实验报告

实验设计

> 要求：使用朴素贝叶斯分类器实现垃圾邮件分类。

将数据读取完之后，使用 `scikit-learn` 包中的 `StratifiedKFold` 进行五折交叉验证。共进行五轮，其中每轮将整体数据的 $\frac{4}{5}$ 数据作为训练集，剩余的 $\frac{1}{5}$ 数据作为测试集。

代码中设计了一个类 `NaiveBayes` ，含有两个标签分别对应的dictionary用于单词计数，实现了两个方法 `train` 和 `predict` 。
其中 `train` 函数输入训练集，无输出。 `predict` 函数输入测试集的邮件内容，输出分类。依次调用完两个函数后比较 `predict` 函数的输出和测试集的标签，进行分析和评估。

除此之外，还实现了两个函数 `check` 和 `important` ， `check` 函数用以判断来去掉一些如乱码之类信息量不大的数据， `important` 函数用来判断含有 `received, subject` 等重要特征信息的句子。接下来简要介绍 `NaiveBayes` 类内的两个方法

`train`

对于每个数据对 `(x, y)` ，前者为文本，后者为标签。首先枚举文本中的每一行，使用 `important` 函数判断该句话的重要性。之后对于句中的每个无空白字符的单词段，使用 `check` 来忽略有乱码有数字的单词，然后在相应标签的dictinary中进行计数，如果处于重要语句中计数 β ，否则计数 1 。在最后会将dictionary中计数只为 1 的单词给抛掉，其中大多为无意义单词，并且可以降低复杂度。

`predict`

对于数据 `x` ，与上同理，枚举文本每一行并判断重要性，对于每个单词检测乱码以及数字，剩下的有效单词用以计算 $P(y|x_1,\dots,x_1) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$ ，为了防止丢精度，故计算 $\log P(y) + \sum_{i=1}^{n} \log P(x_i|y)$ ，因为 $\log$ 为单调函数。与上同理，对于重要的语句，其中的单词会被多次计数，在乘式中表现为多次方，在 $\log$ 函数的和中表现为乘以系数 β 。于此同时，本代码还使用Laplace平滑来解决零概率问题，其中取系数 $\alpha = 1$ ，在概率式中表现为 $P(x_i|y) = \frac{\#\{y=c,x_i=k\}+\alpha}{\#\{y=c\}+M\alpha}$ ，其中 M 为所有 x 在dictionary中的去重后计数。所有根据两个标签计算得到的 $\log P(y) + \sum_{i=1}^{n} \log P(x_i|y)$ 取较大值对应的标签放入结果中。

实验结果

以下为使用五折交叉检验，并采样 $5\%, 50\%, 100\%$ 的训练集数据进行训练后，五次测试平均的指标结果（spam为阳性）：

训练数据	Accuracy	Precision	Recall	F1
5%	0.9672943462318291	0.9596503144476148	0.9450813323005421	0.9514848207106044
50%	0.9876790467106338	0.9775945067561855	0.9865220759101471	0.9820343300728096
100%	0.9904552297115461	0.9822481983335957	0.9899302865995352	0.9860737212244752

实验分析

- 训练集规模：可以发现无论是哪一个指标，都随着训练数据集的规模的增加而升高，说明一般而言，更多的数据可以使泛化效果更好。
- 从最开始的基础结构到目前的结果，依次增加了如下的方案：
  - Laplace平滑，使得 `predict` 可以正常进行。因为零概率问题会导致当该单词在训练集中没有出现过时，最后概率计算为 0，且在本代码中会使 $\log$ 函数出错
  - 去掉一些乱码及带有数字的单词，提升了效率。因为减少了dictionary和需要查询的单词数
  - 在dictionary中去掉出现次数为 1 的单词，提升了效率。因为减少了dictionary和需要查询的单词数
  - 增加了对一些特征的加权处理，如 `received` 所在的语句等，使性能有所提升。这应该是因为许多垃圾邮件都会在邮件的头部有些明显的标志，故对此加权处理的话可以体现该部分的影响

实验讨论

- 对于垃圾邮件分类，precision较为重要，还有待提升
- 若进一步引入验证集，可以对代码内的参数 α, β 进行调整
- 有些数字信息是可以作为特征进行使用的，有待进一步提取文本的特征信息