《数据科学导论》实验大纲

授课学院	计算机学院（国家示范性软件学院）
课程编号	3132132040
课程类型	专业基础课
授课对象	数据科学与大数据技术
授课学时/学分	32 / 2
授课教师	石川

2021 年 10 月

目录

实验一：宿舍管理程序

1.1 实验目的

掌握基本的 Python 语言使用方法，使学生具有具有基本的 Python 程序设计与开发能力。

1.2 实验内容及要求

【内容描述】使用 Python 语言，设计一个小型的学生宿舍管理程序，系统用户为宿舍管理员。

【功能要求】

（1）学生信息：学号、姓名、性别(男/女）、宿舍房间号、联系电话。

（2）系统功能：

1. 可按学号查找某一位学生的具体信息；
2. 可以录入新的学生信息；
3. 可以显示现有的所有学生信息。

【程序要求】

（1）使用函数 列表 字典 字符串 条件循环等解决问题；

（2）程序规模在 80-200 行左右。

1.3 实验环境

采用 python 语言编程实现。

1.4 评分标准

满分 10 分		
程序 8 分	程序的代码是否简洁无冗余	2 分
	系统功能是否完善和鲁棒	2 分

	对于出错是否能够应对和处理和回退	2 分
	界面是否完善	2 分
文档 2 分	是否对程序代码有清晰明确的解释	1 分
	并且对任务分析得当	1 分

实验二：数据预处理的基本方法

2.1 实验目的

理解数据分析的基本过程，使学生掌握基本的数据预处理方法。

2.2 实验内容及要求

【内容描述】利用给出的房屋价格数据集，进行以下任务：

（1）缺失值的检测与缺失值处理；

（2）异常值检测；

（3）特征间的相关性分析；

（4）对 price 属性进行标准化；

（5）根据 price 属性进行离散化；

（6）找出与 price（房价）相关性最高的三个特征，并给出合理的解释。

【实验要求】建议使用 Jupyter Notebook 完成该任务，给出完成任务的过程。

2.3 实验环境

采用 python 语言并利用 Jupyter Notebook 编程实现。

2.4 评分标准

满分 10 分		
程序 8 分	缺失值的检测是否全面	1 分
	缺失值的处理方式是否更加符合实际问题（例如选取 KNN、树等处理方 式填充缺失值等等）	1 分
	异常值检测是否全面	1 分
	处理异常值的方式是否符合实际问题（例如选取平均值等）	2 分

	能否提取出关键特征删去冗余特征	2
	能否对数据进行合适的标准化	1
文档 2 分	是否对程序代码有清晰的解释，并且描述自己的分析结论	1 分
	具有更加深度和发散的思维，思考数据内部之间存在的隐藏的逻辑关系	1 分

实验三：回归算法实现

3.1 实验目的

理解各种回归方法的理论，使学生能够编写简单的回归算法来进行数据分析 。

3.2 实验内容及要求

【内容描述】使用回归算法对多个数据集进行分析预测。

【实验要求】

（1）不得借助现成的工具包调库，例如 Sklearn；

（2）要求至少可以实现一元线性回归；

（3）有能力可实现多元线性回归；

（4）一元线性回归采用 Boston 数据集的住宅平均房间数作为 x，房屋均价作为 y 作为数据集；

（5）多元线性回归采用 Boston 数据集作为数据集；

（6）逻辑回归采用乳腺癌数据集 barestconcer 作为数据集；

（7）调用 Sklean 库函数，比较自己编写的函数和库函数的好坏。

3.3 实验环境

采用 python 语言编程实现。

3.4 评分标准

满分 10 分		
程序 8 分	复现回归算法是否正确	2 分
	逻辑是否清晰	1 分
	实现多种聚类算法加分	3 分
	与调用库函数的结果比对差距大小	2 分

文档 2 分	是否对程序代码有清晰的解释，并且描述自己的分析结论	1 分
	是否说明聚类的实现过程和详细算法逻辑，代码与公式相对应	1 分

实验四：分类算法实现

4.1 实验目的

理解各种分析方法的理论，使学生能够编写简单的分类算法来实现数据分析。

4.2 实验内容及要求

【内容描述】采用乳腺癌数据集进行分类分析，不使用数据集中的标记数据，只使用属性值。比较几种分类方法的性能。

【实验要求】

（1）采用 Python 实现分类算法；

（2）不得借助现成的工具包调库，例如 Sklearn；

（3）至少实现 K-近邻，朴素贝叶斯，逻辑回归，决策树与支持向量机的其中一个算法；

（4）对乳腺癌数据集调用编写的函数进行分类演示；

（5）调用 Sklearn 的库函数进行对比实验；

（6）能力强的可以多实现几种算法。

4.3 实验环境

采用 python 语言编程实现。

4.4 评分标准

满分 10 分		
程序 8 分	复现分类算法是否正确	2 分
	逻辑是否清晰	1 分
	实现多种聚类算法加分	3 分

	与调用库函数的结果比对差距大小	2 分
文档 2 分	是否对程序代码有清晰的解释，并且描述自己的分析结论	1 分
	是否说明聚类的实现过程和详细算法逻辑，代码与公式相对应	1 分

实验五：聚类算法实现

5.1 实验目的

理解各种分析方法的理论，使学生能够编写简单的数据分析代码。

5.2 实验内容及要求

【内容描述】采用乳腺癌数据集进行聚类分析，不使用数据集中的标记数据，只使用属性值。K-均值聚类与 AGENS 聚类设置簇的个数为 2，对三种聚类结果都进行可视化，对比乳腺癌数据集中的真实标记值，比较几种聚类方法的性能。

【实验要求】

（1）不得借助现成的工具包调库，例如 Sklearn；

（2）手写实现对鸢尾花数据集的聚类，自己选择聚类算法；

（3）能力强的可以实现多种聚类算法；

（4）调用 Sklearn 中的库函数，和自己实现的函数进行对比实验。

5.3 实验环境

采用 python 语言编程实现。

5.4 评分标准

满分 10 分		
程序 8 分	复现聚类算法是否正确	2 分
	逻辑是否清晰	1 分
	实现多种聚类算法加分	3 分
	与调用库函数的结果比对差距大小	2 分

文档 2 分	是否对程序代码有清晰的解释，并且描述自己的分析结论	1 分
	是否说明聚类的实现过程和详细算法逻辑，代码与公式相对应	1 分

实验六：集成算法实现

6.1 实验目的

进一步理解集成算法的原理，能用 sklearn 实现集成算法。

6.2 实验内容及要求

【内容描述】采用 Sklearn 实现集成算法，对乳腺癌数据集进行处理，预测分析结果。

【实验要求】

（1）采用集成模块 sklearn.ensemble 分别实现 AdaBoost 算法，Bagging 算法以及随机森林算法手写实现对鸢尾花数据集的聚类，自己选择聚类算法；

（2）对参数进行调优。

6.3 实验环境

采用 python 语言编程实现。

6.4 评分标准

满分 10 分		
程序 8 分	使用的算法时候恰当	2 分
	是否使用多种算法对比	2 分
	是否使用 Adaboost 集成模型	2 分
	参数调优的方式（例网格调参等）	1 分
	结果误差是否小	1 分
文档 2 分	是否对程序代码有清晰的解释，并且描述自己的分析结论	1 分
	是否说明聚类的实现过程和详细算法逻辑，代码与公式相	1 分

	对应	

实验七：价格预测挑战

7.1 实验目的

进一步熟悉数据处理流程，能够编程实现基本的数据分析预测。

7.2 实验内容及要求

【内容描述】

背景：考虑到网上海量的商品数量，对产品的定价难度很大。比如，服装具有较强的季节性价格趋势，受品牌影响很大，而电子产品则根据产品规格波动。因此，如何根据商品提供的文本信息进行合理定价，有效地帮助商家进行商品的销售是一个有意义的问题。文本分析是指对文本信息的表示及特征项的选取。它从文本中抽取的内容并向量化来表示文本信息，能够反映特定立场、观点、价值和利益。

分析目标：通过给出的商品描述、商品类别和品牌信息，并结合训练数据中的商品价格来给新商品定价格。本案例考虑对文本信息的处理方式。

【实验要求】

（1）尝试关键字组合方法进行分析；

（2）尝试更加复杂的特征工程，比如：无意义符号去除；

（3）尝试使用 MLP、GRU、LSTM 和 TextCNN 等方法对文本信息对文本提取高阶特征；

（4）尝试不同特征信息的组合并结合不同模型分别处理，可以使用集成学习的思想；

（5）上述方法单一使用某一种不能取得最好的效果，请尝试使用多种方法组合。

7.3 实验环境

采用 python 语言编程实现 。

7.4 评分标准

满分 10 分		
程序 8 分	是否对缺失值进行处理和填补（如运用 KNN、树等填充方式）	1 分
	是否对特征值进行处理（对于文本数据处理进行处理和编码、停用词等，运用不同方法提取高阶特征）	3 分
	模型是否选择合适的模型进行训练	3 分
	参数调优的方式（例网格调参等）	1 分
文档 2 分	是否对程序代码有清晰的解释，对数据是否有可视性分析，分析数据内在的隐藏规律	1 分
	对提取的特征是否解释原因，分析最后结构的结论	1 分

实验八：信用卡意向预测问题

8.1 实验目的

进一步熟悉数据分析流程，能够编程实现基本的数据分析预测。

8.2 实验内容及要求

【内容描述】

背景：GAMMA 银行是一家私人银行，经营各种银行产品，如储蓄账户、活期账户、投资产品、信贷产品等。 该行还向现有客户交叉销售产品，为此，客户使用不同的通信方式，如电视广播、电子邮件、网上银行推荐、手机银行等。在这种情况下，GAMMA 客户银行希望将其信用卡交叉销售给现有客户。银行已经确定了一组有资格使用这些信用卡的客户。现在，银行正在寻求您的帮助，以确定可能对推荐的信用卡表现出更高意向的客户。

分析目标：通过银行收集到的客户属性数据，预测客户是否对当前推出的信用卡感兴趣。

【实验要求】

（1）两人一组，一组可三人（31+28），可跨班组队；

（2）采用 AUC 作为评价指标。

8.3 实验环境

采用 python 语言编程实现。

8.4 评分标准

满分 10 分		
程序 8 分	是否对缺失值进行查询和分析	1 分
	对缺失值进行填补（如运用 KNN、树等填充方式）	1 分

	是否对某些文本数据进行编码（例如男女性别编码 0、1 等等）	1 分
	是否使用了多种方法对比特点分析	2 分
	是否将不同算法使用集成学习提高效果	2 分
	调参是否合适（例如使用网格调参等等）	1 分
文档 2 分	是否对程序代码有清晰的解释，对数据是否有可视性分析，分析数据内在的隐藏规律	1 分
	对提取的特征是否解释原因，分析最后结构的结论	1 分