《机器学习导论》竞赛作业说明

1 参赛说明

1. 本次竞赛作业仅限单人完成, 不允许组队.
2. 三场比赛可供选择, 可以全部参加, 但仅能选择一场比赛的成绩作为评分依据.
3. 在相应比赛网站提交预测结果时需提供队伍名, 队名中**不能出现学号、学校、院系等太过明显的标识, 命名由字母与数字组成, 不限大小写**. 在比赛截止前会通过课程QQ群统计队伍名, 统计过后不能更改.

2 赛题内容

共有三场比赛可供选择, 每位同学选择一场参加即可. 请注意赛题二在5月底截止, 需要参加的同学请尽快准备.

2.1 赛题一: The ARIEL Space Mission (推荐)

比赛地址为 https://www.ariel-datachallenge.space/ML/documentation/description. ARIEL Space Mission 是欧洲航天局支持的一个科研任务, 其目的是研究太阳系附近的1000颗系外行星的大气层及其化学成分.

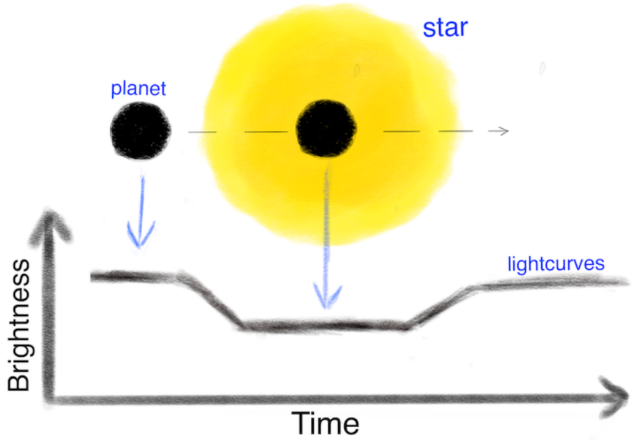

图 1: 当行星经过恒星时, 恒星发出的光被行星遮挡、吸收, 使得观测到的光强度下降; 当行星通过恒星后, 观测到的光强度恢复正常.

如图1所示, 在行星经过恒星时, 可以得到光强度变化曲线. 通常来讲, 半径越大的行星对光的阻挡越严重, 导致光强度变化更大, 所以可以基于该曲线预测行星半径. 每个波长的光都对应一条这样的光强度曲线, 所以可以对行星在每个波长下的观测半径进行预测, 进而分析行星大气层中的化学物质对不同波长的光的吸收情况, 这样就可以得到大气层化学成分的相关信息. 在实际中, 观测数据会受到多种噪声的影响, 如恒星耀斑、仪器噪声等, 这对准确分析行星大气层增加了困难.

该比赛的目标是使用光强度曲线数据、恒星与行星参数预测行星在给定波长下的相对观测半径, 是一个多目标回归任务. 输入输出格式、评价指标等详见官网说明, 该比赛每24小时可提交1次.

预测结果提交截止时间为 **2021.6.30 23:55**, 实验报告中的分数及排名应为这一时刻后的官网结果.

提示:

1. 该比赛数据量较大(完整数据约22G), 可以仅使用部分数据搭建模型, 并在实验报告中注明使用的数据量以便于评分. 经过初步测试, 使用 1% 的数据训练模型得到的模型与使用全部数据训练的模型的性能差距不大.
2. 基于DF21的示例在ecml.ipynb文件中, 经过简单比较, DF21 的性能优于随机森林;
3. 基于神经网络的官方示例https://github.com/ucl-exoplanets/ML-challenge-baseline.

2.2 赛题二: Tabular Playground Series - May 2021

比赛地址为https://www.kaggle.com/c/tabular-playground-series-may-2021. 这是由 Kaggle 提供的练手竞赛, 每月举办一次, 时长一个月, 难度相对较低. 该竞赛的目标是基于输入特征预测商品类别, 是一个多分类任务. 输入输出格式、评价指标等详见官网说明, 该比赛每24小时可提交5次.

预测结果提交截止时间为 **2021.5.31 23:55**, 实验报告中的分数及排名应为比赛结束后的官网private leaderboard结果. 该比赛结束较早, 希望参加的同学请**尽快提交**.

示例代码在kaggle.ipynb文件中.

2.3 赛题三: Tabular Playground Series - June 2021

由Kaggle提供的练手竞赛, 比赛地址待定.

预测结果提交截止时间为 **2021.6.30 23:55**, 实验报告中的分数及排名应为比赛结束后的官网 private leaderboard 结果.

3 赛题引导

经过一个学期的学习, 想必同学们对机器学习的基础知识、一些基本模型已经了如指掌. 但是, 对于比赛而言, 数据处理、特征工程与模型选择、模型训练同样重要. 而且在大多数情况下, 针对某些特征明显的问题, 所有参赛者使用的模型可能都是类似的，最终成绩在很大程度上取决于特征工程. 下面是一些推荐阅读材料, 供同学们参考.

1. Kaggle比赛入门介绍: 1 2
2. Kaggle上的各类问题的Notebook教程
3. 数据科学家初学者教程
4. 书本外的常用模型: 如XGBoost, ARIMA等;
5. Ensembling & Stacking models
6. Kaggle竞赛的获胜解答

4 提交内容

1. 实验报告. 命名格式为“学号_姓名_赛题编号.pdf”, 如“191220000_张三_1.pdf”. 内容应包括: 队伍得分与排名(提供数据和官网截图, 赛题二和赛题三应为 private leaderboard 结果)、建模思路与方法、测试结果等.
2. 代码. 请将项目代码**上传至GitHub**, 并在实验报告中给出项目链接. 要求所有代码放在名为 code 的文件夹下, 并且文件夹下包含一个 main.py 作为入口文件来实现你的模型训练和测试过程, 以及一个 requirements.txt 来指定你所用到的第三方模块, 如这个例子, 使我们可以用如下命令运行你的代码:

   ```
   pip install -r requirements.txt
   python main.py
   ```

3. 提交方式. 将code文件夹与实验报告一起添加到压缩文件“学号_姓名_赛题编号.zip”中, 将该压缩文件上传至教学立方.
4. 作业提交截止时间: **2021.7.1 10:00**, 请注意时间, 在相应比赛预测结果提交截止后尽快提交作业.