信 息 内 容 安 全 实 验 报 告

实验项目名称： 利用朴素贝叶斯过滤垃圾邮件

班级： SC011701

姓名： 程瑞淇（2017302241）

李向娟（2017302235）

王燕颉（2017302234）

曹至杰（2017302233）

成员分工： 李向娟：理论资料查找和整理

讨论代码

曹至杰：资料查找、讨论代码

程瑞淇：代码编写和修改

王燕颉：报告撰写、讨论代码

指导教师： 杨黎斌

实验时间： 2020.3.3

目录

利用朴素贝叶斯筛选垃圾邮件

1、实验目的

①掌握利用朴素贝叶斯筛选垃圾邮件的方法；

②掌握网络舆情系统设计中常用的情感分析方法；

③能够设计合理的指标对性能进行衡量。

2、理论依据

2.1 贝叶斯定理

通常，事件 A 在事件 B(发生)的条件下的概率，与事件 B 在事件 A 的条件下的概率是不一样的；然而，这两者是有确定的关系,贝叶斯定理就是这种关系的陈述。

贝叶斯法则是关于随机事件 A 和 B 的条件概率和边缘概率的。

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

其中 P(A|B)是在 B 发生的情况下 A 发生的可能性。

$$A_1, \cdots, A_n$$

为完备事件组，即

$$\cup_{i=1}^{n} A_i = \Omega, A_i A_j = \phi, P(A_i) > 0.$$

在贝叶斯法则中，每个名词的含义：

P(A)是 A 的先验概率或边缘概率。之所以称为"先验"是因为它不考虑任何 B 方面的因素。

P(A|B)是已知 B 发生后 A 的条件概率,也由于得自 B 的取值而被称作 A 的后验概率。

P(B|A)是已知 A 发生后 B 的条件概率，也由于得自 A 的取值而被称作 B 的后验概率。

P(B)是 B 的先验概率或边缘概率，也作标准化常量（normalized constant）。

按这些术语，Bayes 法则可表述为：

后验概率 = (似然度 * 先验概率)/标准化常量　也就是说，后验概率与先验概率和似然度的乘积成正比。

另外，比例 P(B|A)/P(B)也有时被称作标准似然度，Bayes 法则可表述为：

后验概率 = 标准似然度 * 先验概率。

2.2 拉普拉斯平滑处理

零概率问题即在计算实例的概率时，如果某个量 x，在观察样本库（训练集）中没有出现过，会导致整个实例的概率结果是 0。在文本分类的问题中，当一个词语没有在训练样本中出现，该词语调概率为 0，使用连乘计算文本出现概率时也为 0。这是不合理的，不能因为一个事件没有观察到就武断的认为该事件的概率是 0。

为了解决零概率的问题，法国数学家拉普拉斯最早提出用加 1 的方法估计没有出现过的现象的概率，所以加法平滑也叫做拉普拉斯平滑。假定训练样本很大时，每个分量 x 的计数加 1 造成的估计概率变化可以忽略不计，但可以方便有效的避免零概率问题。

2.3 贝叶斯过滤器的使用过程

贝叶斯过滤器是一种统计学过滤器，建立在已有的统计结果之上。在使用之前对过滤器进行训练并得到初步统计结果。对于收到的一封邮件，在未经统计分析之前，假定它是垃圾邮件的概率为50%。然后对这封邮件进行解析，发现其中包含了A这个词，则计算它是垃圾邮件的概率。我们用W表示"A"这个词，那么问题就变成了如何计算P(C|W)的值，即在某个词语（W）已经存在的条件下，垃圾邮件（C）的概率有多大。根据条件概率公式，马上可以写出P(C|W)=P(W|C)P(C)/P(W)=P(W|C)P(C)/(P(W|H)P(H)/P(H|W))。

公式中，P(W|C)和P(W|H)的含义是，这个词语在垃圾邮件和正常邮件中，分别出现的概率。这两个值可以从历史资料库中得到。另外，P(C)和P(H)的值，前面说过都等于50%。所以，马上可以计算P(C|W)的值。

2.4 ID3决策树算法

决策树是一种依托决策而建立起来的一种树。在机器学习中，决策树是一种预测模型，代表的是一种对象属性与对象值之间的一种映射关系，每一个节点代表某个对象，树中的每一个分叉路径代表某个可能的属性值，而每一个叶子节点则对应从根节点到该叶子节点所经历的路径所表示的对象的值。决策树仅有单一输出，如果有多个输出，可以分别建立独立的决策树以处理不同的输出。

ID3算法流程：

输入：训练数据集D，特征集A，阈值ε

输出：决策树T

(1)若 D 中所有实例属于同-类 Ck ,则 T 为单结点树,并将类 Ck 作为该结点的类标记,返回 T ;

(2)若 A=中,则 T 为单结点树,并将 D 中实例数最大类 Ck 作为该结点类标记,返回 T ;

(3)否则,计算 A 中各特征对 D 的信息增益,选择信息增益最大的特征 Ag ;

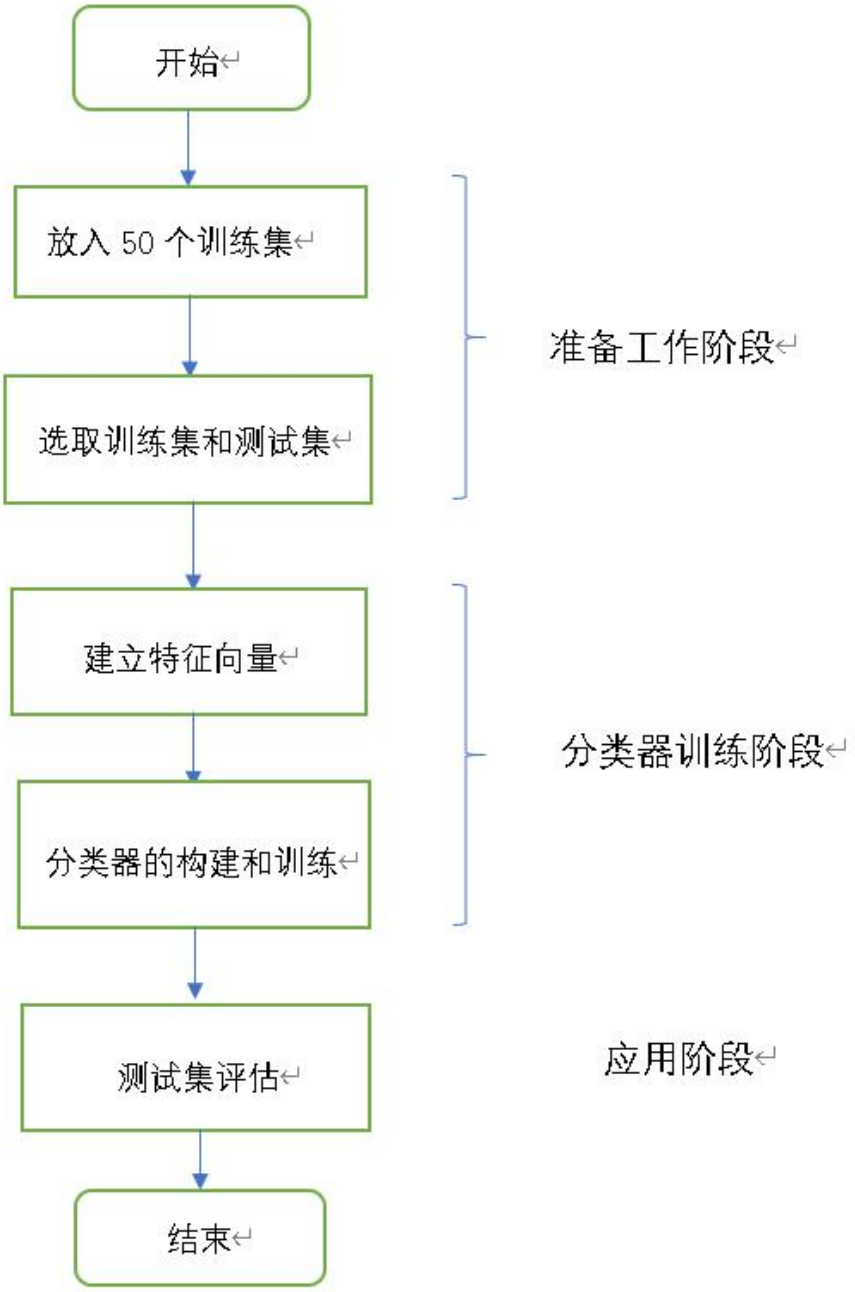
(4)如果 Ag 的信息增益小于阈值 ε ,则置 T 为单结点树,并将 D 中样本数最大的类 Ck 作为该结点的类标记,返回 T ;

(5)否则,对 Ag 的每一个可能值 ai ,分割 D 为若干非空子集 Di ,将 Di 中实例数最大的类作为标记,构建子结点,由结点及其子结点构成树 T ,返回 T ;

(6)对第 i 个子结点,以 Di 为训练集, A-{Ag}为特征集,递归的调用第(1)~(5)步,得到子树 Ti,返回 Ti。

3、代码实现过程

3.1 系统设计流程图



整个朴素贝叶斯分类分为三个阶段：

第一阶段——准备工作阶段，这个阶段的任务是为朴素贝叶斯分类做必要的准备，主要工作是根据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。这一阶段的输入是所有待分类数据，输出是特征属

性和训练样本。这一阶段是整个朴素贝叶斯分类中唯一需要人工完成的阶段，其质量对整个过程将有重要影响，分类器的质量很大程度上由特征属性、特征属性划分及训练样本质量决定。

第二阶段——分类器训练阶段，这个阶段的任务就是生成分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。其输入是特征属性和训练样本，输出是分类器。这一阶段是机械性阶段，根据前面讨论的公式可以由程序自动计算完成。

第三阶段——应用阶段。这个阶段的任务是使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。这一阶段也是机械性阶段，由程序完成。

3.2 数据集收集和训练集、测试集的选取

代码中引用的 email 文件夹分为两个子文件，ham 文件夹保存非垃圾邮件，spam 文件夹保存垃圾邮件。

3.3 训练集和测试集选取

随机选出 5 篇训练集（10%）和 45 篇测试集。对文件进行预处理：因为邮件都为英文文本，因此用非单词字符划分；除去重复单词和长度小于 2 的单词。

3.4 统计词汇数量

建立特征向量，对每一封邮件中的单词出现的次数进行统计。单词计算初值为 0，每出现一次就加 1. 每个测试集的邮件都要构建特征

向量，以列表形式返回。

3.5 分类器的构建和训练

用训练集中的文档进行分类器训练。统计训练集中垃圾和非垃圾邮件的词汇数量的特征向量。计算训练集中单词的词汇频率 P（Ci）并计算垃圾邮件的比率。用拉普拉斯平滑解决零概率问题。

3.6 测试集的评估

判断测试集是否为垃圾邮件，只需对每个邮件判断 p(c0|w)(不是垃圾邮件的概率)与 p(c1|w)（是垃圾邮件的概率）。

如果 p(c0|w)＞p(c1|w)，那么该邮件为非垃圾邮件。

如果 p(c0|w)＜p(c1|w)，那么该邮件为垃圾邮件。

p(ci|w)(i=0 或 1)的计算利用贝叶斯公式通过 p(w|ci)与 p(ci)的计算得出。p(w|ci)为在垃圾邮件（或非垃圾邮件）中的全体向量特征（单词向量特征）出现的概率，p(ci)为训练集中垃圾邮件（或非垃圾邮件）的概率。

4、实验结果及分析

```
分类错误的测试集:  ['yeah', 'ready', 'may', 'not', 'here', 'because', 'jar', 'jar', 'has', 'plane', 'tickets', 'germany', 'for']
错误率为: 0.2

Process finished with exit code 0
```

```
错误率为: 0.0

Process finished with exit code 0
```

因为训练集只有 10%，错误率在 0~20%之间。