信 息 内 容 安 全 实 验 报 告

实验名称： 基于朴素贝叶斯的垃圾邮件过滤器

班级	学号	姓名	承担的主要工作
SC011701	2017302231	王梦涵	朴素贝叶斯原理，贝叶斯分类的常用模型，垃圾邮件过滤的伯努利模型实现与结果分析
SC011701	201730223	罗倩倩	垃圾邮件过滤的多项式模型实现与结果分析，性能比较

基于朴素贝叶斯的垃圾邮件过滤器

一、实验原理

1.1 背景

在概率论与统计学中，贝叶斯定理（Bayes’ theorem）表达了一个事件发生的概率，而确定这一概率的方法是基于与该事件相关的条件先验知识。利用相应先验知识进行概率推断的过程为贝叶斯推断。

1.2 贝叶斯估计

条件概率指在事件B发生的情况下，事件A发生的概率，通常记为$P(A|B)$。

$$P(A|B)=\frac{P(AB)}{P(B)}$$

同理可得：

$$P(B|A)=\frac{P(AB)}{P(A)}$$

由此可推出条件概率的计算公式：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

设Ω为实验E的样本空间，B_1，B_2，…，B_n为一组事件，若

（1） $B_iB_j=\emptyset, i\neq j, i,j=1,2,\cdots,n;$

（2） $B_1\cup B_2\cup\cdots\cup B_n=\Omega$，

则称B_1，B_2，…，B_n为样本空间Ω的一个划分，或称为实验E的一个完备事件组。

设实验E的样本空间为Ω，A为E的事件，B_1，B_2，…，B_n为Ω的一个划分，

且$P(B_i)>0$，$i=1,2,\cdots,n$，则有

$$P(A)=P(B_1)P(A|B_1)+P(B_2)P(A|B_2)+\cdots+P(B_n)P(A|B_n)$$

即全概率公式。

全概率公式的含义在于：设划分中的事件B_1，B_2，…，B_n是事件A发生的全

部“原因”，那么每个“原因”发生的概率与该“原因”导致 A 发生的概率乘积的和即为 A 发生的概率。

由条件概率、乘法公式及全概率公式可得贝叶斯公式：

$$P(B_i|A)=\frac{P(B_iA)}{P(A)}=\frac{P(B_i)P(A|B_i)}{P(A)}$$
$$=\frac{P(B_i)P(A|B_i)}{P(B_1)P(A|B_1)+P(B_2)P(A|B_2)+\cdots+P(B_n)P(A|B_n)}$$

即：

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_j)}\quad i=1,2,\cdots,n$$

贝叶斯公式的意义在于它反映了导致一个事件发生的若干“因素”对这个事件的发生的影响分别有多大。

取$n=2$，并记$B=B_1$，$B_2=\bar{B}$，则全概率公式可以改写为：

$$P(A)=P(B)P(A|B)+P(\bar{B})P(A|\bar{B})$$

贝叶斯公式可以改写为：

$$P(B|A)=\frac{P(B)P(A|B)}{P(B)P(A|B)+P(\bar{B})P(A|\bar{B})}$$

$$P(\bar{B}|A)=\frac{P(\bar{B})P(A|\bar{B})}{P(B)P(A|B)+P(\bar{B})P(A|\bar{B})}$$

这是全概率公式和贝叶斯公式的两个常用形式。

1.3 贝叶斯推断

对条件概率公式进行变形，可以得到如下形式：

$$P(A|B)=P(A)\cdot\frac{P(B|A)}{P(B)}$$

我们把$P(A)$称为“先验概率”，即在 B 事件发生前，对 A 事件概率的一个判断。$P(A|B)$称为“后验概率”，即在 B 事件发生之后，对 A 事件概率的重新评估。$P(B|A)/P(B)$称为“可能性函数”，这是一个调整因子，使得预估概率更接近真实概率。所以，条件概率可以理解为：后验概率=先验概率×调整因子。这就是贝叶斯推断的含义。我们先预估一个先验概率，然后加入实验结果，看这个实验到底是增强还是削弱了先验概率，由此得到更接近事实的后验概率。

在这里，如果“可能性函数”$P(B|A)/P(B)>1$，意味着“先验概率”被增强，事件 A 发生的可能性变大；如果“可能性函数”$P(B|A)/P(B)=1$，意味着事件 B 无助于判断事件 A 的可能性；如果“可能性函数”$P(B|A)/P(B)<1$，意味着“先验概率”被削弱，事件 A 发生的可能性变小。

朴素贝叶斯推断，是在贝叶斯推断的基础上，对条件概率分布做了条件独立性的假设。

1.4 朴素贝叶斯分类

利用贝叶斯进行邮件过滤的核心在于利用后验概率公式：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

这个公式只适用于单个垃圾过滤关键字。

如果包含多个垃圾关键字，可将上面的公式进行推广：

$$P(A|B_1,\ B_2,\ \cdots,\ B_n)=\frac{P(B_1,\ B_2,\ \cdots,\ B_n|A)P(A)}{P(B_1,\ B_2,\ \cdots,\ B_n)}$$

由此，根据扩展的后验概率公式，可知对于收到的一份邮件 E，该邮件出现的单词集（word 1,word 2, …,word n），计算该单词集出现的情况下该封邮件可能是垃圾邮件(Spam)或正常邮件（Ham）的后验联合概率：

$$P(Spam|word_1,\ word_2,\ \cdots,\ word_n)$$

$$=\frac{P(word_1,\ word_2,\ \cdots,\ word_n|Spam)P(Spam)}{P(word_1,\ word_2,\ \cdots,\ word_n)}$$

或

$$P(Ham|word_1,\ word_2,\ \cdots,\ word_n)$$

$$=\frac{P(word_1,\ word_2,\ \cdots,\ word_n|Ham)P(Ham)}{P(word_1,\ word_2,\ \cdots,\ word_n)}$$

由于这两个条件概率分布是一个 n 维空间向量的联合概率，如果每个特征值有 t 种取值，那么可能的情况有t^n次，为指数级，求解该问题是一个 NP 难问题。为简化问题，需要进行朴素而大胆的假设：特征项之间相互独立，由此 n 维的联合概率可简化成 n 个分离的概率乘积，这将问题降到易于计算的多项式级别。在

此独立假设下，联合后验条件概率可变为如下形式：

$$P(Spam|word_1, word_2, \cdots, word_n)$$

$$= \frac{P(word_1, word_2, \cdots, word_n|Spam)P(Spam)}{P(word_1, word_2, \cdots, word_n)}$$

$$= \frac{P(word_1|Spam)P(word_2|Spam)\cdots P(word_n|Spam)P(Spam)}{P(word_1)\mathrm{P}(word_2)\cdots P(word_n)}$$

分子分母均为先验概率乘积，统计计算都比较容易。

1.5 朴素贝叶斯分类的流程

1. 设$x = \{a_1, a_2, \cdots, a_m\}$为一个待分类项，而每个 a 为 x 的一个特征属性。

2. 有类别集合$C = \{y_1, y_2, \cdots, y_n\}$。

3. 计算$P(y_1|x)$, $P(y_2|x)$, …, $P(y_n|x)$。

4. 如果$P(y_k|x) = \max\{P(y_1|x), P(y_2|x), \ldots, P(y_n|x)\}$，则$x \in y_k$。

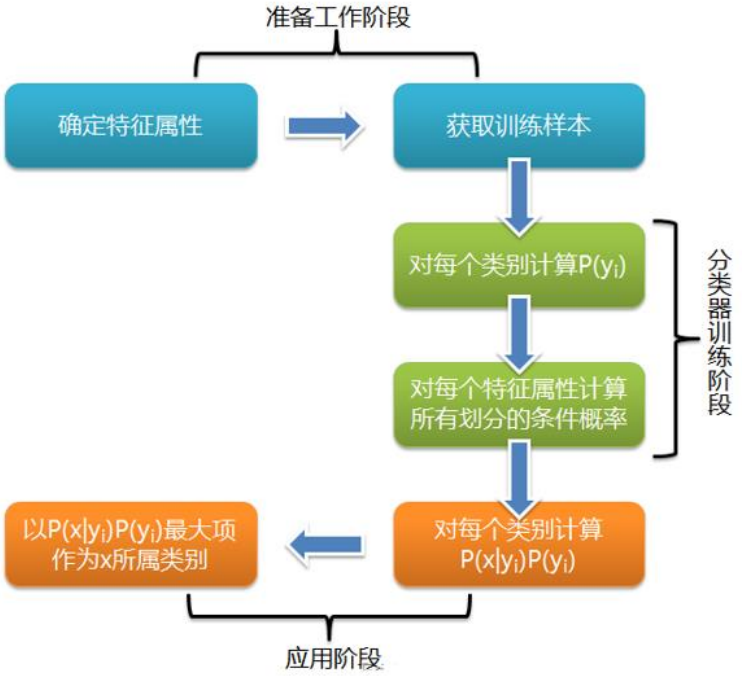
计算第 3 步中的各个条件概率：

1. 找到一个已知分类的待分类项集合，这个集合叫做训练样本集。
2. 统计得到在各类别下各个特征属性的条件概率估计，即 $P(a_1|y_1)$, $P(a_2|y_1)$, …, $P(a_m|y_1)$; $P(a_1|y_2)$, $P(a_2|y_2)$, …, $P(a_m|y_2)$; …; $P(a_1|y_n)$, $P(a_2|y_n)$, …, $P(a_m|y_n)$。
3. 如果各个特征属性是条件独立的，则根据贝叶斯定理有如下推导：

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

因为分母对于所有类别为常数，所以只要将分子最大化即可。又因为各特征属性是条件独立的，所以有：

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i)\ldots P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i)$$



二、实验模型

2.1 高斯模型

当特征是离散变量时，使用多项式模型。多项式模型在计算先验概率$P(y_k)$和条件概率$P(x_i|y_k)$时，会做一些平滑处理，具体公式为：

$$P(y_k)=\frac{N_{y_k}+\alpha}{N+k\alpha}$$

其中，N 是样本总数，k 是类别总数，N_{y_k}是类别为y_k的样本个数，α是平滑值。

$$P(x_i|y_k)=\frac{N_{y_k,x_i}+\alpha}{N_{y_k}+n\alpha}$$

其中，N_{y_k}是类别为y_k的样本个数，n是特征的维数，N_{y_k,x_i}是类别为y_k的样本中第i维特征的值是x_i的样本个数，α是平滑值。

当$\alpha=1$时，称作 Laplace 平滑；当$0<\alpha<1$时，称作 Lidstone 平滑；$\alpha=0$时不做平滑。如果不做平滑，当某一维特征的值x_i没在训练样本中出现过时，会

导致$P(x_i|y_k)=0$，从而导致后验概率为 0。加上平滑就可以克服这个问题。

当特征是连续变量时，运用多项式模型就会导致在不做平滑的情况下很多$P(x_i|y_k)=0$，即使做平滑，所得到的条件概率也难以描述真实情况。所以处理连续的特征变量，应该采用高斯模型。

高斯模型假设每一维特征都服从高斯分布（正态分布）：

$$P(x_i|y_k)=\frac{1}{\sqrt{2\pi\sigma_{y_k,i}^2}}e^{-\frac{(x_i-\mu_{y_k,i})^2}{2\sigma_{y_k,i}^2}}$$

其中$\mu_{y_k,i}$表示类别为y_k的样本中第i维特征的均值；$\sigma_{y_k,i}^2$表示类别为y_k的样本中第i维特征的方差。

2.2 多项式模型

当特征是离散变量时，使用多项式模型。多项式模型在计算先验概率$P(y_k)$和条件概率$P(x_i|y_k)$时，会做一些平滑处理，具体公式为：

$$P(y_k)=\frac{N_{y_k}+\alpha}{N+k\alpha}$$

其中，N 是样本总数，k 是类别总数，N_{y_k}是类别为y_k的样本个数，α是平滑值。

$$P(x_i|y_k)=\frac{N_{y_k,x_i}+\alpha}{N_{y_k}+n\alpha}$$

其中，N_{y_k}是类别为y_k的样本个数，n是特征的维数，N_{y_k,x_i}是类别为y_k的样本中第i维特征的值是x_i的样本个数，α是平滑值。

当$\alpha=1$时，称作 Laplace 平滑；当$0<\alpha<1$时，称作 Lidstone 平滑；$\alpha=0$时不做平滑。如果不做平滑，当某一维特征的值x_i没在训练样本中出现过时，会导致$P(x_i|y_k)=0$，从而导致后验概率为 0。加上平滑就可以克服这个问题。

2.3 伯努利模型

伯努利模型也适用于离散特征的情况，但与多项式模型不同的是，伯努利模型中每个特征的取值是布尔型的，即 true 和 false，或是 1 和 0。以文本分类为例，某个单词在文档中出现过，则其特征值为 1（或 true），否则为 0(或 false)。

伯努利模型中，条件概率$P(x_i|y_k)$的计算方式为：

当特征值$x_i = 1$时，$P(x_i|y_k) = P(x_i = 1|y_k)$；

当特征值$x_i = 0$时，$P(x_i|y_k) = 1 - P(x_i = 1|y_k)$。

三、代码实现

3.1 多项式模型实现

3.1.1 提取邮件，并确定测试集和训练集

涉及函数：get_mail、get_training_set

主要内容：从 ham 和 spam 文件夹中分别提取文件名称列表，然后从中随机抽取 10 份作为测试集，剩下的文件作为训练集。同时标注类型：正常文件记为 1，垃圾邮件记为 0。

3.1.2 邮件内容处理

涉及函数：is_alphabet、standardize、divide、delete_stopword

主要内容：依次提取训练集中的文件内容，进行标准化处理，即除去难以处理的无关字符，只留下单词（其中用到了判断是否属于英文字母的函数 is_alphabet）。接着利用结巴分词进行分词，最后对照停用词表，将与语义相关性不大的一般性单词除去。

3.1.3 计算先验概率

涉及函数：get_priori_possibility

主要内容：分别统计出训练集中的正常与垃圾邮件数量，按照多项式模型公式进行计算。另外，使用了拉普拉斯平滑，即 α = 1。

3.1.4 训练模型

涉及函数：get_Nya

主要内容：对于（训练集中）每个文件的每个单词，统计其分别在正常邮件和垃圾邮件中出现的次数，并将情况分别保存到两个字典中。

其中注意在内容处理后的单词列表仍然可能出现重复，所以应设置列表保存已经计算过的单词，防止重复计数造成的结果偏差。

3.1.5 进行测试

涉及函数：filter

主要内容：依次获取测试集中的文件内容，并依照训练集进行数据预处理，得到有意义的单词列表。接着对每个单词（即特征）计算 p(x|yi)，又因为公式分

母为常数，所以直接以 p(x|yi)与 p(y)的乘积作为决定后验概率 p(yi|x)的因素。比较正常邮件状态下和垃圾邮件状态下的乘积，以较大的那一项所属的类别作为测试项的类别。接着与真真实类别比较，记录错误项和错误数，计算出错误概率并输出。

注意在计算过程中，涉及的数字都比较小，所以在设置精度时应该保证不同的数字不会被四舍五入处理成同样数字。另外仍然要进行拉普拉斯平滑。

3.2 伯努利模型实现

3.2.1 收集数据

email 文件夹下有 ham 和 spam 两个文件夹，其中 ham 文件夹下是 25 封正常邮件，spam 文件夹下是 25 封垃圾邮件。

3.2.2 准备数据

数据集中的邮件均为英文文本，因此以非字母、非数字作为符号，使用 split 函数进行切分。将文本中的字符串解析为字符串列表，并将其整理成不重复的词条列表即词汇表。

3.2.3 构建词集模型/词袋模型

根据上面得到的的词汇表构建词集模型：创建一个元素均为 0 的向量，遍历每个词条，若词条存在于词汇表中则置 1，最后返回文档向量；或根据词汇表构建词袋模型：创建一个元素均为 0 的向量，遍历每个词条，若词条存在于词汇表中则计数加 1，最后返回文档向量。

两种模型中选择一种即可，这里选择词集模型。

3.2.4 训练垃圾邮件过滤器

根据训练文档矩阵和训练类别标签向量，条件概率的分母初始化为 2，做 Laplace 平滑，统计属于垃圾邮件类和正常邮件类的条件概率所需的数据，即 P(W0|1),P(W1|1),…,和 P(W0|0),P(W1|0),…,计算正常邮件类的条件概率数组、垃圾邮件类的条件概率数组和文档属于垃圾邮件类的概率，计算时取对数防止向下溢出。

3.2.5 分类

根据已计算出的正常邮件类的条件概率数组、垃圾邮件类的条件概率数组和文档属于垃圾邮件类的概率，将待分类的词条数组分类。分别计算其属于正常邮件类和属于垃圾邮件类的概率，若属于垃圾邮件类的概率大于属于正常邮件类的概率，则判定其为垃圾邮件，返回 1；否则判定其属于正常邮件类，返回 0。

3.2.6 测试垃圾邮件过滤器

读取每个垃圾邮件和每个正常邮件并分别标记 1 和 0。从 50 封邮件中随机选择 40 封作为训练集，10 封作为测试集。经训练后使用垃圾邮件过滤器将 10 封测试集的邮件进行分类。计数分类错误的邮件，分类完成后计算错误率。

四、结果分析

4.1 多项式模型结果分析

由于是随机抽取文件作为测试集与训练集，因此每次得到的错误率略有不同。经过多次（十次以上）测试，计算出错误率基本可稳定在 0.05，运行时间稳定在 0.878s。

被误判的邮件则集中在正常邮件 17.txt、23.txt、25.txt 中。其中前面两者的情况是邮件中出现了在垃圾邮件中出现率高的词汇，而后者的情况是其词汇没有出现在训练集中，所以仅仅按照先验概率分了类。

4.2 伯努利模型结果分析

我们进行 100 次分类测试，并对错误率取平均值，得到平均错误率为 4.2%。因为是随机选取的训练集与测试集，所以每次测试的结果不同。

在测试中出现过被误判情况的邮件有：ham 下的 16.txt、17.txt、22.txt、23.txt，spam 下的 6.txt、17.txt、25.txt。其中，spam 文件夹下 17.txt 被误判的频率很高。基本是因为其中的很多单词没有出现在训练集中，所以此垃圾邮件容易被误判为正常邮件。

```
分类错误的测试集:  ['home', 'based', 'business', 'opportunity', 'knocking', 'your', 'door', 'don', 'rude', 'and', 'let', 'this', 'chance
错误率：10.00%
分类错误的测试集:  ['home', 'based', 'business', 'opportunity', 'knocking', 'your', 'door', 'don', 'rude', 'and', 'let', 'this', 'chance
错误率：10.00%
分类错误的测试集:  ['yeah', 'ready', 'may', 'not', 'here', 'because', 'jar', 'jar', 'has', 'plane', 'tickets', 'germany', 'for']
错误率：10.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
分类错误的测试集:  ['home', 'based', 'business', 'opportunity', 'knocking', 'your', 'door', 'don', 'rude', 'and', 'let', 'this', 'chance
错误率：10.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
错误率：0.00%
平均错误率: 4.2 %
```

4.3 性能比较

经过对比可得，SVM 的效果较朴素贝叶斯稍好，但同样与语料预处理、训练集的大小与选择有关。

综合比较来看，伯努利模型的错误率更低，而多项式模型的误判情况更少。但多项式情况下的误判基本是对于正常邮件，在实际应用中可能造成更大损失。

4.4 总结

在本次实验中，虽然语料预处理不是主要内容，但仍然对结果有至关重要的影响，例如在英文语料中，应考虑到大小写转化及词形变换等问题。而训练数据的大小与选择、各项模型及其参数（如平滑参数 α ）的选择对结果也有重要影响，在 实际应用时应多次试验，按照实际情况进行修改，以达到最佳过滤效果。