

网络空间先进技术研究院

**人工智能与机器学习**

**课程报告**

项目名称 基于朴素贝叶斯的文本分类

学 院 计算机科学与教育软件学院

专业班级 计算机技术

学生姓名 张川京

学生学号 2111806068

任课教师 仇 晶

2018 年 11月

目录

1 项目目的 3

2 项目环境和条件 3

3 项目原理 3

3.1贝叶斯定义 3

3.2贝叶斯决策论 4

3.3朴素贝叶斯分类器 4

4 项目内容 6

4.1朴素贝叶斯的深度理解 6

4.2朴素贝叶斯在文本分类的应用 6

4.3 python中jieba库的使用及理解 6

4.4 python中sklearn库的使用及理解 7

5 项目过程与内容 7

5.1任务分析 7

5.2数据分析 7

5.3开发步骤 8

5.4关键问题 8

5.5实验结果分析 9

6 项目总结与心得体会 9

6.1 项目总结 9

6.2 心得体会 10

7 参考文献 11

# 项目目的

学习朴素贝叶斯算法的原理，熟悉机器学习中的朴素贝叶斯算法及其在文本分类任务上的应用，在文本分类任务中，又涉及到中文文本的分词、去除停用词、文本的向量表示等文本预处理工作，在文本的向量表示中，最基础的词袋模型、pca降维、TF/IDF等算法都逐一了解，并在实验中尝试，测试实验效果。

# 项目环境和条件

Windows10、python3.7、CPU（i5）

在实验过程中使用了几个成熟的python库，jieba、skelarn、pyplot

# 项目原理

## 3.1贝叶斯定义

在概率论与统计学中，贝叶斯定理（或称贝叶斯法则、贝叶斯规则）描述了一个事件的可能性，这个可能性是基于了预先对于一些与该事件相关的情况的知识。举例来说，如果癌症和年龄有关，那么使用贝叶斯定理的话，相比根本不了解关于此人的任何其他信息，知道了它的年龄的话就可以用来更准确地帮助评估它得癌症与否的概率。

那么其实很明显了，这里的“可能性”也是考虑了与随机事件相关的因素的，所以贝叶斯定理所阐述的也就是后验概率的获得方法。

用数学公式来表述贝叶斯定理：

c表示的是随机事件发生的一种情况。x表示的就是证据（evidence）\状况(condition)，泛指与随机事件相关的因素。

P(c|x)：在x的条件下，随机事件出现c情况的概率。（后验概率）

P(c)：（不考虑相关因素）随机事件出现c情况的概率。（先验概率）

P(x|c)：在已知事件出现c情况的条件下，条件x出现的概率。（后验概率）

P(x)：x出现的概率。（先验概率）

那么，落实到实际的问题当中，我们想获得的核心结果其实也就是P(c|x)，即我们想知道，在考虑了一些现有的因素后，随机事件会以多大概率出现各种情况，通过参考这个结果，我们针对性地作出决策。

而从计算上来说，我们需要同时知道P(c),P(x|c)和P(x)才能算出目标值P(c|x)，而P(x)由于是c无关，而且作为共同的分母，在我们计算c的各种取值的可能性时并不会对各结果的相对大小产生影响。因此可以忽略。

## 3.2贝叶斯决策论

贝叶斯决策理论研究了模式类的的概率结构完全知道的情况。贝叶斯决策理论是解决模式分类问题的一种基本统计途径，假设了决策问题可以用概率的形式来描述并且假设所有有关的概率结构均已知。

假设一个待识别的物理对象用其d个属性观察值描述，称之为d个特征，这组成一个d维的特征向量，而这d维待征所有可能的取值范围则组成了一个d维的特征空间。由于属于不同类的待识别对象存在着呈现相同观察值的可能，这种可能性可用P(ωi|x)表示。如何作出合理的分类就是贝叶斯决策理论所要讨论的问题。

已知总共有c类物体，也就是说待识别物体属于这c类中的一个类别。 把这c类具体为ω1,ω2.....ωc ，先验概率P(ωi)及类条件概率密度函数p(x|ωi)已知的条件下，可以求出后验概率P(ωi|x)。

基于最小错误率的贝叶斯决策论是根据后验概率的大小来的确定待分类对象的类别的，但是有时候将某一类的判别出错会导致很大的风险，比如将癌细胞判别为正常细胞，会扩大癌细胞扩散的风险，这对病人是非常危险的。所以人们又提出了最小风险的贝叶斯决策论，给每一类赋予风险权重，将风险权重与后验慨率做乘积得到风险，在判别时将待分类对象分类给误判风险最小的类别。

## 3.3朴素贝叶斯分类器

贝叶斯分类器是一类分类算法的总称，这类算法均以贝叶斯定理为理论基础。贝叶斯分类器的分类原理是通过先验概率，利用贝叶斯公式计算出后验概率，选择最大后验概率所对应的分类结果。

贝叶斯公式为：。其中，P(c)是先验概率，P(x|c)样本x相对于；类标记c的类条件概率，或称为似然（likelihood）；P(x)是用于归一化的证据因子。对于给定样本x，证据因子与类标记无关，则估计P(c|x)可转换为基于训练数据D来估计鲜艳P(c)和似然P(x|c)。类先验概率P(c)可通过各类样本出现的频率来进行估计。对于类条件概率P(x|c)，一种常用策略是先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布的参数进行估计。

对于朴素贝叶斯分类器，朴素（naive）的含义是各个特征属性之间是相互独立的。例如，在计算p(w|ci)时，我们将特征向量w展开为独立子特征，则转化为，这里我们有假设所有特征都独立，即可以使用以下公式来计算，这就是利用了朴素的原则。（wi和下文的xi是相同含义，表示各个特征属性分量）。

基于以上，朴素贝叶斯分类器的训练过程就是基于训练集D来估计类先验概率P(c)，并为每个特征属性估计条件概率。对于类先6验概率，在有足够独立同分布训练样本的条件下，通过计算各类样本占总样本数的比例来计算。计算条件概率时，对于离散属性而言，，对于连续属性考虑概率密度函数，式中的两个未知参数均值和方差极大似然法估计得到。假设概率密度函数，则参数的极大似然估计为：

这就是说，通过极大似然法得到的正态分布均值就是样本均值，方差就是的均值。

（4）拉普拉斯修正

为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“平滑”（smoothing），常用拉普拉斯修正。具体来说，令N表示训练集D中可能的类别数，Ni表示第i 个属性可能的取值数，修正为 和。

拉普拉斯修正避免了因训练集样本不充分而导致概率估值为零的问题，并且在训练集变大时，修正过程所引入的先验的影响也会逐渐变得可忽略，使得估值逐渐趋向于实际的概率值。

# 项目内容

## 4.1朴素贝叶斯的深度理解

朴素贝叶斯不同于其他SVM，神经网络等机器学习方法，准确来讲朴素贝叶斯并没有训练出一个“模型”，新加入训练样本也只需要在计算概率时加上新加入的样本数据，也并不需要重新训练模型参数，而且朴素贝叶斯的有效训练样本越多，他的预测结果就会越准确。

## 4.2朴素贝叶斯在文本分类的应用

文本分类的第一步需要进行文本的预处理，使用jieba分词之后，为了提高准确率，我们首先根据停用词表、单词长度、是否数字等来进行一遍过滤，再使用TF/IDF来过滤掉不具有代表性的词（即在所有文章中出现频率较高的词）。最终转化为向量表示之后，可以看出文本向量都是稀疏的离散属性。

我们使用的sklern库中的朴素贝叶斯有三种模型：高斯模型、多项式模型、伯努利模型，由于文本分类中，分词之后的结果是离散的数据表示，所以采用多项式模型用来做文本分类最为合适。

## 4.3 python中jieba库的使用及理解

使用jieba分词，支持三种分词模式：精确模式，试图将句子最精确地切开，适合文本分析；全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。在本次实验中我们使用精确模式。

## 4.4 python中sklearn库的使用及理解

全称scikit-learn，是一个开源的机器学习库，机器学习的常见流程为：获取数据、数据预处理、训练模型、模型评估、预测，分类。在sklearn库中也几乎对应这些流程都有常用的方法，主要包括机器学习类型、数据集、效果评估三大类，其中机器学习类型又按照监督学习、半监督学习、无监督学习继续分类，而数据集则按照训练集、测试集、验证集以及交叉验证划分四类，效果评估中主要按照监督学习和无监督学习划分。熟悉sklearn库中的方法对于学习和使用机器学习是非常有必要的。

# 项目过程与内容

## 5.1任务分析

给的文本资料data文件夹下分为文学、体育、校园、女性四个文件夹，每个文件夹下面有若干对应主题的中文txt文本文件，大概数量在1000左右，等于每个txt文件是一个data，而文件夹名字就是对应的label。由于数据量不大，应该不用做降维处理也可以，但是为了熟悉TF/IDF或者PCA等降温方法，我们尝试一下降维处理。之后要依次读取txt文件，将文件夹名作为label，处理为label-data的数据格式之后划分训练集测试集，再进行文本的向量化表示之后就可以进行朴素贝叶斯的训练、测试，查看训练效果。

## 5.2数据分析

每个txt文件大小不一，包含若干标点符号（， 。 “ ” -），数字等，所以必须在文本预处理过程中去除停用词、标点符号、数字等来提高分词的有效性，进而提高朴素贝叶斯的准确率。在读取数据过程中，因为同一主题的文本文件放在一个文件夹下面，可以遍历该文件夹下的所有文本文件，并同一使用文件夹名字作为label。

## 5.3开发步骤

本次开发流程大致为读数据、分词、去停用词、建立词袋、文本向量化表示（one-hot、TF/IDF）、划分测试/训练集、训练朴素贝叶斯、测试、画折线图。

首先读数据时二层遍历，首先遍历data文件夹，有四个文件夹，然后遍历每个文件夹下的所有txt文件，读取txt文件作为最初始的data，对应的文件夹名字作为data的label属性，最终得到一个label\_data\_list，然后使用jieba库进行分词，并根据停用词表、是否数字、长度是否小于5过滤一遍，分别统计词频跟one-hot表示的方法得到两种文本向量表示方式，首先尝试使用PCA对one-hot的文本向量进行降维，由于降维之后出现了负数的情况，在MultinomialNB中无法使用，所以考虑使用TF/IDF将不具有代表性的单词过滤掉（这里过滤掉了在超过40%的文本中都出现过的单词），然后再讲文本用过滤之后的词袋向量表示。再将 label\_data\_list使用shuffle打乱之后固定前500条为测试集，训练集从1000逐步增加到2500，步长50，依次训练sklearn.naive\_bayes的MultinomialNB模型并测试，由于测试集固定，所以可以实验结果有效的反映了朴素贝叶斯模型的准确率随训练集大小的变化情况。最后使用pyplot画折线图表示朴素贝叶斯预测准确率随训练集大小的变化趋势，看是否与朴素贝叶斯的算法原理相吻合。

## 5.4关键问题

1. 要使用utf-8模式，在读文件是可以使用encode/decode函数
2. 读取到的txt文件是data部分，每读取一个txt文件要添加文件夹名作为label属性值建立对应的class\_list
3. 初始的词袋模型有22000多维，这里使用pca降维之后出现非负的情况，无法使用MultinomialNB.fit()来训练，最后采用tf/idf，并将在超过50%的文章中出现的单词都过滤掉。
4. 考虑到朴素贝叶斯算法的特殊性，我们采取固定测试集，逐渐增加训练集的办法来看看实验效果是否与理论一致。

## 5.5实验结果分析

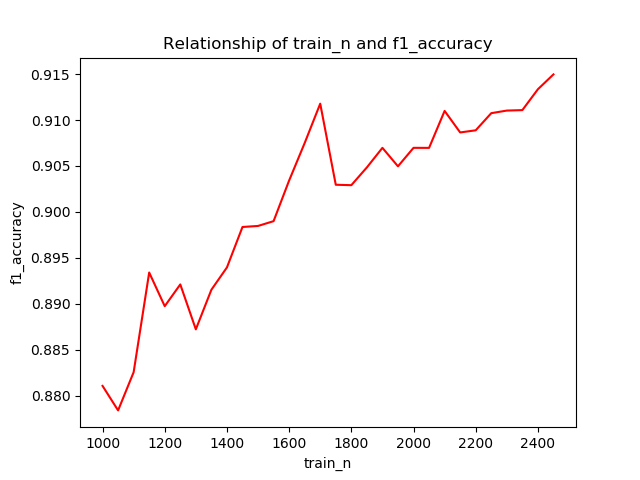


图1

如图1所示，f1评分随着训练集增长呈上升趋势，这与朴素贝叶斯的算法原理是一致的，有效的训练样本越多，朴素贝叶斯的准确率也就越高。由于朴素贝叶斯相对简单，提升性能的关键就落到了文本处理这里，除了要做高效的分词、去停用词外，采取合适的降维手段、调整降维的参数就非常重要，可以逐步寻找最优的降维方法。

# 项目总结与心得体会

## 6.1 项目总结

我在项目中主要负责代码编写工作，原理上看起来很简单的朴素贝叶斯算法，但是在真正落实到工程的时候，也出现过各种细节问题。依次读取的label跟data要一一对应，zip之后才可以shuffle打乱排序，utf-8编码问题，尝试不同的向量表示（0/1表示，词频表示），尝试不同的降维方式（课程演示的时候按词频降序选取1000，后来尝试过TF/IDF，PCA，PCA降维的时候出现负数情况MultinomialNBC.fit()报错），测试准确率的时候也用了accurac跟F1。最终可能由于文本量还是比较少，降维手段仅仅作为一种尝试跟学习，实际效果并没有太大提升。最终选择展示的折线图是F1评分与训练集大小的对应关系，由于朴素贝叶斯的特殊性（简单高效，效果随训练集增加而提升），想验证是否符合这种算法原理。

## 6.2 心得体会

在整个项目流程中，完整的体验了从了解一种算法到落实到具体应用并实现的过程，并且跑通之后后期也尝试了很多不同的细节技术，学习到了很多python中机器学习相关的知识。朴素贝叶斯为什么能够取得这么好的效果呢？分词之后他认为每个单词都是独立的，并没有考虑上下文的语义关系，但是仍然能够取得很好的分类效果，查了相关文章之后，大致可以用奥卡姆剃刀来解释，还有一个文章提出一种解释：有些独立假设在各个分类之间的分布都是均匀的所以对于似然的相对大小不产生影响；即便不是如此，也有很大的可能性各个独立假设所产生的消极影响和积极影响互相抵消，最终导致结果收到的影响不大。而且在使用朴素贝叶斯的过程中，由于算法的简单性，工作重心反而落在了文本预处理的工作中，也了解了一下中文分词的技术，jieba库的使用等等，后续还需要继续加强学习，最好是多动手操作。

# 参考文献

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2017, 147-156

[2] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2015, 47-53

[3] 蒋良孝. 朴素贝叶斯分类器及其改进算法研究[博士学位论文]. 《中国地质大学》，2009年

[4] 周颜军, 王双成, 王辉. 基于贝叶斯网络的分类器研究[J]. 东北师大学报(自然科学版), 2003年02期

[5] 高岩. 朴素贝叶斯分类器的改进研究[D]. 华南理工大学, 2011年

[6] 杜瑞杰. 贝叶斯分类器及其应用研究[博士学位论文].. 上海大学, 2012年

[7] 李静梅, 孙丽华, 张巧荣, 等. 一种文本处理中的朴素贝叶斯分类器[J],《哈尔滨工程大学学报》, 2003年01期

[8] 陈景年.一种用于贝叶斯分类器的文本特征选择方法[J], 计算机工程与应用, 2008, 44(13): 24-26

[9] 朴素贝叶斯分类器(2018-10-01).[Online]. Available URL: https://blog.csdn.net/sinat\_36246371/article/details/60140664

[10] 《机器学习实战》学习总结（三）朴素贝叶斯在文本分类的应用(2018-10-01).[Online]. Available URL: https://blog.csdn.net/wiborgite/article/details/79232647

[11] 朴素贝叶斯算法原理小结(2018-10-01).[Online]. Available URL: https://www.cnblogs.com/pinard/p/6069267.html

[12] 朴素贝叶斯分类器(2018-10-01).[Online]. Available URL: https://zh.wikipedia.org/wiki/%E6%9C%B4%E7%B4%A0%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%88%86%E7%B1%BB%E5%99%A8

[13] 贝叶斯(Bayes)决策理论(2018-10-02).[Online]. Available URL: https://blog.csdn.net/songzitea/article/details/23131609

[14] 机器学习 —— 基础整理（一）贝叶斯决策论；二次判别函数；贝叶斯错误率；生成式模型的参数方法(2018-10-02).[Online]. Available URL: https://www.cnblogs.com/Determined22/p/6347778.html

**课程报告评定表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学生姓名 |  | 学 号 |  | 成绩 |  |
| 专业班级 |  | | | | |
| 项目名称 |  | | | | |
| 指  导  教  师  评  语 | 指导教师：  年 月 日 | | | | |