Task 1 报告

姓名: 王迎旭

学校:中山大学数据科学与计算机学院

1. 完成的任务

- 使用 pandas 读入训练集与测试集,并将文本使用词袋模型转为可供进行计算的向量集
- 理解了 N gram原理与实现方法
- 实现 softmax 分类
- 不调用库函数实现梯度下降法与 loss 函数
- 对测试集与训练集进行划分
- 了解 shuffle 的基本原理与应用
- 个人在校数据挖掘HW2实验报告(内容为:随机梯度下降法、梯度下降法、批量梯度下降法的实现以及对不同 线性机器学习算法学习率的研究)

2. 实验结果

总共进行了 次实验

2.1 实验 1

条件:

- 特征选择: gram = 1, 词袋向量长度为默认向量长度
- 学习率 α = 0.1
- batch = size(训练集行数)
- 训练次数: 1000 (由于训练时间比较久, 计算机资源有限就只把训练次数设置为1000)

实验结果:

```
now step is : 100
now step is : 200
now step is : 300
now step is : 400
now step is : 500
now step is : 600
now step is : 700
now step is : 800
now step is : 900
now step is : 1000
0.5108291682686147
```

如图可知:训练集划分之后,对校正集进行测试准确率为51.0829%

对测试集进行测试之后, 生成的 csv 文件传到 kaggle 之后, 准确率为:

"everything you 'd expect -- but nothing more" *** Overview Data Kernels Dis

Sentiment Analysis on Movie Reviews

Classify the sentiment of sentences from the Rotten Tomatoes dataset 861 teams \cdot 4 years ago

Overview Data	Kernels Disc	ussion Leaderboard	Rules Tea	ım	My Submissions	Late Submission
Your most recent su	bmission					
Name		Submitted		Wait time	Execution time	Score
final.csv		just now		0 seconds	1 seconds	0.51789
Complete						
<u>lump to your positio</u>	n on the leader	board_▼				

2.2 实验2

条件:

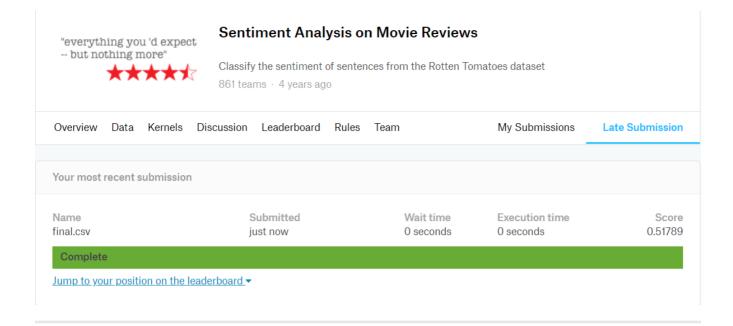
- 特征选择: gram = 2, 词袋向量长度为 15000(如果这里不限制词袋向量长度, 会出现内存错误)
- 学习率 α = 0.05
- batch = size(训练集行数)
- 训练次数: 1000 (由于训练时间比较久, 计算机资源有限就只把训练次数设置为1000)

实验结果:

```
now step is: 100
now step is: 200
now step is: 300
now step is: 400
now step is: 500
now step is: 600
now step is: 700
now step is: 800
now step is: 800
now step is: 900
now step is: 1000
The Accuracy of correction set is 0.5096116878123799
```

如图可知:训练集划分之后,对校正集进行测试准确率为50.9611%

对测试集进行测试之后, 生成的 csv 文件传到 kaggle 之后, 准确率为:



从结果来看,调参对预测结果并没有什么大的影响;

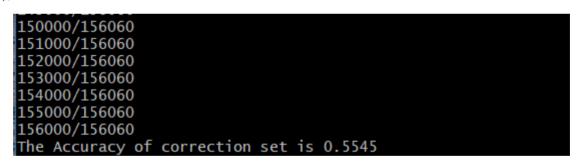
实验 3

为了探求是否模型已经达到调参的极限,我选用了效果最好的标准库中的 **随机森林** 模型来对给定的数据集进行测试,并求出相应的准确率进而来进行对比;

条件:

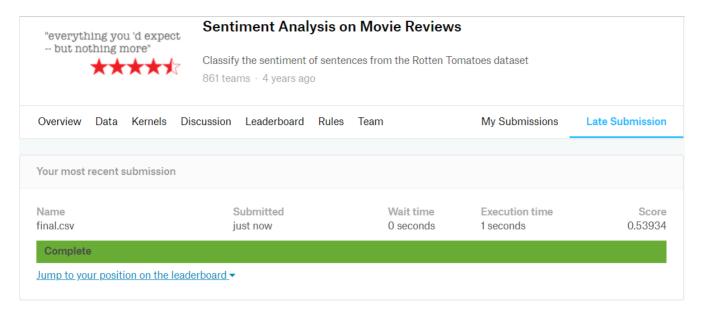
• 随机森林数目为 100

实验结果:



如图可知:训练集划分之后,对校正集进行测试准确率为55.45%

对测试集进行测试之后, 生成的 csv 文件传到 kaggle 之后, 准确率为:



从随机森林的准确率上可以得知,由于模型方法的局限性,softmax分类已经达到的最优的效率,再调参数进行训练意义也并不大,同时由于给定的数据集是5分类问题,准确率低也是正常现象;