



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

12

数据科学导论

Introduction to Data Science

Task2: 实验报告

陈恩红，黄振亚

Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2023.html>

助教: 赵官豪，徐碧涵，覃龙虎
ds_intro2023@163.com

9/29/2023



课程要求与考核方式

13

- 课程目标：用科学的方法研究和应用数据
- 课程要求
 - 文献调研报告 1份
 - 每人一份
 - 提交节点：第9周教学周（2023年11月05日）
 - 实践报告 1份
 - 以小组为单位提交，每小组一份，包含每个人的工作介绍（1-3人）
 - 组队节点：第7周上课（2023年10月19日）
 - 提交节点：第15周（2023年12月16日）
- 考核方式
 - 课堂出勤（30%）+调研报告（30%）+实践报告（40%）
 - 结课：第18周（2024年1月4日）



实验（2023.12.16）

14

- 两个实验方式
 - 实验方式1：参加指定问题的数据比赛（组队，推荐）
 - 实验方式2：自己寻找问题和数据，设计方法，进行实验（独自）

- 重点：大家在实践中熟悉和应用数据科学知识，锻炼团队合作能力，只要在报告中叙述清楚、内容合理即可
 - 学习：分析问题、解决问题、代码实践、团队协作、报告撰写
 - 项目组成员、任务分工和组织、个人总结收获

- 课程（国庆后）：编程与实践基础介绍



实验方式1：大数据竞赛

15

- **大数据竞赛：** **组队(1~3人)**参加给定的比赛，最后将做题思路、结果以及比赛排名以报告形式提交
- **报告内容**
 - 比赛名称
 - 队伍名
 - 问题定义
 - 做题思路，模型设计
 - 比赛排名
 - 团队成员分工
 - 个人总结和感悟



实验方式2： 自选问题实验

16

- 自选问题实验：（1人）结合本学期上课内容，并根据拟定问题，和可用数据集，在该数据集上进行实验并对结果进行评价，将所得结果以报告形式提交。
- 报告内容
 - 问题定义
 - 数据集介绍
 - 模型的设计与实现
 - 实验结果评价
 - 个人总结与感悟



组队要求

17

- 可以单挑，可以组队（1-3人），建议组队
- 组队成员
 - 课上同学
- 注明个人分工
 - 未选本课程的同学无需提交个人感悟
- 组队节点：第7周上课前（2023年10月19日）



实验报告评分要求

18

- 问题介绍与理解
- 数据分析：对问题与数据的分析、特征的处理等情况
- 模型：模型的选择是否合适、是否调参、是否尝试并比较多种模型
- 团队协作：个人分工是否明确合理
- 实验过程：认真度、工作量、思路合理性
- 报告条理：是否条理清晰，内容充足
- 是否迟交
- 是否有抄袭
- （源代码）



比赛平台

19

□ 比赛平台-供了解

□ CCF BDCI (主要)

- <https://www.datafountain.cn/special/BDCI>

□ 天池

- <https://tianchi.aliyun.com/competition/gameList/activeList>

□ Kaggle

- <https://www.kaggle.com/competitions>

□ 会议竞赛

- KDD CUP (“大数据世界杯”、数据挖掘领域 “奥运会”)

- NeurIPS 2023 Competition Track

- <https://neurips.cc/Conferences/2023/CompetitionTrack>

- NeurIPS 2023 Machine Unlearning Competition

- Causal Structure Learning from Event Sequences and Prior Knowledge

9/29/2023



比赛实例

20



□ 天池: <https://tianchi.aliyun.com/competition>

□ ACM MM2021: Robust Logo Detection

- 背景: 商品、视频、音乐等领域知识产权问题逐步受到关注, 存在着非法商户通过一些对抗手段干扰商标识别来逃避侵权
- 数据量: 采用目前规模最大的标识检测数据集Open Brand, 来源于各大电商平台, **包含584,920图片, 1,303,563的标注量**
- 中科院、美国东北大学、中国移动、哈工大、浙大等36490支队伍参赛

□ CVPR 2021 AliProducts Challenge: Large-scale Product Recognition

- 背景: 电商企业面临的大规模、细粒度商品图像识别问题
- 数据量: **300万张图片**, 涵盖了**5万个SKU级商品类别**
- 雅虎、字节跳动、南洋理工大学、香港科技大学、清华大学等623支队伍参赛



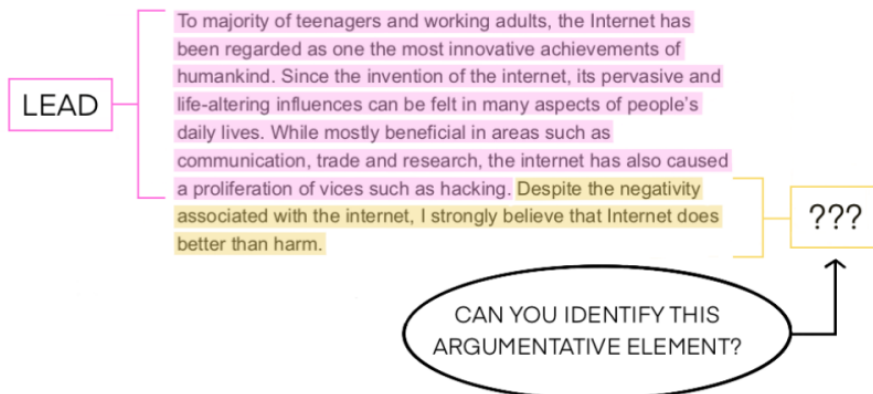
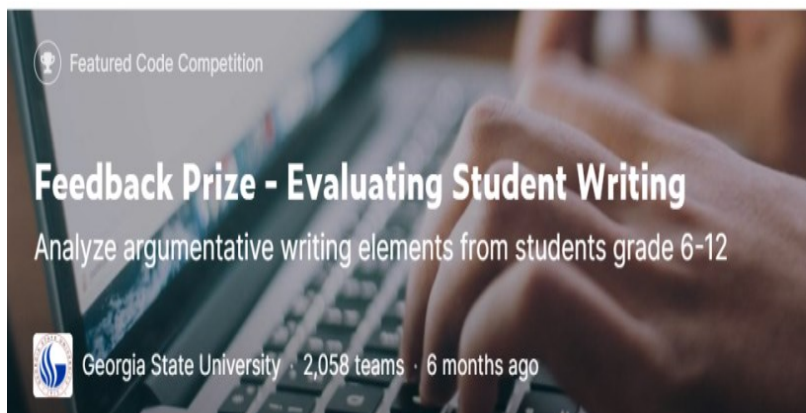
比赛实例

21

□ Kaggle 2022

□ Feedback Prize - Evaluating Student Writing

- 背景：该比赛由佐治亚州立大学 (GSU) 和The Learning Agency Lab提出，专注于开发基于学习的工具和社会公益项目的科学。目前有很多**自动写作**反馈工具，但它们都有局限性，往往不能识别写作结构，比如论文中的导语、立场、论点、论据、反论据等文章元素。
- 要求：将在6 -12年级学生写的文章中找出**学生写作中的元素**，更具体地说，你需要**自动分割文本**，并对议论文的结构元素进行分类。
- 数据量：**15, 600 篇文章**





比赛实例



@ KDD Cup 2021

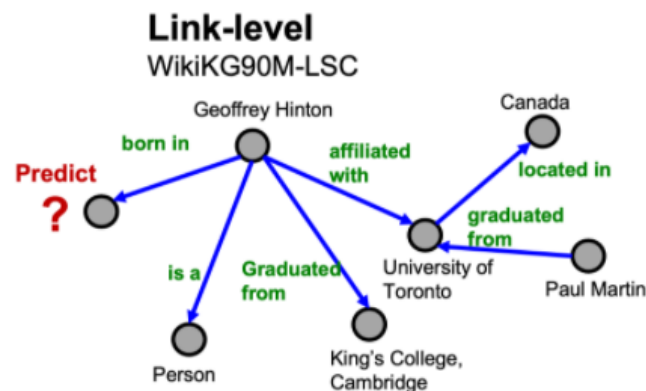
22

□ KDDCup2021 OGB赛道

□ WikiKG90M-LSC

- 背景：大型百科知识图谱(KGs)，如Wikidata和Freebase，通过连接不同实体的三连体表示关于世界的事实知识，例如，(Geoffrey Hinton, citizen of, Canada)。然而，这些大型KGs远远不够完整，缺少实体之间的许多关系信息，因此需要机器学习方法补充
- 数据量：159GB
- DeepMind、微软、UCLA、蚂蚁金服、OPPO、华中科技大学等全球 500 多个顶尖企业、高校和实验室参赛

排名	团队	硬件水平
1	百度	/
2	哈工大	380G memory, Tesla P100 per model
3	中科大	1 TB memory, 8 NVIDIA GEFORCE GTX 2080TI





比赛实例

AAAI2023 Global Knowledge Tracing Challenge

COMPETITION, AAAI • FEB 07, 2023

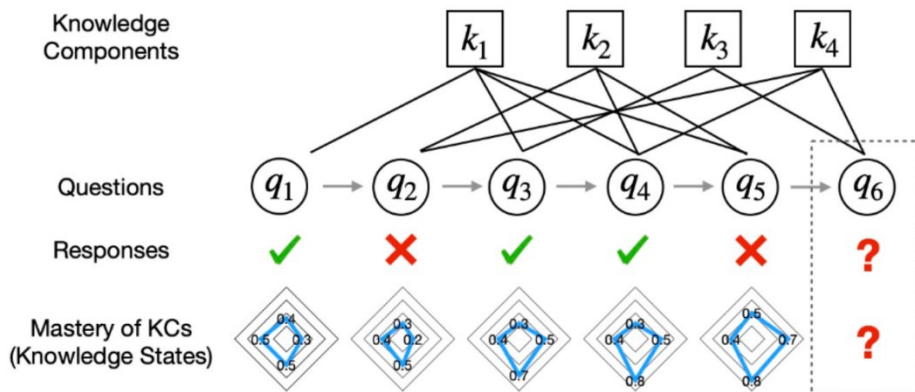
23

AAAI2023 KT比赛

Global Knowledge Tracing Challenge

- 背景：知识追踪（KT）是利用学生的历史学习交互数据来建模他们随时间变化的知识掌握情况，以便预测他们未来的答题表现的任务。这种预测能力可以潜在地为学生提供个性化的服务（比如推荐适合学生能力的试题），这对于构建下一代智能个性化教育至关重要。
- 数据量：18,066名学生在7,652道题目上的5,549,635条答题记录
- 中国科学技术大学、新加坡科技研究局A*STAR、网易等 50多支队伍参赛

排名	团队	学生答题表现预测 (AUC)
1	中科大	0.8178
2	A*STAR	0.8167
3	网易	0.8166





实验题目（推荐赛题）

24

- 课程推荐赛题
- 现提供以下实战题目和若干训练数据集
 - Kaggle比赛 睡眠质量检测-On Hold-\$50,000
 - Kaggle比赛 泰坦尼克幸存者预测
 - CCF-BDCI训练赛 图书推荐系统
 - Kaggle训练赛 预测二氧化碳排放量



实验基本信息

25

□ 数据集：训练集+测试集

2022/08/08 11:01:17	赛题数据	
training_dataset - MD5: e98786b193790857aaa90d90f2b9bfc5		下载
2022/08/08 11:01:01	赛题数据	
test_dataset_A - MD5: 172ae85111abfa5718a7521913be5d5f		下载

□ 常用评价指标

- ✓ 回归任务：RMSE, MAE, NRMSE...
- ✓ 分类任务：ACC, AUC, Recall@K, MRR@K...
- ✓ 主办方自行定义指标：F1、NDCG等

□ A/B榜：评分排名时测试数据分割为A/B两份，分别评分并生成对应排行榜，目的是为了防止对测试数据过拟合

- A榜在“提交开放阶段”对提交结果自动评分并排名,生成A榜
- B榜在“提交截止阶段”对提交结果自动评分并排名,生成B榜，**确定决赛资格**



课程推荐赛题

Kaggle比赛 睡眠质量检测



26

□ Child Mind Institute

- **任务介绍：**这个比赛的目标是检测睡眠开始和醒来。您将开发一个模型，训练在手腕上的加速度计数据，以对一个人的睡眠状态**进行分类**。
- **数据集：**该数据集包含约500个腕带加速度计数据的多天记录，并注释了两种事件类型：**开始（睡眠开始）和醒来（睡眠结束）**。
- **评估方式：**本赛题采用Event Detection AP(<https://www.kaggle.com/code/metric/event-detection-ap/notebook>)进行评价。
- **重要时间节点：**
 - September 5, 2023 - Start Date.
 - November 28, 2023 - Entry Deadline. Team Merger Deadline.
 - December 5, 2023 - Final Submission Deadline.
- **比赛链接：** <https://www.kaggle.com/competitions/child-mind-institute-detect-sleep-states/overview>



课程推荐赛题

Kaggle比赛 睡眠质量检测



27

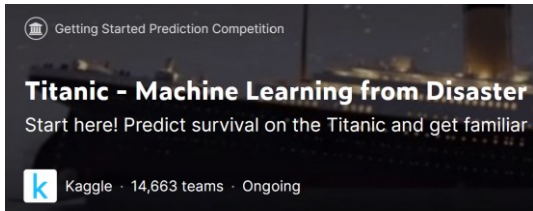
□ 数据示例:

- **train_series.parquet** - Series to be used as training data. Each series is a continuous recording of accelerometer data for a single subject spanning many days.
 - series_id: 序列数据id
 - Step: 整数timestamp
 - Timestamp: ISO 8601 Format Timestamp
 - Anglez: 手臂相对于身体纵轴的角度
 - Enmo:
- **train_events.csv**
 - Series_id: 序列数据id
 - Night: 一个可能行为的枚举
 - Event: 行为类型(onset, wakeup)
 - Step and timestamp:
- 任务分析: 通过train_series.parquet和train_events.csv训练一个模型, 在test_series.parquet上进行预测, 产生提交文件 (详见sample_submission.csv)
 - [Child Mind Institute - Detect Sleep States | Kaggle](#)



课程推荐赛题

Kaggle比赛泰坦尼克幸存者预测



28

□ Kaggle

□ 任务介绍：使用机器学习来创建一个模型，预测哪些乘客在泰坦尼克号沉船中幸存下来。（将乘客分为两类：幸存、死亡）

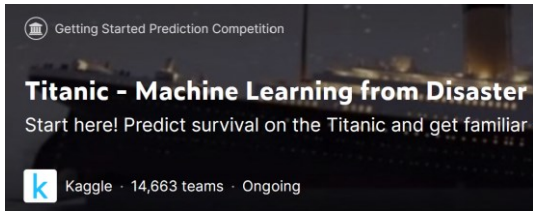
□ 数据集：

变量	定义	键值	输出
survival	Survival	0 = No, 1 = Yes	
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd	输入
sex	Sex		
Age	Age in years		
...	
cabin	Cabin number		
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton	



课程推荐赛题

Kaggle比赛泰坦尼克幸存者预测



29

□ Kaggle

□ Dataset:

- Train.csv、Test.csv
- Submission.csv

□ 链接: <https://www.kaggle.com/competitions/titanic/overview>

提交
文件
示例

PassengerId	Survived
892	1
145	0
896	0
123	1
...	...
800	0



课程推荐赛题

CCF-BDCI训练赛 图书推荐系统

30



- 中国计算机学会
- 任务介绍：本赛题中，为帮助电商系统识别用户需求，为用户提供其更加感兴趣的信息，从而为用户提供更好的服务，需要依据真实的图书阅读数据集，利用机器学习的相关技术，建立一个图书推荐系统。用于**为用户推荐其可能进行阅读的数据**
- 数据集：数据集来自公开数据集Goodbooks-10k，包含网站Goodreads中对10,000本书共约6,000,000条评分。为了预测用户下一个可能的交互对象，数据集已经处理为隐式交互数据集。该数据集广泛的应用于推荐系统中。
- 评估方式：本赛题采用F1进行评价。
- 比赛链接：<https://www.datafountain.cn/competitions/542>



课程推荐赛题

CCF-BDCI训练赛 图书推荐系统



31

- 数据示例：用户-书籍交互矩阵

User_id	Item_id
0	5
0	12352
1	3

- 提交格式要求：

用户ID	用户可能交互的物品ID
User_id	Item_id
39854	5
39864	12352



课程推荐赛题

Kaggle训练赛 预测二氧化碳排放量



32

□ Kaggle

- **任务介绍：**准确监测碳排放能力是应对气候变化的重要步骤。精确的碳排放数据使研究人员和政府能够了解碳排放的来源和模式。尽管欧洲和北美已经建立了广泛的地面碳排放监测系统，但在非洲可用的系统相对较少。**本任务要求参赛选手依据过往二氧化碳排放数据预测未来的排放数据。**
- **数据集：**从卢旺达多个地区挑选了大约497个独特的地点，分布在农田、城市和发电厂周围。这次比赛的数据是按时间划分的;训练数据中包含2019 - 2021年的二氧化碳排放数据，任务是预测2022年至11月的二氧化碳排放数据。
- **评估方式：**本赛题采用RMSE进行评价。
- **比赛链接：** <https://www.kaggle.com/competitions/playground-series-s3e20/overview>



课程推荐赛题

Kaggle训练赛 预测二氧化碳排放量



33

□ 数据示例：

□ 训练数据：train.csv，test.csv

字段	说明	示例
ID_LAT_LON_YEAR_WEEK	地点、时间	ID_-0.510_29.290_2019_00
year	年份	2019
latitude	维度	-0.51
...

输入
78种字段

□ 输出：emission排放量



实验报告（2023.12.16）

34

□ 当前任务

- 第7周，10月19日前完成实验组队和选题，在线表格中填写组队信息和赛题信息
- 注意比赛报名时间
- 填写问卷调查，便于安排编程与比赛介绍

□ 助教：赵官豪，徐碧涵，覃龙虎

- ds_intro2023@163.com
- QQ群：451597440





实验室平台CODIA

35

□ CODIA: <https://code.bdaa.pro>

- 使用人工智能与数据挖掘技术对学习者的编程行为进行多元分析
- 提供智能化、个性化的资源管理、组织、分析服务
- 用户量13000+, 提交量36w+
- 支持学生自学、教师授课、考研复试等





实验室平台CODIA

36

□ CODIA: <https://code.bdaa.pro>

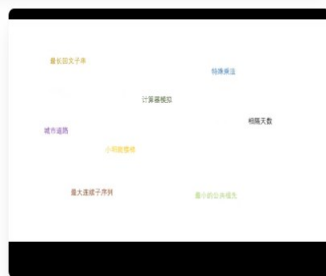
- 2021、2022年春季学期，计算机学院《程序设计II》上机
- 2021年秋季学期，计算机学院《数据结构》上机
- 2021-2023年BDAA实验室考研、夏令营机试



编程到来，智能将来

CODIA 是由大数据分析与应用安徽省重点实验室开发的在线编程平台。掌握编程，是对智能时代的准备，亦是提升竞争力的重要途径。在 CODIA，你可以获得：

- 历年考研及保研机试真题
- 数据分析及算法等课程相关题库
- 中科大校内课程指定在线IDE
- 个性化能力评估和诊断





实验室平台CODIA

37

□ CODIA: <https://code.bdaa.pro>

丰富的题库/题包资源

题目(7188)		题包(105)
状态	题目	
0%	两数之和	
✓	爬楼梯	
	颠倒二进制位	
	位1的个数	
	等差数列划分 II - 子序列	
	环形数组循环	
	神奇字符串	

提供**多语言在线编程环境**、内含丰富的题目/题包资源，方便学生进行编程学习

考研/保研专栏

历年考/保研机试题包 NEW

自由练习

BDAA2024-校内保研机试

限时题包

BDAA2023-保研机试

限时 2 小时

设置考研/保研专栏、助力学生**考研保研复试**

智能题目推荐

为你推荐 ?

[推荐不合适?](#)

【附加选做题】排列金字塔

一般 算法经典思想

[相似习题](#)

拨钟问题

一般 算法经典思想

[相似习题](#)

路径组合

较难 动态规划-基础

[相似习题](#)

根据学生的作答历史，为学生**推荐**相应的题目，为学生提供更高效的学习方式



实验室平台CODIA



38

□ CODIA: <https://code.bdca.pro>

大模型助力的智慧助教

智慧助教

你好，我是CODIA编程智慧助教，你可以询问我任何关于编程的问题

什么是快速排序

A1: 快速排序是一种常用的排序算法，它基于分治的思想。具体来说，快速排序的步骤如下：1. 选择一个基准元素（通常选择数组的第一个元素）。2. 将数组分成两部分，使得左边的元素都小于等于基准元素，右边的元素都大于等于基准元素。3. 对左右两部分分别进行递归地快速排序。4. 合并左右两部分的结果，即得到最终排序后的数组。快速排序的关键在于如何选择基准元素和如何将数组分成两部分。一种常用的方法是使用“挖坑填数”的方式，即先将基准元素挖出形成一个“坑”，然后从数组的右边开始找一个比基准元素小的数填入这个坑，再从数组的左边开始找一个比基准元素大的数填入刚才挖出的坑，重复这个过程直到左右指针相遇，最后将基准元素填入最后一个坑的位置。快速排序的平均时间复杂度为 $O(n \log n)$ ，最坏情况下的时间复杂度为 $O(n^2)$ ，但通常情况下快速排序的效率是很高的。

大模型助力的智慧助教，随时给予学生个性化的专业性解答

智能代码检索

Binary Insertion Sort

搜索

C C++ Python Java

```
#include <stdio.h>
int binarySearch ( int a [ ], int item, int low, int high ) {
    while ( low <= high ) {
        int mid = low + ( high - low ) / 2 ;
        if ( item == a [ mid ] )
            return mid + 1 ;
        else if ( item > a [ mid ] )
            low = mid + 1 ;
        else
            high = mid - 1 ;
    }
    return low ;
}
void insertionSort ( int a [ ], int n ) {
    int i , loc , j , k , selected ;
```

快速查询相关的代码片段。帮助用户更快实现编程目标，避免重复造轮子，提高编程效率

智能诊断分析

能力诊断

诊断不准确?

T = 35



根据学生的作答历史，诊断出学生的能力，为学生提供全面的能力分析



实验室平台CODIA



39

□ CODIA: <https://code.bdca.pro>

自由创建题包

创建题包

自建题包

限时题包 测试作业 限时 20 分钟	创建时间: 2020-09-12 上次修改时间: 2021-07-18 类型: 限时题包 权限: 开放	管理 查看数据
定时题包 测试比赛 限时 2 小时 已截止	创建时间: 2020-09-12 上次修改时间: 2023-03-22 类型: 定时题包 权限: 指定用户可见	管理 查看数据
定时题包 等待比赛 限时 10 分钟 已截止	创建时间: 2020-09-12 上次修改时间: 2021-07-19 类型: 定时题包 权限: 开放	管理 查看数据

教师可自由创建题包，用于课下作业发布、夏令营考研保研机试、随堂测验等情景

题目自动化批改

样例数据 2

输入

[12,3,4,45]
7

输出

[1,2]

解释

取消 保存

通过设置题目测试数据，自动化批改题目并进行智能化打分



实验室平台CODIA

教师用户



40

□ CODIA: <https://code.bdaa.pro>

实时掌握做题情况



题包发布后，实时跟踪用户的答题情况和评分，全方位掌握学生的做题进度和动态

查看用户提交代码

用户: 官方题库
日期: 2022-03-21
耗时: -
是否通过: true

```
#include<bits/stdc++.h>
using namespace std;
int d[500010],L,n,m,l,r;//注意数组不要开小了

bool ok(int x)//x为通过二分算出来的最短跳跃距离的最大值
{
    int tot=0;//tot表示需要移走多少块石头
    int last=0;//last表示上一个石头的位置
    for(int i=1;i<=n;i++)//枚举每个石头
    {
        if(d[i]-last<x) tot++;//d[i]-last表示这个石头到上个没被移走的石头/起点的距离，如果
        else last=d[i];//更新last信息，为下次循环做准备
    }
    return tot<=m;//如果移走石头的个数小于等于m，则x的值是可取的
}

int main()
{
    cin>>L>>n>>m;
```

实时查看用户提交的代码，清晰地了解每个用户的代码编写情况