大作业

练习题

- 选题1：蚂蚁&蜜蜂图片分类
- 示例：

蚂蚁

VS

蜜蜂

- 常用模型：
  - 传统模型：人工构造特征+各种各种分类器
  - 神经网络：CNN（可通过残差连接、BatchNorm等方法改变网络结构，探究它们的效果）
- 参考网址：
  - https://work.datafountain.cn/forum?id=86&type=2&source=1

练习题

- 选题2：命名实体识别（序列标注任务）

- 示例：
  - The former Soviet republic was playing in an Asian Cup finals tie for the first time.

- 常用模型：
  - 传统模型：隐马尔科夫模型HMM、条件随机场CRF
  - 神经网络：RNN/LSTM、Transformer、Bert
    - 注：若没有GPU服务器，不建议使用Transformer、Bert

- 参考网址：
  - https://work.datafountain.cn/forum?id=153&type=2&source=1

练习题

- 选题3：垃圾邮件分类

- 示例：

ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv

- 常用模型：
  - 传统模型：决策树、SVM、逻辑回归、朴素贝叶斯、图模型
  - 神经网络：词向量、TextCNN、RNN/LSTM

- 参考网址：
  - https://work.datafountain.cn/forum?id=71&type=2&source=1

练习题

- 选题4：信用卡评级

- 示例：

RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents
0.88551908	43	0	0.177512717	5700	4	0	0	0	0
0.463295269	57	0	0.527236928	9141	15	0	4	0	2

- 常用模型：
  - 传统模型：决策树、SVM、逻辑回归、朴素贝叶斯、图模型
  - 神经网络：全连接网络

- 参考网址：
  - https://work.datafountain.cn/forum?id=73&type=2&source=1

练习题

- 选题5：航空旅客聚类分析（无监督学习）

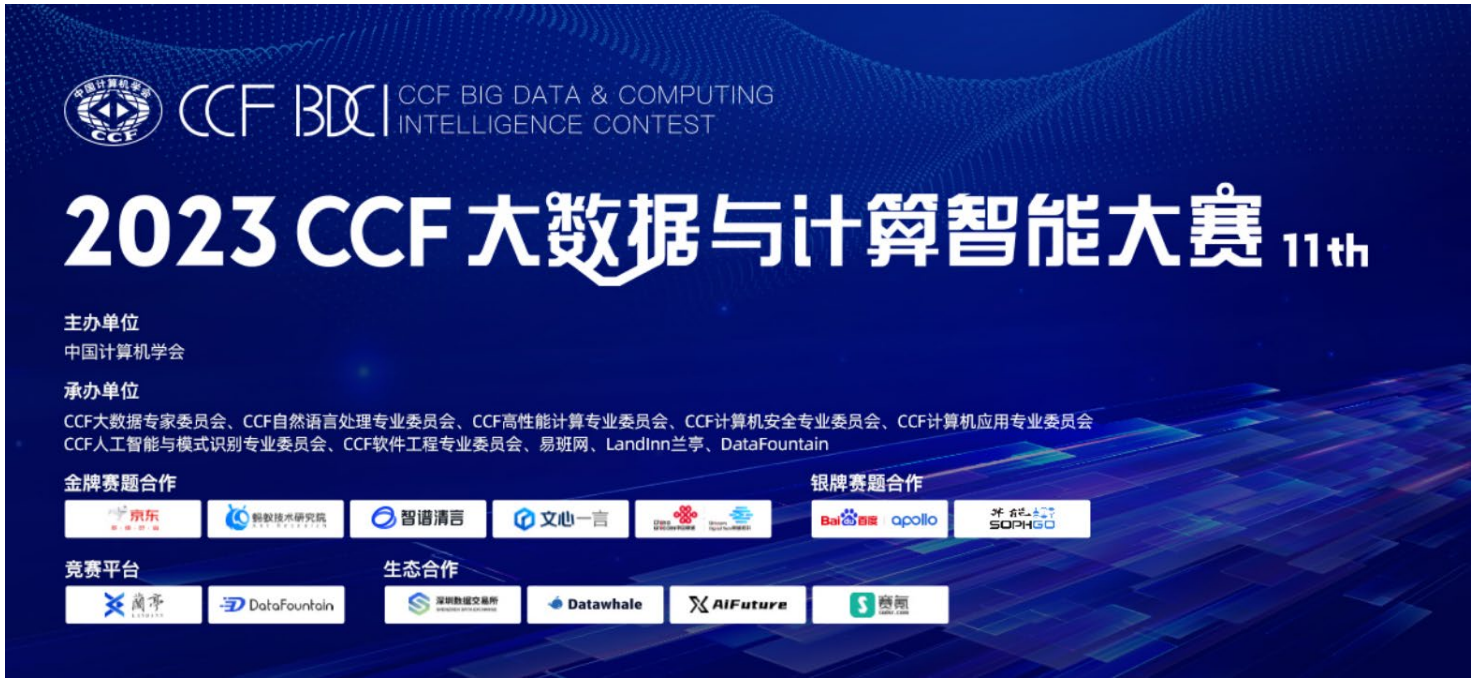
- 示例：

MEMBER_NO	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	AGE	LOAD_TIME	FLIGHT_COUNT	.....
54993	2006/11/02	2008/12/24	男	6	.	北京	CN	31	2014/03/31	210	.....
28065	2007/02/19	2007/08/03	男	6		北京	CN	42	2014/03/31	140	.....
55106	2007/02/01	2007/08/30	男	6	.	北京	CN	40	2014/03/31	135	.....

- 常用模型：
  - K-means、高斯混合模型、层次聚类、密度聚类、谱聚类。。。

- 参考网址：
  - https://work.datafountain.cn/forum?id=67&type=2&source=1

BDCI2023



大赛介绍

CCF大数据与计算智能大赛（CCF Big Data & Computing Intelligence Contest，简称CCF BDCI）由中国计算机学会于2013年创办，是大数据与人工智能领域的算法、应用和系统大型挑战赛事。大赛面向重点行业和应用领域征集需求，以前沿技术与行业应用问题为导向，以促进行业发展及产业升级为目标，以众智、众包的方式，汇聚海内外产学研用多方智慧，为社会发现和培养了大量高质量数据人才。

大赛迄今已成功举办十届，累计吸引全球25个国家，1500余所高校、1800余家企事业单位及80余所科研机构的18万余人参与，已成为中国大数据与人工智能领域影响力最广、参赛规模最大、成熟度最高的综合赛事之一。

2023年，我们将迎来第十一届CCF BDCI，十余年砥砺革新，持续探索数据价值新优势，构筑良性数据生态氛围，CCF BDCI将进一步扩大影响力，关注数字经济技术发展与人才培养，助力推动我国大数据技术及产业生态发展。

BDCI2023赛题

涉及多种数据结构：图像、自然语言、图...
涉及各种行业知识及行业关键问题：金融、医疗、信息安全...
涉及业界最新的模型架构：国产大模型、国产AI框架...

竞技赛

结合业务真实场景应用，选取真实数据开放，包含自然语言处理、数据挖掘、计算机视觉等多种技术领域的竞技赛题。

专题赛

基于承办单位专业领域和技术方向，全年度开展多个独立专题赛发布，赛事组织形式多样、赛程赛制独立定制。

训练赛

中低难度赛题，为高校学生及初学者提供赛练结合促学的平台，并持续发布新的赛题，以满足不同方向、不同阶段的学生及初学者训练学习。

https://www.datafountain.cn/special/BDCI2023/competition

大作业

- **要求：**
  - **选题：**
    - BDCI 2023 中任选一道**竞技赛题**或者**训练赛题**（训练赛题正在陆续公布）
    - BDCI 2023 网站：*https://www.datafountain.cn/special/BDCI2023/competition*
  - **人数：**
    - 每个小组<=5人，选出一位组长，由组长提交电子版大作业
  - **作业提交截止时间：**
    - 本学期结束后的一星期内(具体时间待定)
  - **作业提交邮箱：**
    - 邮箱：ml_ucas_2023@163.com
  - **作业要求：**
    - 每组至少要实现3种方法（注：两种模型+集成方法也算三种方法）
    - 撰写一篇报告
    - 12月份择机选个周末做PPT展示

- **报名方式：**
  - 课程群里的腾讯文档

大作业

- 提交格式：
  - 邮件名称：2023大作业-组号-姓名　　例：2023大作业-3-李四
  - 邮件附件：2023大作业-组号-姓名.zip
    - 将 代码+PPT+报告 打成压缩包
  - 打包目录：
    - 2023大作业-组号-姓名
    - |---　组号-展示.ppt
    - |---　组号-报告.pdf
    - |---　组号-code
    - 　　|--- 代码文件
    - 　　|---　readme.txt
  - 注：代码不要打包数据集、超过100M的模型参数，否则下载容易出问题

大作业

- 额外加分项：
  - 选择CCF BDCI的正式赛题（训练赛道陆续发布中），并在线提交结果
    - 官网：www.datafountain.cn
  - 有现场演示的Demo
  - 用近几年paper中的新模型
  - 自己发明新模型
  - 手动实现模型（注：允许直接调算法库，但不加分）
  - 实现4种以上的模型
  - 在额外数据集上的补充实验
  - 。。。

大作业

- 提示：
  - 1.数据集划分：
    - 若未提供valid集，则需要自己从train中五折交叉划分valid
    - 若valid、test都未提供，则先切出10%的test集，剩余数据按五折交叉切分train、valid
    - 航空旅客聚类是无监督学习，不需要划分train、valid、test
  - 2.模型规模的选择：
    - 数据集都不大，本机无GPU也可训练，有GPU可以加速
    - 若无GPU服务器，不建议使用Bert等大模型
  - 3.可以用模型集成来提高效果
  - 4.非聚类问题，也可以尝试聚类算法
    - 例：正常邮件、垃圾邮件能否天然聚成两簇？

大作业

- 工具推荐：
  - 编程语言： Python（语法简单， package丰富）
  - 算法库：
    - 传统机器学习算法：
      - scikit-learn, scipy
    - 深度学习算法：
      - PyTorch、TensorFlow
      - 注：PyTorch更好上手，TensorFlow分为1.x和2.x两个版本，语法不兼容，网上资料混乱
    - 自然语言处理算法：
      - gensim：词向量
      - Transformers：预训练的Bert
  - 其他：
    - Pandas：处理表格数据
    - jupyter-lab/jupyter-notebook：交互式执行python，便于数据可视化
    - matplotlib：数据可视化
    - tensorboard：可视化监测训练过程中的Loss、Performance曲线
    - tqdm：训练进度条