Artificial Intelligence Project

林东定, 黄行昌

November 22, 2016

1 任务描述

Project使用竞赛制。在不同的数据集(共4个) 和相应的评测指标(共3种) 下，使用相应的算法(算法不限，使用你认为最合适或最优的算法)，实现出分类/回归的结果，最后把最好的预测结果提交到ftp (提交的是预测结果，TA 会帮你跑出评测指标的结果进行排名，不提交则该数据集得分为0)，得到一个排名(所有队伍，包括单挑和双挑，都会有一个排名)，排名越高则得分越高。

1.1 数据集

这次Project包含4个数据集，每个数据集都分为训练集和测试集，鼓励在训练集中使用交叉验证的方式进行调参，并在最后的Presentation 中展示，以下为4个数据集介绍：

Table 1: 数据集描述

数据集	属性	训练集	测试集	输出	回归or分类	评测指标
1.新闻分享	58	23786	15858	是否被分享	二元分类	F1-measure
2.成年人收入	14	30162	15060	是否大于50K	二元分类	F1-measure
3.ISEAR	多少个单词	5271	2395	7种情感中的1种	多元分类	平均正确率
4.蛋白质	16	13902	3477	蛋白质残留量	回归	RMSE

- 1. 新闻分享：该数据集描述了新闻文本属性(58种属性) 和对应是否会被分享的关系，属于二分类问题，评测指标为F1-measure。该数据集大小如下：训练集为23786 行数据，测试集为15858 行数据。对于每行数据，标签为是否被分享。（分享为1，不分享为0）
- 2. 成年人收入：该数据集描述了一个成年人的属性(14种属性) 和对应收入是否大于50K的关系，属于二分类问题，评测指标为F1-measure。该数据集大小如下：训练集大小为30162行数据，测试集大小为15060 行数据。对于每行数据，标签为是否大于50K。（大于50K为1，小于等于50K为0）
- 3. ISEAR：该数据集描述了新闻标题文本和对应情感的关系(跟之前实验使用的semeval 数据集相似)，一共7种情感，对每篇文本，拥有一个对应的文本(已经分词好了)，第一列为标签，第二列为对应文本(跟之前semeval做分类类似)，评测指标为平均正确率。该数据集大小如下：训练集大小为5271个文本，测试集大小为2395个文本。（标签为anger等7种情感）
- 4.蛋白质残留：该数据集描述了9种因素和对应蛋白质残留量的关系，属于回归问题，评测指标为RMSE。该数据集大小如下：训练集大小为13902 个文本，测试集大小为3477 个文本。（标签为残留量）

1.2 算法

不限定使用某种算法，可以使用已学的如KNN,NB,PLA等，也可以根据自己的兴趣阅读论文，使用一些对KNN,NB,PLA等的改进算法或是全新的方法; 还可以使用SVM,神经网络等(不过需要自己实现，不能直接调用现成的库)，最后每个数据集选择出一种你认为最好的方法，并在Presentation中展示。

1.3 排名提交方式

每个组需要在每个数据集各提交一份结果，注意提交的是预测结果的txt文件而不是评测指标(在新闻分享数据集中提交你的0或者1的预测结果，成年人收入也是0或者1 的预测结果，蛋白质残留是输出分享数，问卷调查是输出你的情感值(anger等等))，输出格式为每行一个数字，txt 文件命名为：组号_数据集编号.txt(数据集编号如下：新闻分享为001_1，成年人收入为001_2，问卷调查为001_3，蛋白质为001_4)。
在开放排名之后，每个组每天可以提交上限十次结果（后面加_v1-_v10），请注意按照附件给出的提交格式，每次提交都要交所有的数据集结果哟！！！也就是说就算你只优化了一个part，你也要交里面含有四个txt的压缩包上来哟。我们会把提交的所有版本的ranking结果都告诉你们。

2 时间轴

为了让大家能更好地安排时间完成Project，我们给出了几个关键时间点和时间段的安排，希望大家尽早开始实验，并顺利完成Project 的任务：

Table 2: 时间轴

时间	安排
第12周	发布Project文档和数据集，开始实验
第12-14周· · · 第13周	自主实现代码和评测指标，跑出初步结果
第13周开始	开放提交平台（每一天中午12点会发布昨天提交的Rank更新）
第14-15周	根据排名，优化自己的算法，争取进步，按照顺序做presentation
16周最后一天	提交完整实验材料（如报告，结果等），之后发布ranking最终结果

3 评测指标

我们根据二元，多元分类和回归问题给出三种评测指标：二元分类使用F1-measure，多元分类使用平均正确率(Average Precision)，回归使用RMSE。在分类中，通常以关注的类为正类，其他类为负类(e.g. 在新闻分享中，被分享的新闻为正类，未被分享的为负类；而在新闻标题和情感中，某一种情感为正类时，其他情感都为负类)，分类器在测试数据集上的预测或正确或不正确，共4 种情况，分别记作：

- TP：将正类预测为正类
- FN：将正类预测为负类
- FP：将负类预测为正类
- TN：将负类预测为负类

精确率(Precision)和召回率(Recall)的计算方法：

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

此外F1-measure，即F1值，是精确率和召回率的调和均值，即：

$$\frac{2}{F1} = \frac{1}{P} + \frac{1}{R} \tag{3}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

精确率和召回率都高的时候，F1值也会高，因此F1值越高，分类效果越好，排名越高。

多元分类：使用的评测指标为平均正确率(Average Precision)。在新闻标题和情感数据集中，共有7 种情感，根据公式(1) 分别计算出7个精确率(Precision)，然后取他们的平均值作为评测结果，平均正确率(Average Precision) 的值越高，效果越好，排名越高。

最后是回归的评测指标RMSE(root-mean-square error)，也叫均方根误差，计算方法如下：

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{5}$$

其中n表示n个回归预测值，$\hat{y}_t$表示你的每个回归预测值，y_t表示真实的回归值。因此，RMSE的值越小，表示误差越小，回归效果越好，排名越高。

4 Presentation

做presentation的时候，不一定要求已经有了一个最终版本，但是要看到每一个部分必须都要已经完成了一些工作，并且要有一些自己的想法

- 内容：PPT展示使用的算法，结果，改进方法或优化，测试方法，分工等
- 时间：第14-15周，每组展示不超过5分钟(展示时间4分钟，提问时间1分钟)
- 人数：每个组都需要齐人，每人都需要发言
- 顺序：当堂分配

5 最终评分标准

- 排名得分(50%)，其中四个数据集的排名的权重按顺序分别为(新闻分享25%, 成年人收入25%, 问卷调查25%, 蛋白质残留25%)，综合4个数据集的排名给出一个综合排名
- Presentation(20%)，会根据每个人的贡献评分，小组成员不一定是同一个分数
- 实验报告(25%)，一个组一份
- 加分点(5-20%)，包括：1 有足够的理论依据证明你的优化方式是可信的，2.新算法，即非实验课实现过的算法

6 提交内容

注意这个是最终版，要和日常排名分开。提交样式也在附件给出。

- 实验报告：每个组一份，命名为：组号_report.pdf
- Presentation PPT：每个组一份，命名为：组号_presentation.pptx

- 结果文档：每个组各提交一个文件夹，文件夹名称为组号_result.zip，里面包含了4个txt，注意提交的是预测结果而不是评测指标(比如在新闻分享数据集中提交你的0和1 的预测结果)，命名为：组号_数据集编号.txt，格式：一个预测标签一行。
- 实验源码：每个组一个文件夹，命名为：组号_code.zip（不一定是cpp文件，可以是py文件等等，按你实现的代码种类来）

7 非常重要的注意事项

来自东定宝宝的忠告！不要使用任何压缩软件来压缩文件！否则我们写的ranking代码会解压出奇怪的东西。请使用原生的打包格式，windows和mac都有系统自带的打包方式，穷人决定只演示windows 的打包方法，如Figure 1所示：

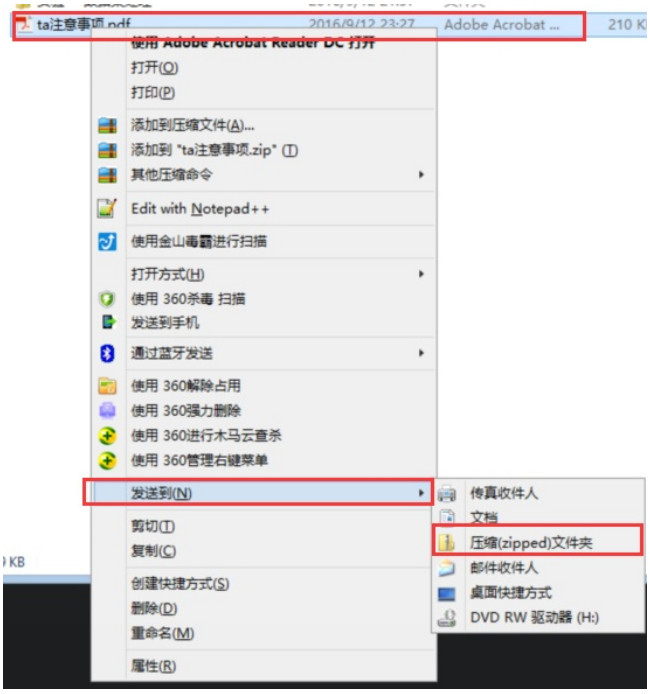

Figure 1: windows原生打包方式

还有，请大家直接对四个文件进行全选然后压缩，要保证你的压缩包打开直接是四个文件，而不是一个文件夹，正确示范如Figure 2 和Figure 3 所示：

错误示范示范如Figure 4 和Figure 5 所示：

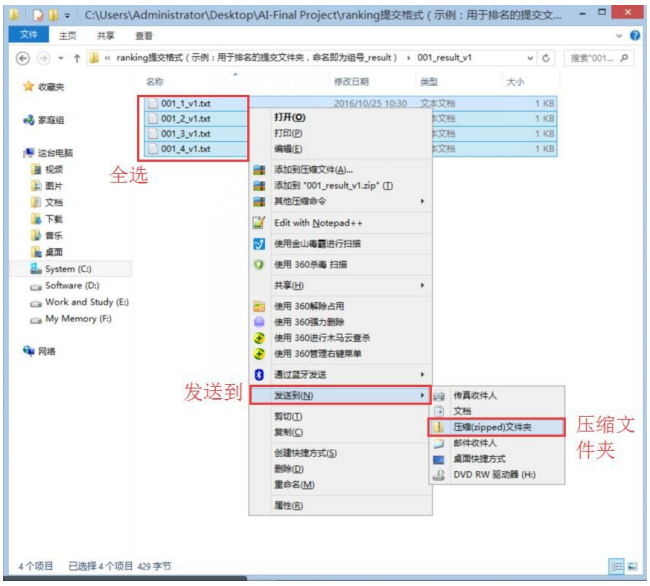

Figure 2: 正确步骤1

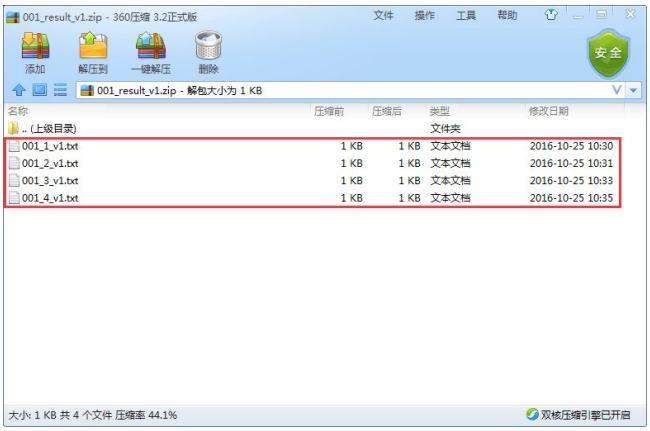

Figure 3: 正确步骤2

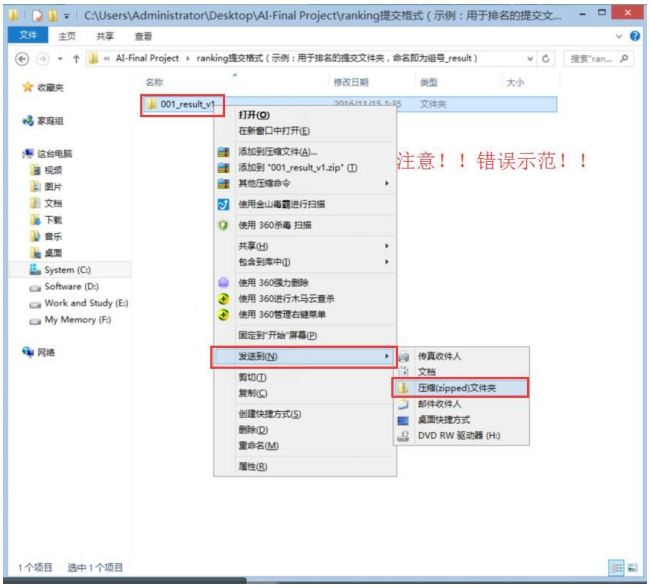

Figure 4: 错误步骤1

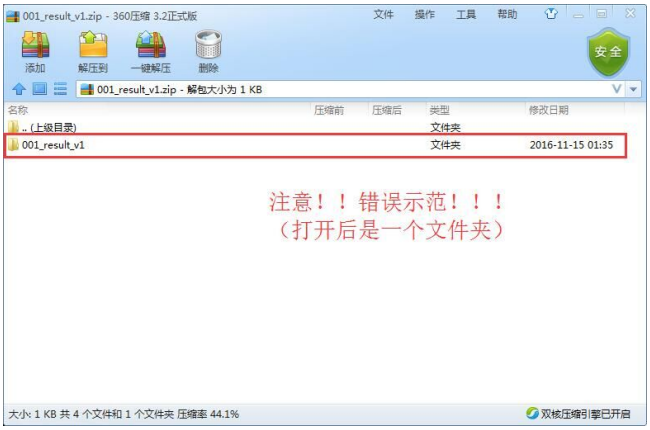

Figure 5: 错误步骤2