数据科学导论

Introduction to Data Science

课程实验

（训练题目不在课程考核范围之内）

刘 淇

Email: qiliuql@ustc.edu.cn

课程QQ群：1158258327

实验（2020.12.22）

- 以下方式
  - 1.参加指定问题的数据比赛
  - 在完成1的基础上，有余力的同学，也可以自己寻找问题和数据，自行设计方法，进行实验
- 实验部分重点是为了让大家在实践中熟悉数据科学知识，锻炼团队合作能力，只要在报告中叙述清楚、内容合理即可。
  - 认真度、工作量、思路是否清晰完备、个人收获、实际结果、报告的格式是否专业、能否提交源码
  - 项目组成员、任务分工和组织、个人总结收获
  - 会指定高年级学长（研究生）给大家指导

实验方式1

- 组队(1~3人)参加给定的大数据相关比赛，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

实验方式2

- 结合本学期上课内容，分组(1~3人)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。
- 报告内容包括
  - 问题定义
  - 数据集介绍
  - 模型的设计与实现
  - 实验结果评价

实验报告评分要求

- 对问题与数据的分析、特征的处理等情况
- 模型方面：模型的选择是否合适、是否调参、是否尝试并比较多种模型
- 报告条理：是否条理清晰，内容充足
- 个人分工是否明确合理
- 是否迟交
- 是否有抄袭嫌疑

比赛平台

- 比赛平台-供了解
  - CCF BDCI
    - https://www.datafountain.cn/special/BDCI2020/competition
  - 天池
    - https://tianchi.aliyun.com/competition/gameList.htm?spm=5176.100065.5610717.11.ba5d2bdpinhVA
  - 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 2020 Competition Track
      - https://neurips.cc/Conferences/2020/CompetitionTrack

实验题目

- 现提供以下实战题目和若干训练数据集：
  - **BDCI-20比赛题目：** 企业非法集资风险预测
  - **BDCI-20比赛题目：** 大数据时代的Serverless工作负载预测
  - **BDCI-20比赛题目：** 路况状态时空预测

  - 推荐的训练数据集（不是本课程考核的内容）：
    - UCI数据集：社区犯罪率预测
    - UCI数据集：森林覆盖类型预测
    - UCI数据集：个人收入预测
    - CVPR-17公开数据集：面向图像情感识别

BDCI-20比赛题目：企业非法集资风险预测

- **任务介绍**：非法集资严重干扰了正常的经济、金融秩序，容易引发社会不稳定。本次比赛中，选手需要利用企业数据，构建机器学习模型并预测企业是否存在非法集资风险。
- **数据集**：数据集包含约25000家企业数据，其中约15000家带标签企业作为训练集，剩余数据作为测试集。数据包括了企业基本信息、企业年报、企业纳税情况、企业变更情况、新闻舆情等。
- **评估方式**：本赛题采用AUC进行评价。
- **重要时间节点**：9月29日开始报名，10月13日开放A榜提交，12月5日停止A榜提交，12月6日B榜提交。
- **比赛链接**： https://www.datafountain.cn/competitions/469

BDCI-20比赛题目:大数据时代的Serverless工作负载预测

- **任务介绍**：云计算时代，Serverless软件架构可根据业务工作负载进行弹性资源调整，这种方式可以有效减少资源在空闲期的浪费以及在繁忙期的业务过载。在弹性资源调度的背后，工作负载预测是重要一环。本次比赛中，我们需要预测未来的CPU利用率和作业数。
- **数据集**：本次赛题数据来自华为云数据湖探索。数据每5分钟会进行一次采集，包含集群队列信息，CPU、内存使用量等。
- **评估方式**：要求选手预测未来5个时间点的数据。
  由于需要预测两个目标, 因此采用加权 MSE作为评估指标。
- **重要时间节点**：9月29日开始报名，10月13日开放A榜提交，12月5日停止A榜提交，12月6日B榜提交。
- **比赛链接**： https://www.datafountain.cn/competitions/468

B榜提交只能提交一天，大家要及时提交。

BDCI-20比赛题目：路况状态时空预测

- **任务介绍**：精准预估未来的路况状态对出行决策, 缓解城市拥堵等场景有至关重要的作用。本次比赛中，选手需要根据滴滴提供的路段的实时和历史路况状态特征, 道路基本属性以及路网拓扑关系图, 预测未来一段时间内路段的路况状态 (即畅通, 缓行和拥堵几类状态）。
- **数据集**：本次比赛提供了2019年7月1日至2019年7月31日西安市各路段的实时和历史路况信息, 以及道路属性和路网拓扑信息。
- **评估方式**：由于路况包含畅通, 缓行, 拥堵三种状态, 因此采用加权 F1 Score 作为算法评价指标。
- **重要时间节点**：9月29日开始报名，10月13日开放A榜提交，12月5日停止A榜提交，12月6日B榜提交。
- **比赛链接**： https://www.datafountain.cn/competitions/466

训练：UCI数据集：社区犯罪率预测

- **数据链接：**
  http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime
- **数据简介：**美国境内社区数据，整合了1990年美国人口普查的社会经济数据， 1990年美国LEMAS调查的执法数据和1995年FBI UCR的犯罪数据。

特征类型	实例数量	特征数量	任务类型	缺失值
多变量	1994	128	回归	有

注：特征值已经过归一化

- **任务目标：**预测每10,000人的暴力犯罪数量（对应变量：ViolentCrimesPerPop）

训练： UCI数据集：森林覆盖类型预测

- **数据链接：**
  http://archive.ics.uci.edu/ml/datasets/Covertype
- **数据简介：**数据为来自美国地质调查局（USGS）、美国林务局（USFS）和资源信息系统（RIS）的制图变量（无遥感数据），研究区包括位于科罗拉多州北部罗斯福国家森林的四个荒野地区。

特征类型	实例数量	特征数量	任务类型	缺失值
多变量	581012	54	分类	无

注：数据为原始形式（未缩放），且包含定性自变量（荒野区域和土壤类型）的二进制（0或1）数据列。

- **任务目标：**预测森林覆盖类型（对应变量： Cover_Type）

训练： UCI数据集：个人收入预测

- 数据链接： http://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
- 数据简介：该数据集包含从美国人口普查局进行的1994年和1995年当前人口调查中提取的加权人口普查数据。实例权重（instance weight MARSUPWT）表示由于分层抽样，每个记录所代表的人口中的人数。

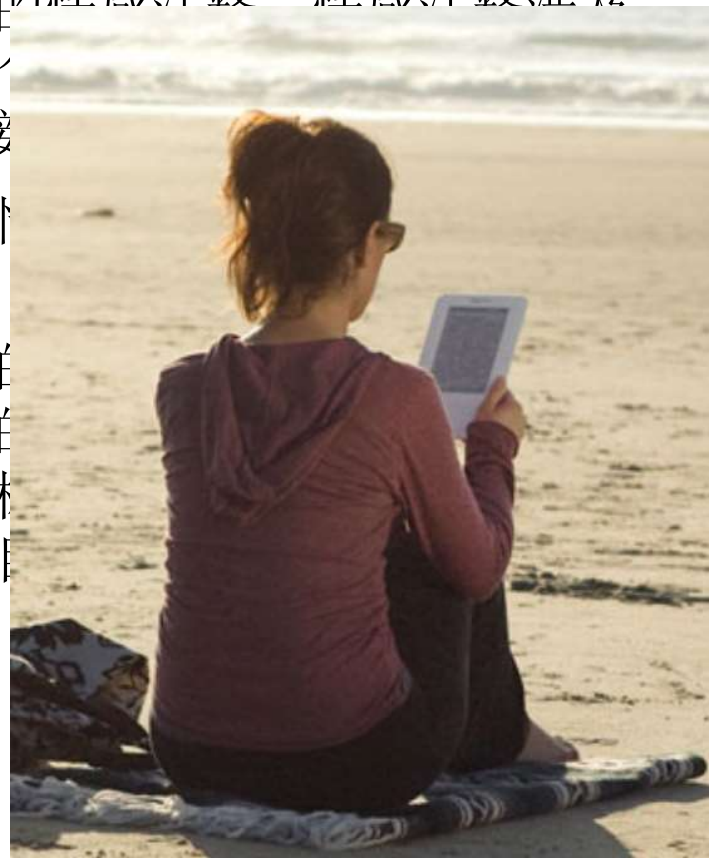
特征类型	实例数量	特征数量	任务类型	缺失值
多变量	299285	40	分类	有

注：此数据集已分好训练集和测试集

- 目标：预测个人总收入（对应变量： total person income PTOTVAL）

训练： CVPR-17公开数据集：面向图像情感识别任务的EMOTIC Dataset

- **数据链接：**http://sunai.uoc.edu/emotic/?tdsourcetag=s_pctim_aiomsg
- **数据简介：**在CVPR17文章"Emotion Recognition in Context"中公开，包括人在不同真实环境下的数字图像和对应的情感注释，情感注释涉及26种情感类别，以及与情感深度相关联的三个Arousal和Dominance（具体含义请参考链接
- **注意：**该数据集有别于面部表情数据集，其中心、难过、愤怒等。
- **任务目标：**根据一个人所处的环境和正在做的其在当前环境下的情感类型。例如某个图像的阅读，那我们可以给与其平静、专注的情感机器识别出不同环境下人类情感的能力，其目表情、肢体动作的图像情感识别任务的。



训练： QM9数据集：分子属性预测

- **数据链接：** https://github.com/geekinglcq/QM9nano4USTC
- **数据简介：** 该数据集包括了13万有机分子的构成,空间信息及其对应的属性. 它被广泛应用于各类数据驱动的分子属性预测方法的实验和对比.
- 除了原始数据外,我们还给出了一些有效的预处理/特征工程方案,如CM,HOB,BAML等.

特征类型	实例数量	特征数量	任务类型	缺失值
多变量	133，885	/	回归	有

- **目标：** 预测分子能量（对应变量：U0）

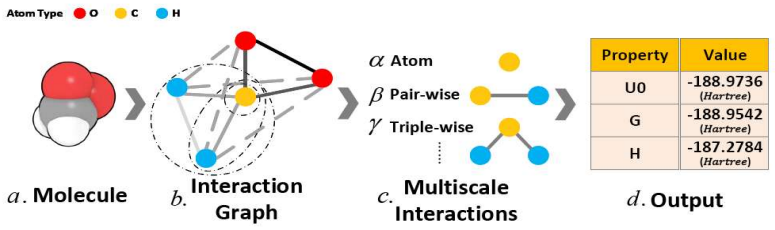


Figure 1: Illustration of the process of a molecule (CH_2O_2) via our method.

训练： QM9数据集：分子属性预测

腾讯量子实验室发起Alchemy竞赛聚焦机器学习预测分子性质

文章来源：企鹅号 - 机器之心

alchemy.tencent.com

腾讯量子实验室公开自研的分子量子性质数据集，发起Tencent Alchemy 2019竞赛，关注算法的泛

Alchemy Contest

What is the Alchemy Dataset?

The Tencent Quantum Lab has recently introduced a new molecular dataset,

现在任务：

10月25日前完成实验组队和选题，并把相关信息发给助教

顾垠 gy128@mail.ustc.edu.cn

课程**QQ**群： 1158258327

注：也可以与未选修该课程的同学组队