1 数据挖掘2023 课程作业

1.1 在10,000,000个样本点中聚类，聚为1,000,000类（40分）

1.1.1 数据介绍

1. 数据文件：Task1-聚类数据.zip，其中含有10,000,000个样本，每个样本为一行，第一列为样本号，剩余64列为样本特征。
2. 教师持有该数据的聚类参考标准。

1.1.2 任务介绍

将以上数据聚为1,000,000个类，每类10个样本，并按照样本号顺序输出所聚类别号，类别号取值范围[1,1000000]。

1.1.3 评分标准

1.1.3.1 聚类方法评分，占20分

聚类方案、代码以及ppt、介绍

1.1.3.2 聚类结果评分，占20分

聚类得分：$score_i = \frac{\sum(\text{参考标准类内距离})}{\sum(\text{预测聚类内距离})}$

根据聚类得分，计算聚类排名：$Rank_i$

聚类成绩：$\frac{100-rank_i}{100} \times 20$

1.1.4 输出要求

1. 介绍聚类方案
2. 聚类结果另存一个文件

   输出文件Task1.out说明:格式为文本，将数据每一个样本聚类类别写入该文件，每个样本一行，一行只有一个整数，取值范围[1,1000000]，表示聚类类别。
3. 簇数量最多为1000000，如果超过1000000，将超出部分的簇合并为到第1000000簇

1.2 垃圾邮箱地址检测（30分）

1.2.1 数据说明：

1. email*.txt，共50个文件，每个文件50000000个邮箱地址，这些都是垃圾邮件地址
2. check.txt，一个文件，内有30000000个邮箱地址，这些是待检测的邮件地址

1.2.2 任务介绍

检查check.txt文件中的每一个邮箱地址，如果出现在email*.txt文件中，说明该邮件地址为垃圾邮件地址，输出到文件Task2.out，一个邮件地址占一行。

1.2.3 评分标准

1.2.3.1 检测方案（10分）

1.2.3.2 检测结果评分标准（20分）

精确度（Precision）：

$$\{\text{Precision}\} = \frac{\{\text{True Positives}\}}{\{\text{True Positives}\}+\{\text{False Positives}\}}$$

召回率（Recall）：

$$\{\text{Recall}\} = \frac{\{\text{True Positives}\}}{\{\text{True Positives}\}+\{\text{False Negatives}\}}$$

F1 分数：

$$\{\text{F1}\} = 2 \times \frac{\{\text{Precision}\} \times \{\text{Recall}\}}{\{\text{Precision}\}+\{\text{Recall}\}}$$

根据F1得分，计算检测排名：$Rank_i$

检测成绩：$\frac{100-rank_i}{100} \times 20$

1.2.4 3.3 输出要求

1. 介绍检测方案，代码
2. 输出文件Task2.out说明:格式为文本，输出check.txt文件中的垃圾邮箱地址，并整个方案执行的时间
3. 请事先检查Task2.out的输出数据格式是否正确

1.3 分类应用，给出300类DNA，进行分类预测（30分）

1.3.1 2.1 数据介绍

文件Task3-分类数据.zip提供以下三个数据集：

1. task.3.train.data.csv，训练数据特征文件，一行为一个样本，第一行为特征名称，第一列为训练样本号，剩余16列为特征
2. task.3.train.label.csv，训练数据标签文件，一行为一个样本，第一行为特征名称，第一列为训练样本号，第二列为样本标签即样本分类，其中样本号与task.3.train.data.csv的样本号相同的为同一样本。
3. task.3.test.data.csv，测试数据特征文件，一行为一个样本，第一行为特征名称，第一列为测试样本号，剩余16列为特征。

教师持有该测试数据的分类参考标准。

1.3.2 2.2 任务介绍

根据训练样本数据，进行训练模型，并对测试数据特征文件中每一个测试样本进行预测分类。

1.3.3 2.3 输出要求

1. 介绍分类方案以及代码
2. 输出文件Task3.out说明

格式为文本，将测试数据每一个样本预测的分类写入该文件，每个样本一行，一行只有一个整数，取值范围[1,300]，表示预测的分类。

1.3.4 2.3 评分标准

1. 根据分类结果，计算F1值，本组客观分数=$\frac{\text{本组}F1}{\text{全班最高}F1\text{值}} \times 25$
2. 分类方案、代码以及介绍共5分

1.4 课程作业其他说明

1. 提交时间：2023年6月10日下午4：00之前，由学委收集后，通过邮件将电子版发送给教师，2023年6月11日上课前提交纸质实验报告
2. 不进行分组，每位同学一份课程作业
3. 演示方式：
    1. 上台讲解课程作业
    2. 需要制作PPT
4. 课程作业提交方式：2023年6月11日，提交纸质实验报告
5. 提交材料内容：
    1. 需要提交实验报告、代码以及输出结果，将这些内容打包。一共四个文件
        1. 2024DM_学号_姓名_实验报告.pdf(任务1-任务4实验报告)
        2. 2024DM_学号_姓名_Task1.out
        3. 2024DM_学号_姓名_Task2.out
        4. 2024DM_学号_姓名_Task3.out
    2. 压缩包文件名命名：2024DM_学号_姓名_课程作业.zip。
    3. 输出结果文件必须严格按照本文档的要求。
    4. 实验报告第一页使用给定封面
6. 课程作业评分占总成绩的40%
7. 如有更新，会在QQ群中通知