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/DS2024.html

助教：肖桐
ds_intro2024@163.com

课程要求与考核方式

- 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
    - 独立完成，每人一份
    - 提交节点：第10周教学周（2024年11月05日）
  - 实践报告 1份（初步计划）
    - 组队完成，每小组一份，包含每个人的工作介绍（2人）
    - 组队节点：第7周上课（2024年10月15日）
    - 提交节点：第15周教学周（2024年12月15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12月31日）

实验（2024.12.15）

- 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）**
    - 参加课程推荐的比赛（推荐）
    - 自己寻找正在进行的比赛
  - 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，**报告中叙述清楚、内容合理**
  - 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

实验

- **大数据竞赛**：组队(1~2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

组队要求

- 可以单挑，可以组队（1-2人），建议组队
- 组队成员
  - 课上同学
- 注明个人分工
- **组队节点：第7周上课前（2024年10月15日）**

实验报告评分要求

- 问题介绍与理解
- 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
  - **需提交代码及运行说明**，提交的代码必须可运行，也必须**附有自己运行时的log文件**。如代码或log文件雷同判定为抄袭。
  - log文件要求记录代码运行的中间结果（如数据处理过程和训练时的loss值），要求每一条log都记录相应的时间。示例：

```
2024-09-22 19:10:49 - INFO - Epoch: 1
2024-09-22 19:10:49 - INFO - Learning Rate: 4e-05
2024-09-22 19:10:51 - INFO - Epoch: 1, Step: 0, Train Loss: 0.0102, Precsion: 0.0059, Recall: 0.7903, F1: 0.0117, A
2024-09-22 19:10:59 - INFO - Epoch: 1, Step: 100, Train Loss: 0.4072, Precsion: 0.0101, Recall: 0.7451, F1: 0.0199,
2024-09-22 19:11:07 - INFO - Epoch: 1, Step: 200, Train Loss: 0.1537, Precsion: 0.0215, Recall: 0.6254, F1: 0.0415,
2024-09-22 19:11:14 - INFO - Epoch: 1, Step: 300, Train Loss: 0.0834, Precsion: 0.0357, Recall: 0.4798, F1: 0.0664,
2024-09-22 19:11:22 - INFO - Epoch: 1, Step: 400, Train Loss: 0.0502, Precsion: 0.0539, Recall: 0.3615, F1: 0.0938,
```

实验报告内容要求细则

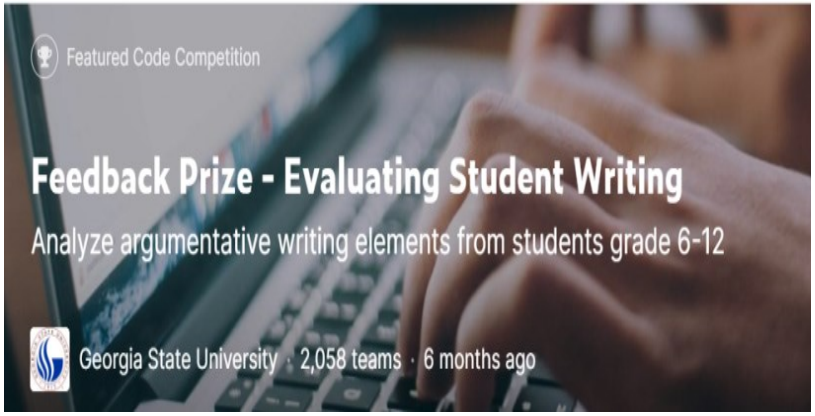
- 数据分析：对问题与数据的分析、特征的处理，需说明将原数据每一行转化为模型输入的完整处理过程。
- 模型：机器学习模型的选择依据以及模型原理说明、模型的输出形式。
- 训练：训练中是否以及如何调参、是否尝试并比较多种模型。
- loss：报告中需要讲述loss函数的含义（即为什么最小化这个函数就可以达到分类、回归预测的目的）。
- 最终模型的预测结果展示与输入输出样例展示。
- 模版不限

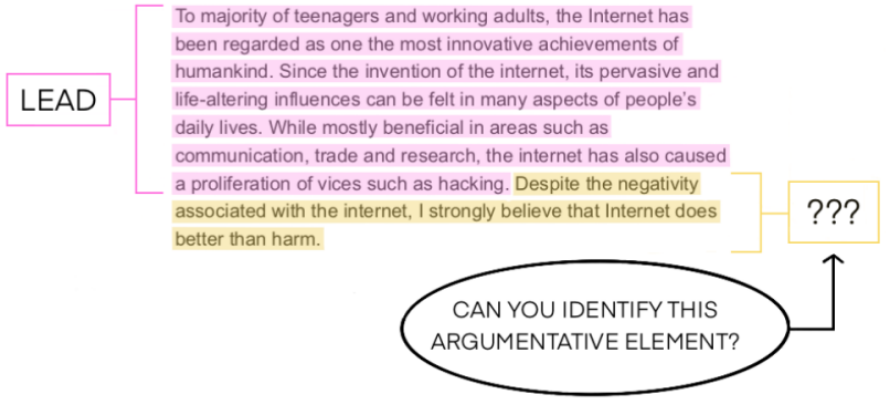
比赛平台

- 比赛平台-供了解
  - CCF BDCI
    - 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 2024 Competition Track
      - https://neurips.cc/Conferences/2024/CompetitionTrack
        - Generative AI and Large Language Models
        - Multiagent Systems and Reinforcement Learning

比赛实例

- 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**

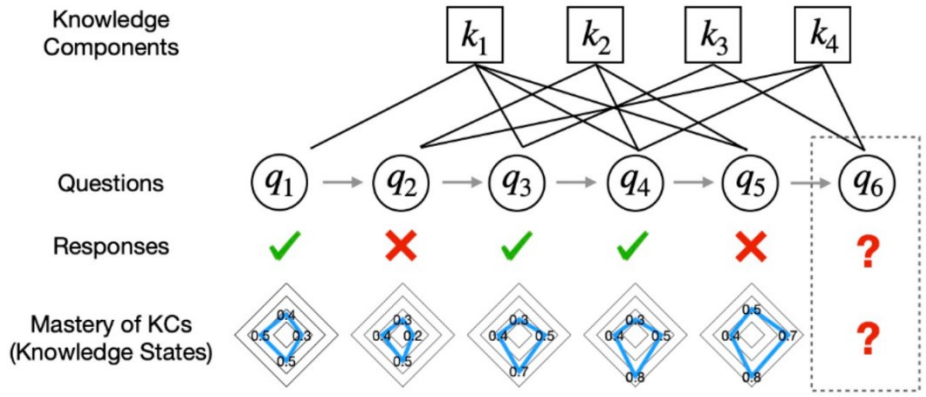




比赛实例

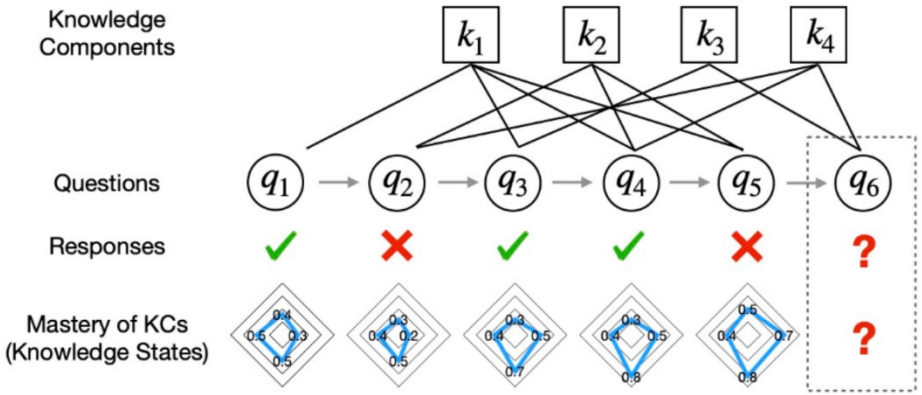
- AAAI2023 KT比赛
  - Global Knowledge Tracing Challenge
    - 背景：知识追踪（KT）是利用学生的历史学习交互数据来建模他们随时间变化的知识掌握情况，以便预测他们未来的答题表现的任务。这种预测能力可以潜在地为学生提供个性化的服务（比如推荐适合学生能力的试题），这对于构建下一代智能个性化教育至关重要。
    - 数据量：18,066名学生在7,652道题目上的5,549,635条答题记录
    - 中国科学技术大学、新加坡科技研究局A*STAR、网易等 50多支队伍参赛

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



比赛实例

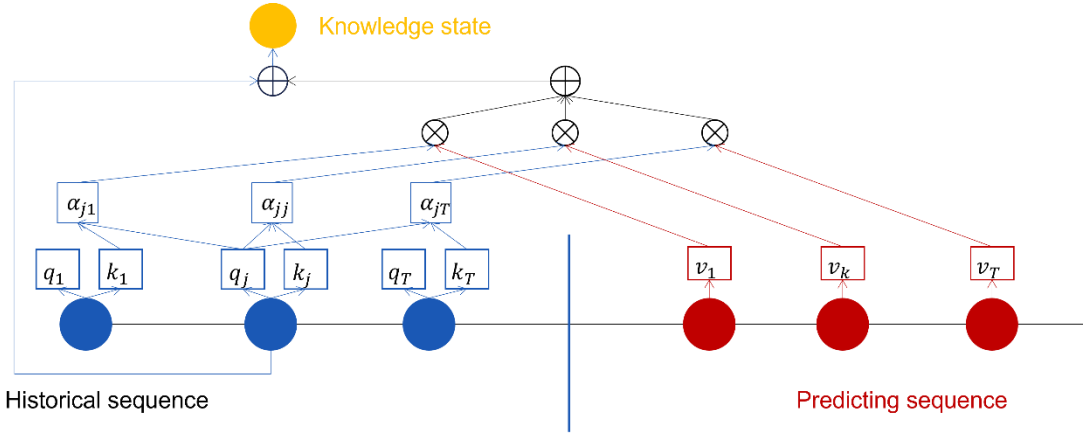
- 解决方案 --- 数据方面
  - 除了最基本的问题-答案对，还使用了以下信息来增强交互表征
    - 问题类别
    - 问题本身内容长度
    - 答案的长度与相对长度
    - 问题相关的知识点，以及该知识点在整个层级知识点树中的位置
    - 历史 8 次互动和未来 8 次互动的间隔



比赛实例

- 解决方案 --- 模型方面
  - Cross-Attention 机制，计算预测问题 e_i 与历史问题之间的注意力权重。将注意力权重应用于学生的历史知识状态，以预测问题 e_i 的表现

$$\alpha'_{ij} = softmax(\boldsymbol{e}_i \cdot \boldsymbol{e}_j)$$



  - 交互位置权重重新分配 --- 对于更近的交互给予更大的权重

$$\beta_j^1 = \frac{j}{\sum_{j'=1}^{T} j'}$$

$$\beta_j^2 = \frac{T - j}{\sum_{j'=1}^{T} (T - j')}$$

$$\alpha''_{ij} = \alpha'_{ij} + \beta_j^1 - \beta_j^2$$

比赛实例

- ICML 2024 数学自动推理比赛
  - Automated Math Reasoning
    - 背景：数学推理是人类智慧中最为高级的形式之一。人类开发了形式化的语言，用以严格描述数学问题并推导数学知识。机器学习领域的研究者们也在努力开发具有数学推理能力的神经模型，旨在达到与人类相似的推理水平。
    - 要求：
      - 数学证明自动形式化：给定一个问题陈述及其自然语言的证明，生成相应的形式化陈述和证明。
      - 数学证明自动去形式化：给定一个问题陈述的形式化陈述和证明，生成相应的自然语言问题陈述及其自然语言的证明。

比赛实例

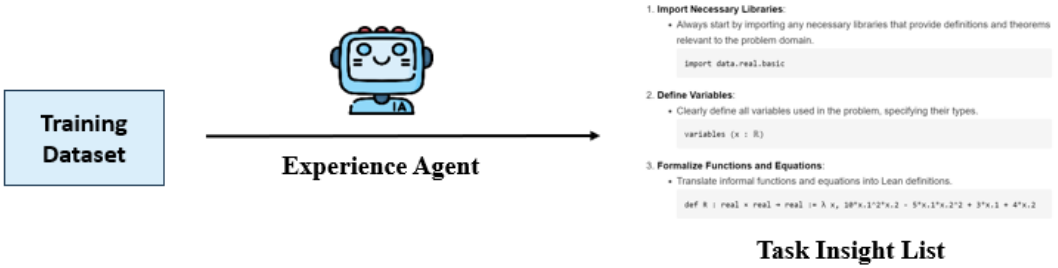
- 数据示例

{
"**problem_name**": "correct_by_msg__ELEM_theorem_proving_1st_grade_15_round2",
"**informal_statement**": "If a two-digit number is formed with 4 in the tens place and 3 in the ones place, prove that the number is 43.",
"**informal_proof**": "We know that the number in the tens place represents tens and the number in the ones place represents ones. So, if 4 is in the tens place, it represents 4 * 10 = 40. If 3 is in the ones place, it represents 3. So the total number is 40 + 3 = 43.",
"**formal_proof**": "def ten (n : ℕ) : ℕ := n * 10\ndef one (n : ℕ) : ℕ := n\ntheorem two_digit_number : ten 4 + one 3 = 43 :=\nbegin\n have h1 : ten 4 = 40, by refl, \n have h2 : one 3 = 3, by refl, \n rw [h1, h2],\n exact add_comm 3 40 \nend"
}

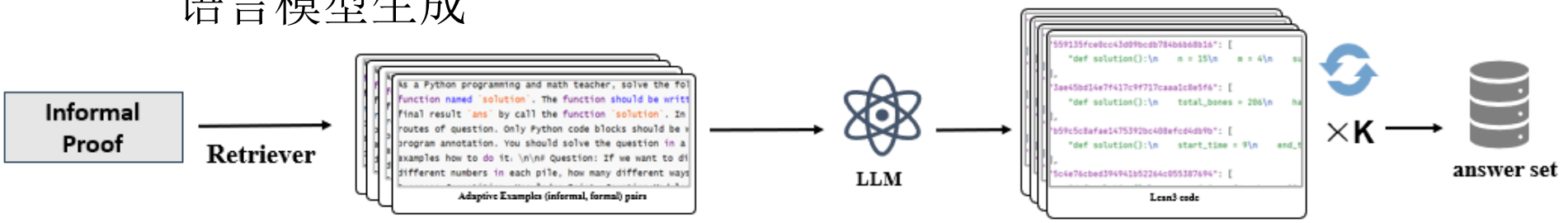
- 结果评估
  - Rouge-L，BLUE，Lean 3 代码通过率
  - Rouge-L，BLUE

比赛实例

- 解决方案
  - 1. 使用大语言模型（GPT-4）总结训练数据集中的经验和方法



  - 2. 在训练集中寻找与当前问题类似的例子，以此为示例辅助大语言模型生成



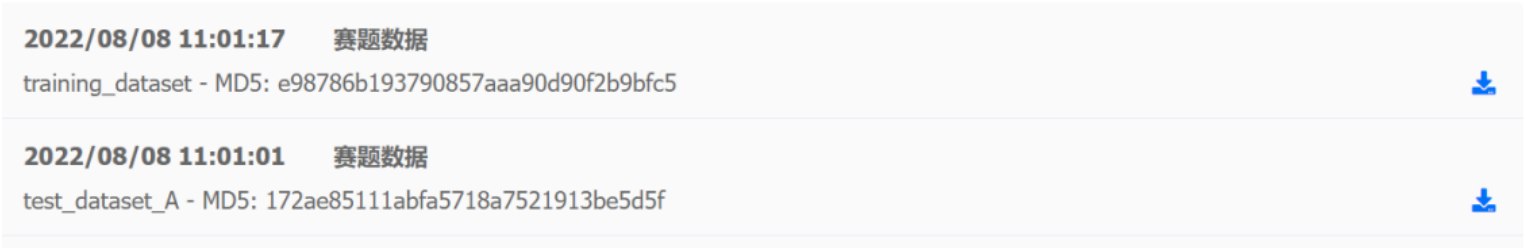
  - 3. 大语言模型生成多个推理路径后，根据验证结果（如检测代码是否通过编译、模型自验证等）和多数投票决定最终答案

实验题目（推荐赛题）

- 现提供以下实战题目和若干训练数据集
  - Kaggle比赛 泰坦尼克星舰幸存者预测
  - Kaggle比赛 青少年互联网使用问题程度预测
  - Kaggle训练赛 洪水发生概率预测

实验基本信息

- 数据集：训练集+测试集



- 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--泰坦尼克星舰转移预测

- Kaggle
- **任务介绍**：使用机器学习来创建一个模型，根据乘客的个人信息预测哪些乘客在泰坦尼克星舰毁灭时被转移到二次元。（**将乘客分为两类：未被传送、被传送**）
- **评价指标**：分类正确率，详见这里
- **数据集**：

变量	定义	键值	
Transported	是否被转移	0 = No, 1 = Yes	输出
CryoSleep	是否冷冻睡眠	True, False	
Cabin	所属舱室	deck/num/side	
Age	年龄	integer	
…	…	…	输入
Name	乘客姓名	String	

课程推荐赛题
Kaggle比赛--泰坦尼克星舰转移预测

- Kaggle
  - Dataset：
    - train.csv、test.csv
    - sample_submission.csv
  - 比赛主页链接：
    - https://www.kaggle.com/competitions/spaceship-titanic

提交文件示例

PassengerId	Transported
0013_01	False
0018_01	False
0019_01	True
0021_01	False
…	…
0023_01	True

课程推荐赛题
Kaggle比赛--青少年互联网过度使用程度预测

- **Kaggle - Child Mind Institute - 正在进行**
- **任务介绍**：这个比赛的目标是检测青少年的互联网过度使用程度。你需要开发一个模型，根据收集的个人信息数据和身体活动数据，预测过度使用互联网的早期迹象。
- **数据集**：分为两个部分：青少年的个人信息数据，以及来自体动记录仪的多天身体活动数据（序列数据）。你的目标就是根据这些数据判断青少年过度使用互联网的程度（如0-正常，...，3-严重）。
- **评估方式**：本赛题采用 Quadratic Weighted Kappa 进行评价（可见比赛主页了解）。
- **重要时间节点**：
  - September 20, 2024 - Start Date.
  - December 20, 2024 - Final Submission Deadline.
- 比赛链接： https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use

- 1st Place - $ 15,000
- 2nd Place - $ 10,000
- 3rd Place - $ 8,000
- 4th Place - $ 7,000
- 5th Place - $ 5,000

课程推荐赛题
Kaggle比赛--青少年互联网使用问题程度预测

- **数据示例：**
  - **series_train.parquet - 用于训练的身体活动序列数据. 每一个 series 都是多天连续的身体加速度记录。**
    - step: id
    - X, Y, Z: 佩戴式手表沿着每个标准轴的加速度测量值
    - light: 环境光强度（以lux为单位）
    - Anglez：手臂相对于身体纵轴的角度
    - ……
  - **train.csv**
    - id：序列数据id
    - Basic_Demos-Age：受试者年龄
    - Physical-BMI: 受试者 BMI 值
    - ……
    - 更多数据域详见比赛主页的 data_dictionary.csv 文件
  - 任务分析：通过series_train.parquet和train.csv训练一个模型，在series_test.parquet和test.csv上进行预测，产生提交文件（详见sample_submission.csv）
    - Child Mind Institute - Problematic Internet Use | Kaggle

课程推荐赛题
Kaggle训练赛--洪水发生概率预测

- **Kaggle**
- **任务介绍：**洪水检测是指识别、监测并向相关部门或个人发出洪水发生或可能发生的警报。该比赛的目的为根据包括地区地形、森林砍伐程度、人工管理水平等多个纬度数据判断当地洪水的发生概率。
- **数据集：**数据集中一共有21个columns作为特征，其中包括地区地形、森林砍伐程度、人工管理水平等数据，预测当地发生洪水的概率。该概率为0-1之间的浮点值，因此本质上是一个回归预测任务。
- **评估方式：**本赛题采用R2指标进行评价。
- **比赛链接：** https://www.kaggle.com/competitions/playground-series-s4e5

课程推荐赛题
Kaggle训练赛--洪水发生概率预测

- **数据示例：**
  - 训练数据：train.csv，test.csv

字段	说明	示例
MonsoonIntensity	季风强度	0-16的数值，表示强度
TopographyDrainage	地形排水	0-18的数值
RiverManagement	人工管理水平	0-16的数值
…	…	…

输入
21种字段

  - 输出：FloodProbability 洪水发生概率

实验报告（2024.12.15）

- 当前任务
  - **第7周，10月15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**
  - **对于正在进行的比赛，注意比赛报名时间**
  - **填写问卷调查，**对大家编程能力进行统计
- 实验提交
  - 实验报告要求提交pdf格式文件
  - 将实验报告、代码和log文件压缩打包，命名为“姓名-学号-课程实验.zip”发送至课程邮箱，邮箱主题为“姓名-学号-课程实验”
  - 课程邮箱：ds_intro2024@163.com