期末大作业
Final-Exam Project

Data Mining, Spring 2018

数据科学与计算机学院

中山大学

Grading Scheme

- Homework: 30% (each 10%)
- Midterm programming assignment: 20%
- Final-exam project: 50% (due on 2018/07/18)
  - Kaggle competition (25%)
  - Final written report (25%)

Final-Exam Project with kaggle

- kaggle public webpage
  - https://www.kaggle.com/c/datamining2018-final-exam
- kaggle public link to join competition
  - https://www.kaggle.com/t/f0ce95e58a56487aa89cb14706754fbe

训练数据

1. **train_data.mat**

This file includes all the training data, which could be loaded by Matlab or Python. There are two variables in this file: 'train_feat', 'train_label',

```
'train_feat': a 76240x6812 matrix, with 76240 instances and each instance is the feature with 6812 dimensions.
```

```
'train_label': a 76240x1 matrix, which are the corresponding labels for the instances of 'train_feat'.
```

训练数据

4. **train_data_split.zip**

We split the 'train_feat' into 397 parts according to 'train_label', and save the features of each class into the files 'train_feat_class_000.mat' to 'train_feat_class_396.mat'. So the train_data_split.zip contains these 397 files.

测试数据

2. **test_data_raw.mat**

This file includes all the testing data. There is only one variable in this file: 'test_feat'.

```
'test_feat': a 19850x6812 matrix, with 19850 instances and each instance is the feature with 6812 dimensions. Each category contains 50 instances, and there are 397 categories.
```

测试数据的预测类标格式

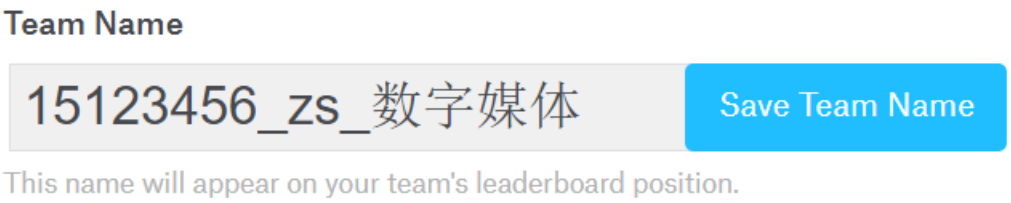
3. **sampleSubmission.csv**

The file namely "sampleSubmission.csv" contains the test image data IDs in the first column and the predicted class labels (integers from 0 to 396) in the second column. Please be remembered to add "image" and "label" to the first row of the "sampleSubmission.csv" file. Please name the csv file for submission as "学号_姓名第一个字母_专业", e.g. “15123456_zs_数字媒体”. The format of the 2 to 19851 rows is as follows:

```
img_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2018/07/18
- 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

- 三个baseline
  - 60分以上：分类准确率高于10%
  - 80分以上：分类准确率高于18%
  - 90分以上：分类准确率高于22%

#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
		baseline3.csv			0.24332		
		baseline2.csv			0.20100		
		baseline1.csv			0.13618		

- 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。

Final Written Report

- 提交方法
  - 以电子邮件方式提交PDF文档到邮箱地址，DataMining_2018@126.com
  - 邮件标题格式：“report_学号_姓名”
  - 提交截至时间： 2018/07/18
- 实验报告包括两部分
  - 分类算法介绍与分析
    - 课堂上所介绍的算法，或者相关研究领域的算法（请引用相关的参考文献）
    - 分析所用算法理论上的优缺点、时间复杂度、内存需求等
  - 实验结果分析
    - 对比不同算法的分类效果和计算复杂度

Q&A

- 有关数据格式、实验报告等问题，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
- 期末大作业助教：
  - 谢国添, 1224617026@qq.com