Big Data Analysis and Mining

大数据分析与挖掘

实验指导书

广东财经大学信息学院

2025 年 10 月

实验要求

一、实验意义和目的

本实验课程配合《大数据分析与挖掘》理论课同步开设，其中包括验证型、设计型和综合型实验。本实验课程着眼于理论与应用的结合，注重培养学生的实际动手能力，增加学生对大数据背景下数据挖掘中的算法和技术的了解。通过实验教学，使学生掌握基于 Python 语言的数据分析和挖掘算法在实际中的应用。

本实验课程要求学生完成适当的上机实习，并写出相应的实验报告。验证和设计题单独完成，综合题任选一题。验证型题目使学生熟悉大数据分析分析挖掘平台的搭建和使用。设计型题目使学生掌握大数据分析与挖掘中算法的基本理论以及在实际中的应用，培养基本的应用能力。综合型题目在于提高学生分析问题、解决问题的能力，加强学生对数据挖掘的整体理解。

二、实验内容安排

实验内容安排如下：

实验项目编号	实验性质	实验要求	实验项目名称	学时	备注
1	验证	必修	数据分析与挖掘平台的搭建	2	
2	验证	必修	Python 编程实践（函数，复用，组合类型等）	2	
3	设计	必修	数据探索与数据预处理	2	
4	设计	必修	分类与回归实践	4	
5	设计	必修	聚类算法实践	2	
6					

三、实验过程要求

实验前要充分做好准备工作：

1．复习和掌握与本实验有关的知识内容；

2．预习、思考实验内容；

3．对实验内容进行分析和设计。

实验过程中，实验者必须服从指导教师和实验室工作人员的安排，遵守纪律与实验制度，爱护设备及卫生。在指定的实验时间内，必须到实验室内做实验。

对于上机过程中出现的问题，尽量先独立思考和解决；对于难以解决的问题可以和同学交流或询问老师；对于同一个实验题目，可以考虑多种方法来实现，然后比较并选择出一种较为有效的方法来实现。

对于设计型和验证型实验，实验时一人一组，独立上机。

四、实验报告要求

实验后，应及时整理出实验报告，实验报告提交电子文档，实验报告具体内容见附录 1：实验报告模板。

实验项目四　电力窃漏电用户自动识别

一、实验类型

实验类型为验证型，2 个学时，必修。

二、实验目的

1、　基于 Python，了解分类算法的使用。

三、基础知识

1、　Python 开发平台的搭建。

2、　Python 使用入门。

3、　Python 数据分析工具。

4、　K 近邻算法和决策树。

四、实验环境

1、　操作系统：Windows XP 以上操作系统，Linux 操作系统、Unix 操作系统或 Mac 操作系统。

五、实验内容

1、　实验内容见教材（《Python 数据分析与挖掘实战》张良均等著（第二版，白色封面的那版））Page 161 的 6.3 上机实验。（第二题中将 LM 神经网络改为 K 近邻算法）。

6.3 上机实验

1. 实验目的

- 掌握拉格朗日插值法进行缺失值处理的方法。
- 掌握 LM 神经网络和 CART 决策树构建分类模型的方法。

2. 实验内容

- 用户的用电数据存在缺失值，数据见“ test/data/missing_data.xls”，利用拉格朗日插值算法补全数据。
- 对所有窃漏电用户及正常用户的电量、告警及线损数据和该用户在当天是否窃漏电的标识，按窃漏电评价指标进行处理并选取其中 291 个样本数据，得到专家样本，数据见“ test/data/model.xls”，分别使用 LM 神经网络和 CART 决策树实现分类预测模型，利用混淆矩阵和 ROC 曲线对模型进行评价。

注意 数据 80% 作为训练样本，剩下的 20% 作为测试样本。

六、实验要求

1、撰写实验报告（模板参照附录 1），要求报告美观，可读性好。内容包括：

（1）实验目的：要求实验目的明确。

（2）实验环境：要求说明实验的软、硬件配置环境。

（3）实验原理：要求能理解实验的基本原理。

（4）实验步骤和结果：要求实验步骤正确、完整；实验结果清晰、明确。

（5）实验总结：包括实验结果评价，实验中遇到的主要问题的分析与处理，要求评价合理，问题描述清楚具体，分析透彻，处理正确。

2、所有实验项目均需提交源代码(复制)及运行结果截图。实验报告交给组长，下下周三晚 10 点前提交。实验报告命名：学号姓名实验报告 3。