7 观察以上折线图，进行分析和讨论。