项目简介与要求

徐泽南

2018.9.26

Project

- 两个任务
  - 二多元分类，以 **分类正确率（ Accuracy ）** 作为评测指标
  - 回归，以 **皮尔逊相关系数（Pearson Correlation）**作为评测指标
- 竞赛制
  - 组队后，每个队伍每天可提交自己的结果到 ftp，TA 会跑 rank，然后把 rank 的情况发给大家（群里excel），如果提交的结果是空的则分数为 0，分数越高排名靠前
  - Rank由Project验收前（10.21）和验收后截止组成（10.24）
  - rank占实验总评 **30%**

Rank 提交结果说明

- 关于二元分类，数据全都是txt格式，用utf-8编码
- 所以提交的二分类标签也请用txt格式，utf-8编码
- 训练集标签就是0或者1，提交1/0即可，不用1.00，一行一个结果

- 关于回归，数据是存储在excel表格中
- 但是提交的回归结果也请用txt格式，utf-8编码，一行一个结果，可能结果是7.32

Rank注意

- 关于分类和回归会有两个文件夹，请对应文件夹提交
- 测试集是 N 行，提交上来的行数也必须是N行
- 文件名出现非法字符（命名格式，学号_次数）
- 每天最多提交10个文件（学号_1.txt,学号_2.txt,学号_9.txt）
- 学号请输入正确，不然无法确认是谁的
- 请不要压缩，直接提交txt文件即可
- 内容包含有非结果字符，比如多了一行“正确率”字样，多了一列文本序号等等
- 编码出错，TA调用的时候使用函数是open(path, r,encoding=utf-8)，请统一用utf-8编码

数据集介绍

- 每个任务都会提供一个数据集
- 二元分类：影评预测正负
- 提示，请不要作弊或者上网搜原数据集，每个数据集会有一个state-of-the-art基准线，如果你超过了目前世界上最好的效果，在验收的时候会详细让你解释模型如何设计。（如果效果真的好，会帮助你发论文）与其搜索答案胆战心惊故意搞错部分答案打rank，不如认真训练模型。
- 有作弊嫌疑的分数会全部要求提交模型，然后TA亲自跑，发现抄袭整组一律0分

训练集**有效**行数	测试集行数	**输入单位**	输出
24000	**6000**	**句子**	**1 或 0**

数据集介绍

- 五分类：电影评论预测类别。

训练集**有效**行数	测试集行数	**输入单位**	输出
24000	**6000**	**句子**	**1或0**

数据集介绍

- 回归：酒店评分预测，给予酒店的一些特征，预测顾客的分数。
- 这里有一个小加分点，有一行的特征是这个酒店具备的一些标签，关于这一个特征如何处理。可以选择忽略，也可以选择将标签表示成one-hot，标签数很少，总数不会超过10种情况，具体如何使用可以仔细思考

训练集**有效**行数	测试集行数	**有效**属性个数	输出
80000	**20000**	**7**	**酒店频评分**

算法

- 没有规定使用某种算法
- 学过的算法：KNN, NB, PLA, DT, LR, NN
- 全新的算法：SVM, SVR, …
- 挑战性算法：CNN，RNN，LSTM，・・・
- 鼓励大家尝试新算法
- **所有算法，都必须是自己实现的，不可以调用现成库（但是如果使用CNN，RNN，LSTM算法，允许调用CNN，RNN，LSTM框架）**
- 比如想在 NN 里面用 PLA，PLA部分也要自己实现
- 在 Project 报告中，将你使用的所有方法展示出来

验收

- 内容：团队如何分工，自行测试的方法，每个任务使用的方法，结果，改进思路。
- 时间：**每组验收约 15 分钟，包括提问**
- 要求：每个组的所有成员都要验收

评分标准

- 排名 30%（小组算分，同一个小组所有成员相同）
- 验收 30%（验收以小组为单位，但是每个人单独算分）
- 报告 40%（个人为单位，单独算分）
- 加分（5~20%）（小组算分）
  - 尝试新算法
  - 或者分数较高且算法有创新
  - 除此之外的一些优秀的情况

最终提交

- 报告提交DDL：（10月21日晚上11点）23:00:00
- 提交内容：
  - 实验报告，每个人一份，命名为：组号_学号_姓名拼音_report.pdf，如“2_15350000_xiaoming_report.pdf”
  - 无论之前是否已经有成绩，都要提交一份最后的结果跑最终rank。
  - 源码zip，包含多个文件，命名为：组号_code.zip，如“2_code.zip”，里面包含一个 **readme** 文件，阐述各文件用途

组队名单上报

- **9月30日晚0点**前确认分组
- 一组3~4人，评分标准**没有区别**，推荐组队完成
- 确认分组后，上交一份 txt 文件到 ftp “**组队信息**”文件夹，到时候没有组队的同学会直接强制组队。
- txt命名为：组长学号. txt
- txt中包含以下内容：
  - 组内各成员学号，姓名
  - 队伍名字（自己定一个，会在 rank 的时候出现）

提交 Rank 重点注意事项

- 只需要提交**结果！结果！**
- 不要多一列文章序号，不要多一行文字介绍
- Test 几行有效数据，答案就提交几行，请提交前自行确认。

- 每天都可以提交！记得自己存好最佳 rank 的结果文件，最后上交

提交 Rank 要求

- 每天 ftp 的结果文件夹会每天清空，**请严格按照要求，否则无rank，浪费一天的等待**。

- 每天可以提交十个版本，**多于十个版本的不会处理**，用**0~9**区分，就算如果只有一个版本，也要加“0”。