机器学习 Project 2 文本分类

--10215501435 杨茜雅

摘要：

文本分类是一个机器学习领域经典的话题。朴素贝叶斯(Naive Bayes) 是一种构建分类器的机器学习算法，朴素指该分类算法假定样本每个特征与其他特征都不相关。朴素贝叶斯分类器的一个优势在于只需要根据少量的训练数据估计出必要的参数，且计算量相较其他机器学习算法而言较小。本实验所采用的数据集为卡内基·梅隆大学 Text Learning Group 的 20newsgroup 数据集，共涉及 20 个网络新闻话题。本实验采用三种不同分布朴素贝叶斯模型，通过整理数据、探索文本数据、文本特征提取、划分训练集测试集、归一化、构建模型来解决文本分类的问题。先使用三种不同分布的朴素贝叶斯模型先进行预测，分别是 Multinomial，Complement，Bernuolli，分别在这三种模型下通过五重交叉验证，选择最优超参数，并且对一些重要指标与分类精度，以及交叉验证结果（误差对比）进行了可视化展示。最后用 complement 最好的模型参数来看最后的结果，达到了 90%的准确率。

关键词：文本分类，朴素贝叶斯，向量化，误差曲线

|

数据集:

http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html

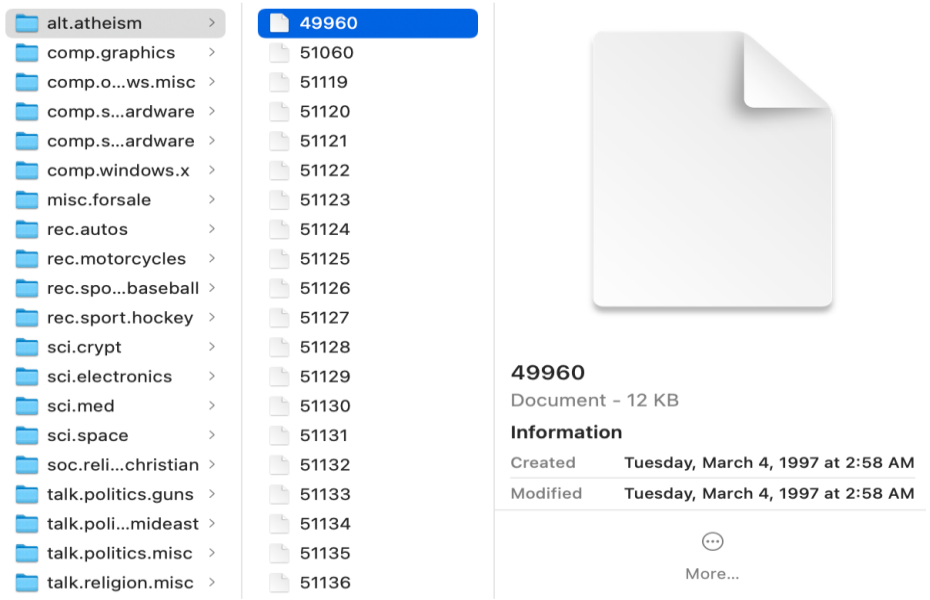
实验任务:

20000 个文档分成 20 类，五重交叉验证结果，不要使用网站上的代码

本实验的目标在于使用朴素贝叶斯分类器对 20 newsgroups 数据集进行文本分类，并使用交叉验证方法找到最优超参数，再进行后续分析。

一、20 newsgroups 数据集

本次实验所用的数据集为卡内基·梅隆大学 Text Learning Group 的 20newsgroup 数据集，数据集内包含了 20 个网络新闻话题文档，每个文档里面有 1000 个项目，所以总共是 20000 个文档。由于下载下来是文本的形式，并不是现成的.csv 的格式，需要进行整理。



每份新闻数据都包含长短不一的文本内容，如下所示，这个示例新闻数据是关于枪支控制的政治类新闻。

53293

```
Xref: cantaloupe.srv.cs.cmu.edu misc.headlines:41568 talk.politics.guns:53293
Newsgroups: misc.headlines,talk.politics.guns
Path: cantaloupe.srv.cs.cmu.edu!rochester!udel!darwin.sura.net!wupost!uunet!murphy!jpradley!magpie!manes
From: manes@magpie.linknet.com (Steve Manes)
Subject: Re: Gun Control (was Re: We're Mad as Hell at the TV News)
Organization: Manes and Associates, NYC
Distribution: na
Date: Thu, 1 Apr 1993 19:56:03 GMT
Message-ID: <C4tM1H.ECF@magpie.linknet.com>
Followup-To: misc.headlines,talk.politics.guns
X-Newsreader: TIN [version 1.1 PL9]
References: <1993Mar31.211339.29232@synapse.bms.com>
Lines: 38

hambidge@bms.com wrote:
: In article <C4psoG.C6@magpie.linknet.com>, manes@magpie.linknet.com (Steve Manes) writes:

: >: Rate := per capita rate.  The UK is more dangerous.
: >: Though you may be less likely to be killed by a handgun, the average
: >: individual citizen in the UK is twice as likely to be killed
: >: by whatever means as the average Swiss.  Would you feel any better
: >: about being killed by means other than a handgun? I wouldn't.
:
: >What an absurd argument.  Switzerland is one-fifth the size of the
: >UK with one-eigth as many people therefore at any given point on
: >Swiss soil you are more likely to be crow bait.  More importantly,
: >you are 4x as likely to be killed by the next stranger approaching
: >you on a Swiss street than in the UK.

: You are betraying your lack of understanding about RATE versus TOTAL
: NUMBER. Rates are expressed, often, as #/100,000 population.
: Therefore, if a place had 10 deaths and a population of 100,000, the
: rate would be 10/100,000.  A place that had 50 deaths and a population
: of 1,000,000 would hav a rate of 5/100,000.  The former has a higher
: rate, the latter a higher total.  You are less likely to die in the
```

经过整理后的数据全貌：

In [8]:
```
df
```

Out[8]:

	Unnamed: 0	data	target
0	0	Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:49...	alt.atheism
1	1	Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:51...	alt.atheism
2	2	Newsgroups: alt.atheism\nPath: cantaloupe.srv....	alt.atheism
3	3	Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:51...	alt.atheism
4	4	Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:51...	alt.atheism
...	...	...	...
19992	19992	Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:54...	talk.religion.misc
19993	19993	Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:54...	talk.religion.misc
19994	19994	Xref: cantaloupe.srv.cs.cmu.edu talk.religion....	talk.religion.misc
19995	19995	Xref: cantaloupe.srv.cs.cmu.edu talk.religion....	talk.religion.misc
19996	19996	Xref: cantaloupe.srv.cs.cmu.edu talk.abortion:...	talk.religion.misc

19997 rows × 3 columns

数据集有 19997 行，3 列，每行表示一个新闻，第一列不需要，第二列是 data，对应的新闻的文本内容，第三列是新闻所属的类别。

查看第一条新闻的内容:

```
In [9]: df['data'][0]
```

```
Out[9]: 'Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:49960 alt.atheism.moderated:713 news.answers:7054 alt.answers:126\nPath: cantaloupe.srv.cs.cmu.edu!crabapple.srv.cs.cmu.edu!bb3.andrew.cmu.edu!news.sei.cmu.edu!cis.ohio-state.edu!magnus.acs.ohio-state.edu!usenet.ins.cwru.edu!agate!spool.mu.edu!uunet!pipex!ibmpcug!mantis!mathew\nFrom: mathew <mathew@mantis.co.uk>\nNewsgroups: alt.atheism,alt.atheism.moderated,news.answers,alt.answers\nSubject: Alt.Atheism FAQ: Atheist Resources\nSummary: Books, addresses, music -- anything related to atheism\nKeywords: FAQ, atheism, books, music, fiction, addresses, contacts\nMessage-ID: <19930329115719@mantis.co.uk>\nDate: Mon, 29 Mar 1993 11:57:19 GMT\nExpires: Thu, 29 Apr 1993 11:57:19 GMT\nFollowup-To: alt.atheism\nDistribution: world\nOrganization: Mantis Consultants, Cambridge. UK.\nApproved: news-answers-request@mit.edu\nSupersedes: <19930301143317@mantis.co.uk>\nLines: 290\n\nArchive-name: atheism/resources\nAlt-atheism-archive-name: resources\nLast-modified: 11 December 1992\nVersion: 1.0\n\n                                Atheist Resources\n\n                Addresses of Atheist Organizations\n\n                                        USA\n\nFREEDOM FROM RELIGION FOUNDATION\n\nDarwin fish bumper stickers and assorted other atheist paraphernalia are\navailable from the Freedom From Religion Foundation in the US.\n\nWrite to:  FFRF, P.O. Box 750, Madison, WI 53701.\nTelephone: (608) 256-8900\n\nEVOLUTION DESIGNS\n\nEvolution Designs sell the "Darwin fish".  It\'s a fish symbol, like the ones\nChristians stick on their cars, but with feet and the word "Darwin" written\ninside.  The deluxe moulded 3D plastic fish is $4.95 postpaid in the US.\n\nWrite to:  Evolution Designs, 7119 Laurel Canyon #4, North Hollywood,\n          CA 91605.\n\nPeople in the San Francisco Bay area can get Darwin Fish from Lynn Gold --\ntry mailing <figmo@netcom.com>.  For net people who go to Lynn directly, the\nprice is $4.95 per fish.\n\nAMERICAN ATHEIST PRESS\n\nAAP publish various atheist books -- critiques of the Bible, lists of\nBiblical contradictions, and so on.  One such book is:\n\n"The Bible Handbook" by W.P. Ball and G.W. Foote.  American Atheist Press.\n372 pp.  ISBN 0-910309-26-4, 2nd edition, 1986.  Bible contradictions,\nabsurdities, atrocities, immoralities... contains Ball, Foote: "The Bible\nContradicts Itself", AAP.  Based on the King James version of the Bible.\n\nWrite to:  American Atheist Press, P.O. Box 140195, Austin, TX 787
```

查看一共 20 个类别:

```
In [10]: np.unique(df.target)
```

```
Out[10]: array(['alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc',
                'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware',
                'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycles',
                'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt',
                'sci.electronics', 'sci.med', 'sci.space',
                'soc.religion.christian', 'talk.politics.guns',
                'talk.politics.mideast', 'talk.politics.misc',
                'talk.religion.misc'], dtype=object)
```

接下来看样本均不均衡，这是要代入朴素贝叶斯模型之前需要关注的:

```
In [11]: for i in ['alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc',
                   'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware',
                   'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycles',
                   'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt',
                   'sci.electronics', 'sci.med', 'sci.space',
                   'soc.religion.christian', 'talk.politics.guns',
                   'talk.politics.mideast', 'talk.politics.misc',
                   'talk.religion.misc']:
             print(i,(df.target==i).sum()/len(df.target))
```

```
alt.atheism 0.050007501125168774
comp.graphics 0.050007501125168774
comp.os.ms-windows.misc 0.050007501125168774
comp.sys.ibm.pc.hardware 0.050007501125168774
comp.sys.mac.hardware 0.050007501125168774
comp.windows.x 0.050007501125168774
misc.forsale 0.050007501125168774
rec.autos 0.050007501125168774
rec.motorcycles 0.050007501125168774
rec.sport.baseball 0.050007501125168774
rec.sport.hockey 0.050007501125168774
sci.crypt 0.050007501125168774
sci.electronics 0.050007501125168774
sci.med 0.050007501125168774
sci.space 0.050007501125168774
soc.religion.christian 0.049857478621793266
talk.politics.guns 0.050007501125168774
talk.politics.mideast 0.050007501125168774
talk.politics.misc 0.050007501125168774
talk.religion.misc 0.050007501125168774
```

可以看到总体是比较均衡的。

二、文本特征提取

使用 **TF-IDF** 向量计数

在开始分类之前，必须先将文本编码成数字，一般常用的方法是单词计数向量计数和 TF-IDF 向量计数方法。

1、 单词计数向量，在这种技术中，一个样本可以包含一段话或者一篇文章，这个样本如果出现了 10 个单词，就会有 10 个特征，每个特征代表一个单词，特征的取值代表这个单词在这个样本中总共出现了几次，是一个离散的，代表次数的整数。在 sklearn 中，单词计数向量计数可以通过 feature_extraction.text 模块中的 CountVectorizer 类实现。

2、 可以预见，如果使用单词计数向量，可能会导致一部分常用词频繁出现在矩阵中并且占有很高的权重，对分类来说，这明显是对算法的一种误导。为了解决这个问题，比起使用次数，使用单词在句子中所占的比例来编码单词，这就是 TF-IDF 方法，词频逆文档频率，是通过单词在文档中出现的频率来衡量其权重，IDF 的大小与一个词的常见程度成反比，这个词越常见，编码后为它设置的权重就会倾向于越小，从而来压制频繁出现的一些无意义的词。在 sklearn 中，使用 feature_extraction.text 中类 TfidVectorizer 来执行这种编码

在实验初期，两种方法都用了，但是 TF-IDF 向量计数效果要好一些，所以就使用 TF-IDF 向量计数进行文本特征提取。

三、划分训练集测试集

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=0)
```

四、归一化

在本实验集上，归一化并没有什么实质的效果。

五、构建模型

1、先使用三种不同分布的朴素贝叶斯模型先进行预测，分别是 Multinomial，Complement，Bernuolli

模型介绍：

朴素贝叶斯有各种各样的假设，除了“朴素”的假设即假设变量之间相互独立的假设，还有对于概率分布的假设，包括多项式朴素贝叶斯 multinomialNB、伯努利朴素贝叶斯 BernuolliNB（二项分布，数据集中存在多个特征，但每个特征都是二分类的，可以用布尔变量表示，由于本数据集中有 20 个类，那么可以使用类中专门用来二值化的参数 binarize 来改变数据）、高斯朴素贝叶斯和补集朴素贝叶斯。

1、多项式朴素贝叶斯适用于二项分布、多项分布，擅长分类型变量，多项式朴素贝叶斯的特征矩阵经常是稀疏矩阵，所以经常被用于文本分类。

2、伯努利朴素贝叶斯和多项式朴素贝叶斯非常相似，常用于处理文本分类数据，但由于伯努利朴素贝叶斯处理二项分布，所以更在意存在与否，而不是出现多少次，这是两者的根本不同。在文本分类的情况下，伯努利朴素贝叶斯可以使用单词出现向量而不是单词计数向量来训练分类器，在文档较短的数据集上，伯努利朴素贝叶斯效果更好。它适用于二项分布，数据集中存在多个特征，但每个特征都是二分类的，可以用布尔变量表示，由于本数据集中有 20 个类，那么可以使用类中专门用来二值化的参数 binarize 来改变数据，在本实验中也试着采用 binarize，但效果反而更差，所以就不进行该项操作了。

3、高斯朴素贝叶斯是不接受稀疏矩阵的，而这里恰恰就是稀疏矩阵，所以在本次实验中就无法使用高斯朴素贝叶斯。

4、补集朴素贝叶斯是多项式朴素贝叶斯的改进，它能够解决样本不平衡的问题，并且能够一定程度上忽略朴素假设。

超参数：平滑系数 alpha，是为了防止训练数据中出现过的一些词汇没有出现在测试集中导致 0 概率，之后会采用五折交叉验证得到这三种模型当超参数取哪个值的时候表现最优。

当平滑系数 alpha 默认为 1 时的实验结果：

```
Multinomial
        Accuracy:0.877
                          precision    recall  f1-score   support

             alt.atheism       0.83      0.82      0.82       255
           comp.graphics       0.95      0.73      0.82       265
 comp.os.ms-windows.misc       0.93      0.89      0.91       270
comp.sys.ibm.pc.hardware       0.82      0.86      0.84       259
   comp.sys.mac.hardware       0.81      0.94      0.87       234
          comp.windows.x       0.93      0.86      0.89       261
            misc.forsale       0.97      0.81      0.88       250
               rec.autos       0.89      0.93      0.91       221
         rec.motorcycles       0.95      0.94      0.95       225
      rec.sport.baseball       0.97      0.94      0.95       262
        rec.sport.hockey       0.95      0.95      0.95       261
               sci.crypt       0.84      0.97      0.90       230
         sci.electronics       0.94      0.82      0.88       267
                 sci.med       0.98      0.90      0.94       247
               sci.space       0.90      0.97      0.93       260
  soc.religion.christian       0.92      1.00      0.96       285
      talk.politics.guns       0.68      0.96      0.80       215
   talk.politics.mideast       0.91      0.98      0.94       250
      talk.politics.misc       0.78      0.70      0.73       248
      talk.religion.misc       0.65      0.58      0.61       235

                accuracy                           0.88      5000
               macro avg       0.88      0.88      0.87      5000
            weighted avg       0.88      0.88      0.88      5000

00:00:619732
Complement
        Accuracy:0.907
                          precision    recall  f1-score   support

             alt.atheism       0.83      0.84      0.84       255
           comp.graphics       0.93      0.86      0.89       265
 comp.os.ms-windows.misc       0.93      0.91      0.92       270
comp.sys.ibm.pc.hardware       0.86      0.86      0.86       259
   comp.sys.mac.hardware       0.87      0.92      0.90       234
          comp.windows.x       0.92      0.91      0.92       261
            misc.forsale       0.96      0.90      0.93       250
               rec.autos       0.94      0.94      0.94       221
         rec.motorcycles       0.96      1.00      0.98       225
      rec.sport.baseball       0.98      0.97      0.98       262
        rec.sport.hockey       0.95      0.99      0.97       261
               sci.crypt       0.95      0.99      0.97       230
         sci.electronics       0.95      0.89      0.92       267
                 sci.med       0.97      0.96      0.97       247
               sci.space       0.91      0.98      0.95       260
  soc.religion.christian       0.91      1.00      0.95       285
      talk.politics.guns       0.82      0.94      0.88       215
   talk.politics.mideast       0.91      0.99      0.95       250
      talk.politics.misc       0.85      0.71      0.77       248
      talk.religion.misc       0.68      0.56      0.61       235

                accuracy                           0.91      5000
               macro avg       0.90      0.91      0.90      5000
            weighted avg       0.91      0.91      0.90      5000

00:00:522055
Bernuolli
        Accuracy:0.841
                          precision    recall  f1-score   support

             alt.atheism       0.84      0.78      0.81       255
           comp.graphics       0.84      0.83      0.83       265
 comp.os.ms-windows.misc       0.97      0.12      0.22       270
comp.sys.ibm.pc.hardware       0.79      0.91      0.85       259
   comp.sys.mac.hardware       0.74      0.99      0.85       234
          comp.windows.x       0.80      0.85      0.83       261
            misc.forsale       0.47      0.96      0.63       250
               rec.autos       0.89      0.96      0.93       221
         rec.motorcycles       0.96      0.96      0.96       225
      rec.sport.baseball       0.98      0.97      0.97       262
        rec.sport.hockey       1.00      0.94      0.97       261
               sci.crypt       0.96      0.93      0.94       230
         sci.electronics       0.94      0.93      0.94       267
                 sci.med       1.00      0.85      0.92       247
               sci.space       0.98      0.87      0.92       260
  soc.religion.christian       0.99      0.97      0.98       285
      talk.politics.guns       0.85      0.91      0.88       215
   talk.politics.mideast       0.93      0.81      0.87       250
      talk.politics.misc       0.82      0.69      0.75       248
      talk.religion.misc       0.64      0.63      0.64       235

                accuracy                           0.84      5000
               macro avg       0.87      0.84      0.83      5000
            weighted avg       0.87      0.84      0.83      5000

00:00:865652
```

为了对比更加清晰，用表格显示上面的结果：

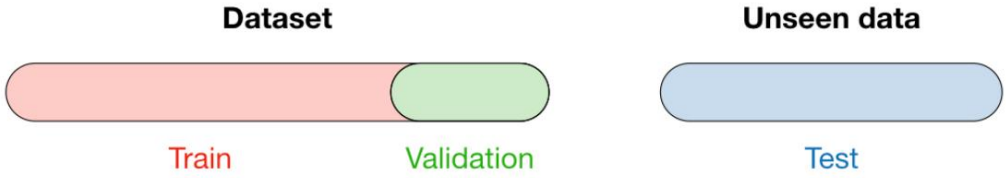
模型	用时	accuracy
Multinomial	61ms	0.877
Complement	52ms	0.907
Bernuolli	86ms	0.841

在平滑系数 alpha 默认为 1 时，Complement 不仅 Accuracy 最高而且用时最短，表现最好

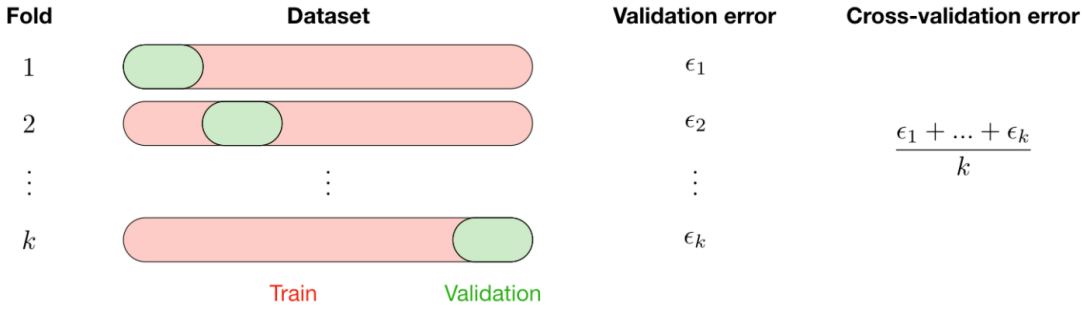
2、分别在这三种模型下通过交叉验证，选择最优超参数

交叉验证

交叉验证（Cross Validation）是在机器学习建立模型和辅助模型超参数选择时常用的技巧。交叉验证指的就是重复使用数据，把得到的样本数据按一定的准则进行切分，组合为不同的训练集和测试集，用训练集来训练模型，用测试集来评估模型预测的好坏。在此基础上可以得到多组不同的训练集和测试集，某一轮的训练集中的样本在下一轮可能成为测试集中的样本，这就是所谓的“交叉”。



交叉验证用在数据量不充足时可以方便的帮助我们进行超参选择。交叉验证的过程中一般随机把数据分成三份，如图 1.3 所示，一份为训练集（Training Set），一份为验证集（Validation Set），最后一份为测试集（Test Set）。用训练集来训练模型，用验证集来评估模型预测的好坏和选择模型及其对应的参数。



本次实验采用五折交叉验证，5-fold cross validation，将数据集分为 3 部分，training set、validation set 和 test set，一般划分的比例是 3:1:1，三者之间没有交集。

超参数 alpha 的取值范围：0.00001，0.0001，0.0005，0.001，0.005，0.01，0.05，0.1，0.2，0.5，0.7，0.9，1。由于有三个模型，所以先要画出随着 alpha 的变化，三种不同模型的平均 accuracy 的变化，进而选择较优的模型，然后针对该模型，画出对应的随着 alpha 变化，对应的交叉验证图，选择平均准确率最高且方差较小的 alpha。也就是不断增大 alpha，对于

每一个 alpha，计算验证集上的平均 accuracy 和 standard deviation，选取 accuracy 较大且表现稳定的 alpha。
由于参数的搜索空间比较大，而且又是五折交叉验证，比较耗时，对于更大型的数据集合深度学习来说，较少使用，本实验的数据集并不大，所以耗时还能接受。
交叉验证过程：

```
alpha: 1e-05
------------fold0.0-----------
1 val_accuracy1: 0.88375
1 val_accuracy2: 0.86575
1 val_accuracy3: 0.90875
------------fold1.0-----------
2 val_accuracy1: 0.8725
2 val_accuracy2: 0.85775
2 val_accuracy3: 0.9045
------------fold2.0-----------
3 val_accuracy1: 0.878
3 val_accuracy2: 0.8655
3 val_accuracy3: 0.91225
------------fold3.0-----------
4 val_accuracy1: 0.87425
4 val_accuracy2: 0.863
4 val_accuracy3: 0.91125
------------fold4.0-----------
5 val_accuracy1: 0.873
5 val_accuracy2: 0.86475
5 val_accuracy3: 0.907
alpha: 1e-05 accuracy1: 0.8763
alpha: 1e-05 accuracy2: 0.86335
alpha: 1e-05 accuracy3: 0.9087500000000001
```

```
alpha: 0.0005
------------fold0.0-----------
1 val_accuracy1: 0.882
1 val_accuracy2: 0.87325
1 val_accuracy3: 0.906
------------fold1.0-----------
2 val_accuracy1: 0.87925
2 val_accuracy2: 0.87325
2 val_accuracy3: 0.9085
------------fold2.0-----------
3 val_accuracy1: 0.87775
3 val_accuracy2: 0.86875
3 val_accuracy3: 0.90625
------------fold3.0-----------
4 val_accuracy1: 0.8715
4 val_accuracy2: 0.8655
4 val_accuracy3: 0.899
------------fold4.0-----------
5 val_accuracy1: 0.8805
5 val_accuracy2: 0.87025
5 val_accuracy3: 0.9045
alpha: 0.0005 accuracy1: 0.8782
alpha: 0.0005 accuracy2: 0.8702
alpha: 0.0005 accuracy3: 0.9048499999999999
```

```
alpha: 0.0001
------------fold0.0-----------
1 val_accuracy1: 0.879
1 val_accuracy2: 0.86575
1 val_accuracy3: 0.911
------------fold1.0-----------
2 val_accuracy1: 0.882
2 val_accuracy2: 0.8715
2 val_accuracy3: 0.917
------------fold2.0-----------
3 val_accuracy1: 0.8805
3 val_accuracy2: 0.86625
3 val_accuracy3: 0.9145
------------fold3.0-----------
4 val_accuracy1: 0.86975
4 val_accuracy2: 0.859
4 val_accuracy3: 0.90425
------------fold4.0-----------
5 val_accuracy1: 0.8765
5 val_accuracy2: 0.8575
5 val_accuracy3: 0.90425
alpha: 0.0001 accuracy1: 0.87755
alpha: 0.0001 accuracy2: 0.8640000000000001
alpha: 0.0001 accuracy3: 0.9102
```

```
alpha: 0.001
------------fold0.0-----------
1 val_accuracy1: 0.8895
1 val_accuracy2: 0.88175
1 val_accuracy3: 0.91375
------------fold1.0-----------
2 val_accuracy1: 0.88175
2 val_accuracy2: 0.87725
2 val_accuracy3: 0.9115
------------fold2.0-----------
3 val_accuracy1: 0.88575
3 val_accuracy2: 0.872
3 val_accuracy3: 0.909
------------fold3.0-----------
4 val_accuracy1: 0.88375
4 val_accuracy2: 0.87
4 val_accuracy3: 0.91025
------------fold4.0-----------
5 val_accuracy1: 0.882
5 val_accuracy2: 0.8815
5 val_accuracy3: 0.91125
alpha: 0.001 accuracy1: 0.88455
alpha: 0.001 accuracy2: 0.8765000000000001
alpha: 0.001 accuracy3: 0.9111500000000001
```

```
alpha: 0.05
-------------fold0.0------------
1 val_accuracy1: 0.90075
1 val_accuracy2: 0.90125
1 val_accuracy3: 0.893
-------------fold1.0------------
2 val_accuracy1: 0.8955
2 val_accuracy2: 0.897
2 val_accuracy3: 0.90375
-------------fold2.0------------
3 val_accuracy1: 0.887
3 val_accuracy2: 0.888
3 val_accuracy3: 0.90025
-------------fold3.0------------
4 val_accuracy1: 0.89725
4 val_accuracy2: 0.89425
4 val_accuracy3: 0.898
-------------fold4.0------------
5 val_accuracy1: 0.89425
5 val_accuracy2: 0.8945
5 val_accuracy3: 0.89575
alpha: 0.05 accuracy1: 0.89495
alpha: 0.05 accuracy2: 0.8949999999999999
alpha: 0.05 accuracy3: 0.89815
```

```
alpha: 0.1
-------------fold0.0------------
1 val_accuracy1: 0.896
1 val_accuracy2: 0.89975
1 val_accuracy3: 0.893
-------------fold1.0------------
2 val_accuracy1: 0.89675
2 val_accuracy2: 0.9015
2 val_accuracy3: 0.89225
-------------fold2.0------------
3 val_accuracy1: 0.8965
3 val_accuracy2: 0.89225
3 val_accuracy3: 0.8855
-------------fold3.0------------
4 val_accuracy1: 0.89
4 val_accuracy2: 0.8985
4 val_accuracy3: 0.9015
-------------fold4.0------------
5 val_accuracy1: 0.9005
5 val_accuracy2: 0.90075
5 val_accuracy3: 0.8915
alpha: 0.1 accuracy1: 0.89595
alpha: 0.1 accuracy2: 0.89855
alpha: 0.1 accuracy3: 0.89275
```

```
alpha: 0.2
-------------fold0.0------------
1 val_accuracy1: 0.89825
1 val_accuracy2: 0.903
1 val_accuracy3: 0.89775
-------------fold1.0------------
2 val_accuracy1: 0.89975
2 val_accuracy2: 0.90375
2 val_accuracy3: 0.88475
-------------fold2.0------------
3 val_accuracy1: 0.8925
3 val_accuracy2: 0.8985
3 val_accuracy3: 0.9105
-------------fold3.0------------
4 val_accuracy1: 0.89275
4 val_accuracy2: 0.90325
4 val_accuracy3: 0.8885
-------------fold4.0------------
5 val_accuracy1: 0.89375
5 val_accuracy2: 0.8985
5 val_accuracy3: 0.88575
alpha: 0.2 accuracy1: 0.8954000000000001
alpha: 0.2 accuracy2: 0.9014
alpha: 0.2 accuracy3: 0.89345
```

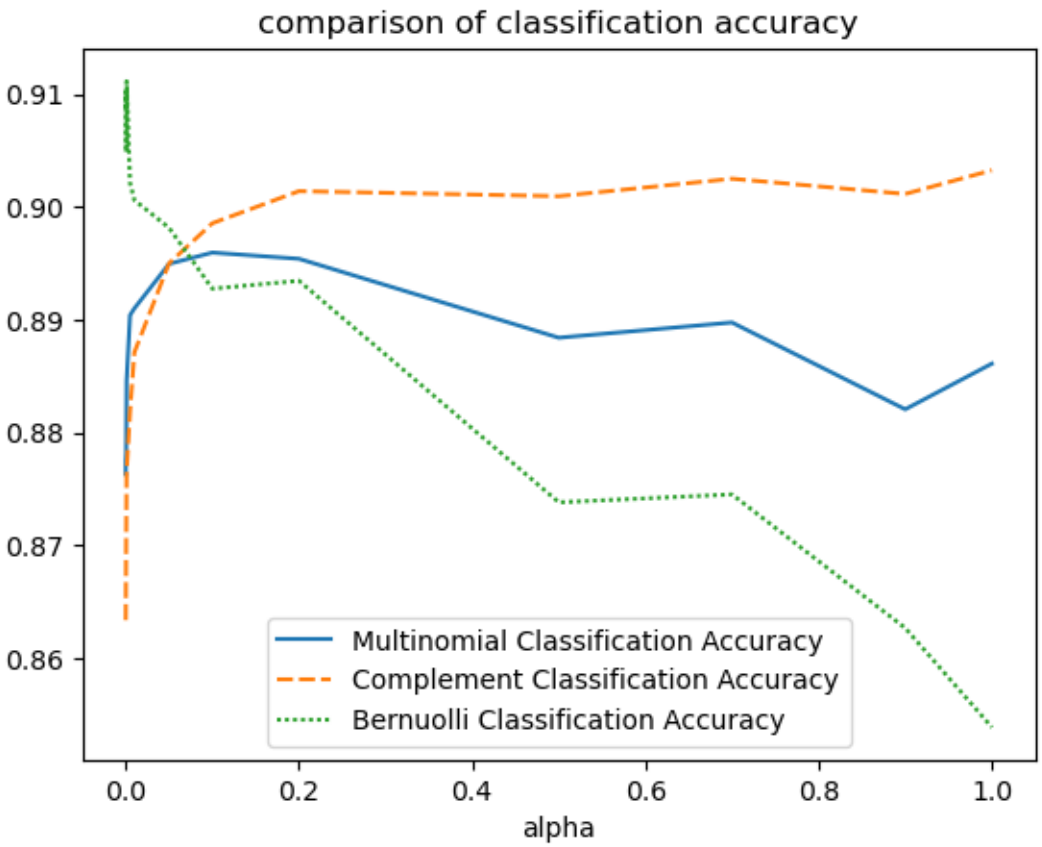
```
alpha: 0.5
-------------fold0.0------------
1 val_accuracy1: 0.88825
1 val_accuracy2: 0.903
1 val_accuracy3: 0.8625
-------------fold1.0------------
2 val_accuracy1: 0.883
2 val_accuracy2: 0.89775
2 val_accuracy3: 0.866
-------------fold2.0------------
3 val_accuracy1: 0.8875
3 val_accuracy2: 0.9015
3 val_accuracy3: 0.8845
-------------fold3.0------------
4 val_accuracy1: 0.891
4 val_accuracy2: 0.90275
4 val_accuracy3: 0.8905
-------------fold4.0------------
5 val_accuracy1: 0.89225
5 val_accuracy2: 0.89975
5 val_accuracy3: 0.8655
alpha: 0.5 accuracy1: 0.8884000000000001
alpha: 0.5 accuracy2: 0.9009500000000001
alpha: 0.5 accuracy3: 0.8737999999999999
```

```
alpha: 0.7
-------------fold0.0------------
1 val_accuracy1: 0.89525
1 val_accuracy2: 0.89875
1 val_accuracy3: 0.87025
-------------fold1.0------------
2 val_accuracy1: 0.888
2 val_accuracy2: 0.9035
2 val_accuracy3: 0.879
-------------fold2.0------------
3 val_accuracy1: 0.8845
3 val_accuracy2: 0.9025
3 val_accuracy3: 0.8675
-------------fold3.0------------
4 val_accuracy1: 0.89375
4 val_accuracy2: 0.905
4 val_accuracy3: 0.89275
-------------fold4.0------------
5 val_accuracy1: 0.88725
5 val_accuracy2: 0.90275
5 val_accuracy3: 0.863
alpha: 0.7 accuracy1: 0.8897499999999999
alpha: 0.7 accuracy2: 0.9025000000000001
alpha: 0.7 accuracy3: 0.8745
```

```
alpha: 0.9
-------------fold0.0------------
1 val_accuracy1: 0.87875
1 val_accuracy2: 0.90525
1 val_accuracy3: 0.88625
-------------fold1.0------------
2 val_accuracy1: 0.88575
2 val_accuracy2: 0.89925
2 val_accuracy3: 0.854
-------------fold2.0------------
3 val_accuracy1: 0.88475
3 val_accuracy2: 0.9
3 val_accuracy3: 0.855
-------------fold3.0------------
4 val_accuracy1: 0.8865
4 val_accuracy2: 0.90575
4 val_accuracy3: 0.86175
-------------fold4.0------------
5 val_accuracy1: 0.8745
5 val_accuracy2: 0.8955
5 val_accuracy3: 0.85625
alpha: 0.9 accuracy1: 0.8820499999999999
alpha: 0.9 accuracy2: 0.90115
alpha: 0.9 accuracy3: 0.86265
```

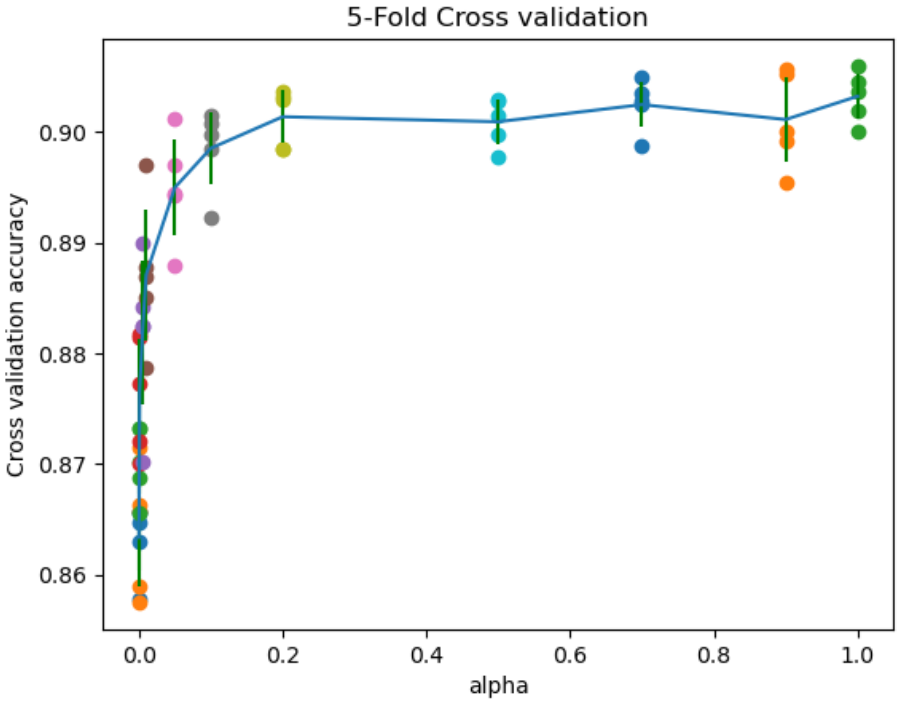
```
alpha: 1
------------fold0.0-----------
1 val_accuracy1: 0.889
1 val_accuracy2: 0.906
1 val_accuracy3: 0.83475
------------fold1.0-----------
2 val_accuracy1: 0.8885
2 val_accuracy2: 0.9
2 val_accuracy3: 0.863
------------fold2.0-----------
3 val_accuracy1: 0.882
3 val_accuracy2: 0.902
3 val_accuracy3: 0.86425
------------fold3.0-----------
4 val_accuracy1: 0.88575
4 val_accuracy2: 0.9045
4 val_accuracy3: 0.8465
------------fold4.0-----------
```

结果：



由此可以看到，Multinomial 和 Bernuolli 的 classification accuracy 均是先升高再下降，而 complement 模型则是先快速升高后趋于稳定，且最后的准确率是 complement 模型最高，所以接下来用 complement 最好的模型参数来看看最后的结果。

选取平均准确率最高且方差较小的 alpha 作为最后模型的参数:



六、最后达到的效果

	precision	recall	f1-score	support
alt.atheism	0.77	0.86	0.81	212
comp.graphics	0.87	0.89	0.88	177
comp.os.ms-windows.misc	0.93	0.91	0.92	203
comp.sys.ibm.pc.hardware	0.88	0.89	0.88	198
comp.sys.mac.hardware	0.96	0.92	0.94	212
comp.windows.x	0.93	0.88	0.91	188
misc.forsale	0.94	0.91	0.93	221
rec.autos	0.97	0.93	0.95	214
rec.motorcycles	0.95	1.00	0.97	204
rec.sport.baseball	0.97	0.95	0.96	211
rec.sport.hockey	0.91	0.98	0.94	183
sci.crypt	0.96	0.98	0.97	193
sci.electronics	0.94	0.91	0.92	182
sci.med	0.97	0.94	0.95	205
sci.space	0.93	1.00	0.96	212
soc.religion.christian	0.91	1.00	0.95	194
talk.politics.guns	0.82	0.96	0.88	189
talk.politics.mideast	0.88	0.98	0.93	198
talk.politics.misc	0.85	0.70	0.77	198
talk.religion.misc	0.70	0.49	0.57	206
accuracy			0.90	4000
macro avg	0.90	0.90	0.90	4000
weighted avg	0.90	0.90	0.90	4000

最后达到 90%的准确率（因为在调参前就阴差阳错选到了最好的 alpha,最后证实 1 就是最好的参数），效果还不错。

七、总结

本次实验采用三种不同分布朴素贝叶斯模型，通过整理数据、探索文本数据、文本特征提取、划分训练集测试集、归一化、构建模型来解决文本分类问题。先使用三种不同分布的朴素贝叶斯模型，分别是 Multinomial，Complement，Bernuolli 先进行预测，再分别通过交叉验证，选择最优超参数，最后用 complement 最好的模型参数来看最后的结果，达到了 90%的准确率。对于本次实验的数据集来说，Complement 是最适合的贝叶斯模型，不仅运行速度是最快的而且准确率是最高的，但是对于不同的数据集，可能 Bernuolli、Multinomial 的表现会更优秀，这个要根据具体情况确定。