暨南大学本科作业专用纸

数据挖掘课程作业

学院 信息科学技术学院 系 计算机科学系 成绩评定

学生姓名 陈宇 学号 2020101642

一. 简述数据预处理的方法有哪些？

答：

1. 数据清洗：去除噪声、缺失值和重复数据等
2. 特征选择/提取：挑选最相关的特征或通过计算构造新的特征
3. 数据归一化：将多个特征的取值范围统一到相同的尺度，使得不同特征对模型的影响权重相同
4. 数据标准化：将数据转化为均值为 0、方差为 1 的标准正态分布
5. 数据集划分：将原数据集划分为训练集、验证集和测试集等
6. 特征降维：通过主成分分析等算法将原始高维数据转化为低维特征向量
7. 数据增强：在原有数据集的基础上，通过旋转、平移、裁剪、翻转等操作生成新的样本，以增加数据量和丰富样本分布。

二. 驾驶员的年龄范围为 18 岁至 70 岁，使用 Min-Max 标准化公式，将 40 岁驾驶员的年龄值变换到[0,1]之间的数值。

答：由 min-max 公式可得：

$$y = \frac{40 - 18}{70 - 18}$$

解之得：

$$y = 0.42$$

三. 简述混淆矩阵，准确率、精准率和召回率。

答：

混淆矩阵：是一些通过模型在测试集上预测结果的统计，包括模型正确分类正样本数量-TP、模型正确分类负样本数量-TN、模型将负样本错误预测为正样本的数量-FP、模型将睁眼本错误预测为负样本数量-FN

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

准确率：所有样本中被正确预测的比率，分类模型总体判断的准确率。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

精确率：预测为正类的准确率。

$$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

召回率：又叫查全率，它是针对原样本而言的，它的含义是在实际为正的样本中被预测为正样本的概率。

$$Recall = \frac{TP}{TP + FN}$$