CSA 云行工作室《数据工程综合实践》23 级第 1 期

学习内容与考核安排

报名方式：QQ 群 641172776，联系各组助教报名，报名截止日 2023 年 12 月 18 日 22:00

报名对象：2021 级、2022 级、2023 级计算机类专业本科生

培训教材：R1. 数据工程综合实践实验指导书（见 QQ 群文件）

培训视频资源：http://list.youku.com/albumlist/show/id_51986962.html

培训数据源：阿里天池新人实战赛——O2O 优惠券使用预测

https://tianchi.aliyun.com/getStart/introduction.htm?spm=5176.100066.0.0.518433af02Tyd5&raceId=231593

培训助教团队：

A. 沈宇彤（22 级数据科学与大数据技术拔尖人才实验班）

B. 黄臻炫（22 级数据科学与大数据技术拔尖人才实验班）

C. 陈昊宇（22 级数据科学与大数据技术拔尖人才实验班）

D. 王鹏勋（22 级数据科学与大数据技术拔尖人才实验班）

E. 邹沛汐（22 级数据科学与大数据技术拔尖人才实验班）

F. 朱恒（22 级数据科学与大数据技术拔尖人才实验班）

G. 李昀臻（22 级数据科学与大数据技术）

H. 陈裕萍（23 级硕士研究生）

I. 杨敏（23级硕士研究生）

•每位助教所在组接收学员数为：5-8人

•上述培训教材、示例代码（见QQ群文件）和视频资源（含8个视频）是《数据工程综合实践》学习的核心资源。学员结合上述教材、示例代码、视频自学是本次培训的主要学习方式。

•培训期间考核总成绩在60分以下者，列入CSA云行工作室未完成培训学员名单，在校期间禁止再次报名参加任何CSA云行工作室相关培训和招新活动。

• 完成CSA云行工作室相关培训并考核合格的同学，有义务至少担任一期的培训组长，指导和帮助新学员完成培训任务。

《数据工程综合实践》学习任务固定考核时间（提交实验报告和代码时间）：

启动时间：2023年12月19日

任务一：2024年02月05日

任务二：2024年03月01日

任务一：学习内容及要求

1. Python机器学习库与未集成算法的安装
2. 机器学习基本概念
3. 基础、中级Python 编程
4. 第三方 Python 库

5. 020 优惠券使用预测算法基本框架

6. 020 优惠券使用预测项目解读

7. 数据分析与预处理方法

8. 数据划分与打标方法

9. 020 优惠券使用预测 Baseline

#**机器学习参考书籍：**周志华：机器学习，清华大学出版社，2016

#**机器学习参考视频资源：**吴恩达机器学习系列课程，B 站视频链接：

https://www.bilibili.com/video/BV164411b7dx?p=1

任务一考核内容：

- 自行练习实验指导书：实验任务 1-实验任务 33（不需要交实验任务报告）
- 完成实验指导书：课后作业（二.3）、课后作业（五.2、五.3）、课后作业（八）
- 完成 K-means 算法的代码实现（同时提交源代码）及数据（至少测试 5 个数据集，数据集来源建议采用 UCI 数据集）测试

任务一（考核截止时间：2024 年 02 月 05 日）考核当日须提交：

1.“任务一实验报告”，实验报告书写格式请参考：A1.任务一实验报告模板，**请将此 ms-word 文档命名为“实名-任务一.doc”**

2.“K-means 代码”，**请将此 py 文件命名为“实名-kmeans.py”**（如有多个文件请命名为“实名-kmeans-1.py”、“实名-kmeans-2.py”……）

实验报告和 py 代码请按时发送至各组组长邮箱，由组长统一收集后发送给负责教师。过期者本次任务直接记 0 分

#**本次实验 K-Means 代码将进行查重，**并按代码相似比例（两个百分比取低值，超过 30%开始扣除，相似 30%-10%，相似 35%-15%，相似 40%-20%）给予 K-

Means 部分考核成绩扣除 20%-80%的处罚。

PS:任务一考核成绩占总成绩 50%

任务二：学习内容及要求

1. 020 优惠券使用预测阿里天池平台线上实战

任务二考核内容：

- 阿里天池平台线上测评：提交技术报告（含所有历史提交记录截图、个人天池平台 ID）和源代码

任务二（考核截止时间：2023 年 03 月 01 日）考核当日须提交：

1.“任务二实验报告”，实验报告书写格式请参考：A2.任务二实验报告模板，**请将此 ms-word 文档命名为“实名-任务二.doc”**

2.“O2O 代码”,**请将此 py 文件命名为“实名-O2O.py”**（如有多个文件请命名为“实名-O2O -1.py”、“实名-O2O -2.py”……）

实验报告和 py 代码请按时发送至各组组长邮箱，由组长统一收集后发送给负责教师。过期者本次任务直接记 0 分

#本次实验报告文档、O2O 代码将进行人工审查和查重。

人工审查代码不合格，被认定存在欺诈行为；或者代码无法复现线上成绩的，本次考核成绩记 0 分。

查重过程中出现问题的学生，将根据相似代码行数（超过 100 行开始扣除，100 行-20%，150 行-30%，200 行-40%，250 行 50%，300 行-60%，350 行-70%，400 行-80%，450 行-90%））以及代码相似比例（两个百分比取低值，超过 30%开始扣除，相似 30%-10%，相似 35%-15%，相似 40%-20%）给予本次考核成绩扣除 10%-100%的处罚（相似代码行数、代码相似比例取扣除比例高的一项）。

•特别注意，实验指导书配套的全部任务代码和 Baseline 代码仅供学习和思路参考，照着实验指导书敲一遍代码或者直接复制粘贴代码，查重是必然无法通过的。《数据工程综合实践》的要求是在理解实验指导书的基础上，用自己的思路去完成数据分析任务！

PS:任务二考核成绩占总成绩 50%

阿里天池竞赛平台“020 优惠券使用预测”线上测评评分细则：

考核内容	100-90 分	90-80 分	80-70 分	70-60 分	小于 60 分
阿里天池竞赛平台线上评测 AUC 评分（权重 1.0）	0.800<AUC 评分<=0.805	0.790<AUC 评分<=0.800	0.780<AUC 评分<=0.790	0.760<AUC 评分<=0.780	0.715<AUC 评分<=0.760

答辩考核

实践训练结束将单独组织对总成绩在 80 分以上同学的答辩考核，考核以组为单位，每组在现场随机抽取 1-2 人进行（报名“行业大数据挖掘及应用”微专业且总成绩 80 分以上的同学 100%抽取答辩），每人答辩时间 15 分钟（讲述时间 5 分钟）。答辩考核安排 3-5 位评委判定答辩人完成的项目任务是否真实、合理、有效。如有 2 位（含）以上评委在答辩考核环节判定“不通过”，则总成绩记 0 分，并列入 CSA 云行工作室未完成培训学员名单。

总成绩在 80 分以上同学未按规定时间出席答辩考核，成绩记 0 分。