Mid-Term Programming Exercise

Data Mining, Spring 2018

数据科学与计算机学院

中山大学

Programming Exercise with kaggle

- kaggle public webpage
  - https://www.kaggle.com/c/datamining-2018-mid-term
- kaggle public link to join competition
  - https://www.kaggle.com/t/4778d5cccacd4972bf08a059f5529e85

训练数据

1. **train.csv**

This file includes the training data IDs in the first column, 128 real-value features of all the training data in the 2nd to 129th columns, and the labels in the last column. The number of training samples is 6000. Please be noted that there are 6 classes in total. So the class label of each training sample in the last column is an integer varied from 1 to 6. The format of each row is as follows:

```
ID, feature1, feature2, ..., feature128, label
```

测试数据

2. **test_raw.csv**

This file includes the test data IDs in the first column and 128 real-value features of all the test data in the 2 to 129 columns, while the labels of the test data are not available. The number of test samples is 7910. The format of each row is as follows:

```
ID, feature1, feature2, ..., feature128
```

测试数据的预测类标格式

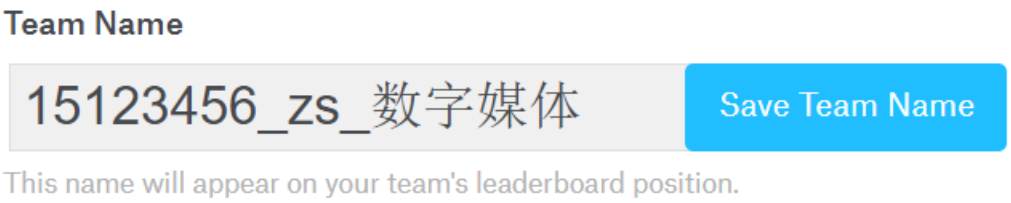
3. **sampleSubmission.csv**

The file namely "sampleSubmission.csv" contains the test data IDs in the first column and the predicted class labels (integers from 1 to 6) in the second column. Please be remembered to add "ID" and "Pred" to the first row of the "sampleSubmission.csv" file. The format of the 2 to 7911 rows is as follows:

```
ID, label
```

基本规则

- 每个人单独参赛，不允许组队
- 每人每天最多提交2次结果
- 独立完成，不能share结果
  - 系统可以自动检测share结果的情况，一旦发现会扣分
- 比赛有效时间：现在到2018/06/15
- 为方便统计成绩，大家参赛时创建team name的时候，请按照格式
  - 学号_姓名第一个字母_专业
  - 例如，张三的学号是15123456，专业是软件工程（数字媒体），则命名为“15123456_zs_数字媒体”



关于分数计算

- 两个baseline
  - 如果一位同学提交的结果高于这个baseline1（分类准确率为0.65908），则分数为60分以上
  - 如果一位同学提交的结果高于这个baseline2（分类准确率为0.88200），则分数为80分以上

#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
		baseline2			0.88200		
		baseline1			0.65908		

- 分数按排名和分类准确率（accuracy）算标准分

关于leaderboard上的排名

- 大家提交结果后可以看到你的方法的准确率和排名
- 请注意：这个准确率只是在30%的test data上的统计结果
- 当比赛结束后，系统会公布你提交的结果在全部test data上的错误率，从而决定你的最终排名。

Q&A

- 有关数据格式等问题，可以请教TA
- 对于一些常见问题，TA会在kaggle比赛系统的forum里回答
- 大家也可以在forum或微信群里提问和讨论
- 期中程序作业助教：
  - 谢国添, 1224617026@qq.com