大数据导论大作业

大作业：特定疾病的回归和分类

叶允明

计算机科学与技术学院

哈尔滨工业大学（深圳）

大数据导论大作业

- 一、题目
- 二、数据
- 三、评估指标
- 四、任务要求与说明

一、题目

- 本实验旨在通过某种患病病人的临床数据和体检指标来预测人群指示病情程度的指标。
- 需要设计高效，且解释性强的算法来精准预测病情指标。
- 全部编程实现

二、数据—任务I

- 实验任务I数据为训练集文件d_train.csv，测试集d_test.csv
- 每个文件第一行是字段名，之后每一行代表一个个体。
- 训练集文件共包含42个字段，包含数值型、字符型、日期型等众多数据类型，部分字段内容在部分人群中有缺失，其中第一列为个体 id 号。
- 训练集文件的最后一列为标签列，既需要预测的目标值。
- 测试集文件的标签列为空，需要将预测结果上传至Kaggle。
- 提交说明：提交一个d_model.py 即预测的模型文件

二、数据—任务II

- 实验任务II数据为训练集文件f_train.csv，测试集文件f_test.csv
- 每个文件第一行是字段名，之后每一行代表一个个体，部分字段名已做脱敏处理。
- 训练集文件共包含85个字段，部分字段内容在部分人群中有缺失，其中第一列为个体 id 号。
- 训练集文件的最后一列为标签列，既需要预测的是否患病的类标。
- 测试集文件的标签列为空，需要将预测结果上传至Kaggle。
- 提交说明：提交一个f_model.py 即预测的模型文件

三、评估指标—任务Ⅰ

- 对于任务Ⅰ，需要提交对每个人的指标预测结果，以小数形式表示，保留小数点后三位。该结果将与个体实际检测到的结果进行对比，以均方误差为评价指标，结果越小越好，均方误差计算公式如下：

$$L = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

- 其中 n 为总人数， $\hat{y}_i$ 为预测的第 i 个人的指标值， y_i 为第 i 个人的实际指标检测值。

三、评估指标—任务Ⅱ

- 对于任务Ⅱ，需要提交对每个人是否患病的预测结果，以整数形式表示类别，取值为0或者1。该结果将与个体实际检测到的是否患病情况进行对比，以F1为评价指标，结果越大越好，F1计算公式如下：

$$F1 = \frac{2 * P * \mathrm{R}}{(P + R)}$$

- 其中P为准确率，计算公式如下：

$$P = \frac{\text{预测正确的正样本数}}{\text{预测的正样本数}}$$

- R为召回率，计算公式如下：

$$R = \frac{\text{预测正确的正样本数}}{\text{总正样本数}}$$

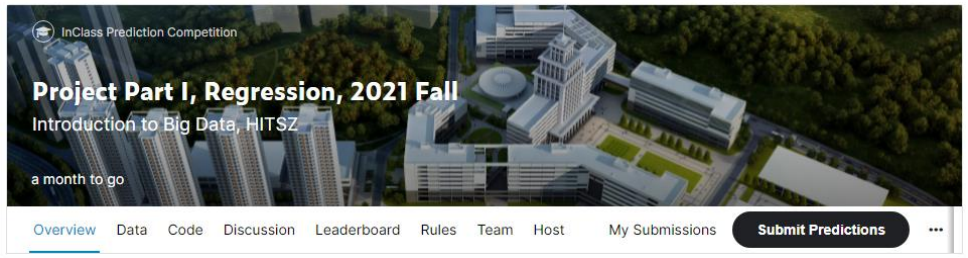
其中正样本数定义为数值为1的样本数

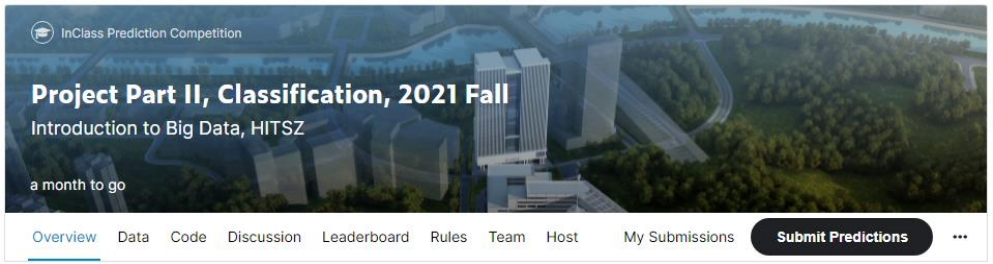
四、任务要求与说明

- 完成以下要求：
  - 在任务I编程完成对患病指标值的预测；
    - ✓ 预测结果请提交至Kaggle：

      https://www.kaggle.com/t/d6c4c37231844495b6cbe276d0a9271c
  - 在任务II编程完成对是否患病的预测；
    - ✓ 预测结果请提交至Kaggle：

      https://www.kaggle.com/t/1d61b5b2e8ef4f73aab131e8861800fa
  - 撰写一个总的实验报告 。





四、任务要求与说明

- Kaggle使用提示：
  - 每队仅限一人，请将队伍名称设置为学号
  - 可以在截止时间前多次提交预测结果，但在截止前需要选择一次提交记录作为最终结果，在对应提交记录一栏中勾选Use for Final Score即可。
  - 由于Public Leaderboard没有测试数据，因此提交后会显示Score为0。实际预测结果的Score会在截止时间过后在Private Leaderboard中显示。
  - 提交样例可以在data栏下载，如右图所示

id	Predicted
0	5
1	5
2	5
3	5
。。。	。。。

  - id列表示训练集对应个体，Predicted列表示对应的预测值
    请各位同学注意大家所提交的结果文件的第一行应为id，Predicted。
  - 需另外提交代码，要求提交的代码可复现预测结果。可以采用设置随机数种子的方式实现。
    - ✓ 例：np.random.seed(1024)

四、任务要求与说明（重点事项）

- 大作业报告中文书写，内容包含：实验目的、实验内容、实验过程、实验结果与分析；
- 大作业报告使用统一封面；
- 任务I和任务II数据于14周周一（2021.11.29）发布，请将预测结果提交至Kaggle；
- 最后需要提交的材料：
  - ✓ 任务I的代码d_model.py
  - ✓ 任务II的代码f_model.py
  - ✓ 大作业报告一份。
- 以上全部放到同一个文件夹里，统一命名为“学号_姓名_大作业”，并打包为zip文件；
- 群内“202100607_哈小深_大作业”文件夹为代码提交样例，命名格式错误将酌情扣分；
- 模型需按要求定义好对应的接口，具体参照代码提交样例；
- 开发环境版本务必参照群文件中的requirements.txt，统一使用 numpy、pandas、sklearn ；
- 大作业截止日期：19周周三（2022.1.5）22:00（UTC+08:00）前。