Report for Exp1NavieBayes

计65 王起旺 2016050029

1、源代码设计

数据处理

BuildSet.py中对数据进行初始化的处理，主要的操作包括将6w+条数据分成五折，实现为通过random.random()的值进行分配，将分配好的数据以路径为id存放在./dataset内。

统计词频计算准确率

将上一步分好的数据进行五折交叉验证，每次取一份为验证集，其余四份为训练集，统计训练集中所有中文字词的出现次数，以及'http'出现的次数，对验证集计算在ham和spam中取log计算预测概率。

代码结构

./BuildSet.py
----> 数据分割
.
.
.

./train.py
----> loadSet(setName)
加载分割好的数据，setName为数据集名字
---->train(sample,laplace)
五折交叉验证计算正确率将max模型存放在./trainData/data
sample为对于数据集中每一条数据进行统计的概率，laplace为参数

2、结果以及分析

目前在随机划分的验证集中正确率大概为0.9738,相较来说是比较低的，目前所做的就是拉普拉斯平滑，乘积取log叠加减少精度损失，匹配"http"添加到字词列表，再加一些邮件地址以及其他之类的feature可能准确率会有所提升。

3、issue相关

1、the size of training set

我选取了sample=1、=0.5、=0.2分别进行五折交叉验证，
得到准确率为

```
sample=0.5

cur ac is ------------- 0.9689832986992997
Average is ---------  0.9657347237763201

sample=0.2

cur ac is ------------- 0.9648272146540445
Average is ---------  0.9615550238677724


sample=1

cur ac is ------------- 0.9722927730316324
Average is ---------  0.9708542870550552
```

可见训练集的大小对结果的准确率确实有影响，而且数据集越大往往越准确，但是对结果的影响并不是很大，分析原因很可能是实现了拉普拉斯平滑的功劳，即使当数据集较小时，也可以通过给其赋一个较小的值来减少对结果的影响，取消拉普拉斯平滑后也印证了上述的分析，训练集的大小对结果产生了很大的影响。

2、zero-probabilities

零概率的问题就和训练集的大小有关，当训练集比较小的时候可能会发生一些小概率的事件，比如出现某些特殊的字词只出现某一份样本中，这样就会导致在其他样本中该字词的概率为零导致零概率的产生。于是解决此问题可以对将出现的零概率对其赋一个极小的值来阻止零概率的产生，我在train.py中实现了laplace平滑处理，train(sample,laplace)中第二个参数的值就是拉普拉斯中的参数值，经过实验发现取值为1e-80时效果较好。此时五折交叉验证平均概率为0.970854。

3、specific features

目前只实现了根据"http"这一特征作为额外特征加入字词列表中，因为经过对比发现垃圾邮件中包含有大量字符http，而非垃圾邮件中则不然，不加http这一特征之前准确率大概为0.94，实现后为0.9784,可见实验结果也足以证明其作用。
另外因为时间问题暂时没有实现的一些其他特征，比如发件人的邮箱地址，时间等等都是一些很有特征的线索，相信增加之后会对准确率有更大的提升，以及这些特征举足轻重的作用使我们可以增加他们的权重，使得增加其对最后结果的影响。