阅卷人	得分

2021-2022 学年第一学期
数据科学产业实践
期末论文

利用 sklearn 进行垃圾邮件分类的分析

姓　　名：王子恒

专业班级：2019 级经济统计学 1 班

学　　号：2019102520121

完成时间：2022.12.20

利用 sklearn 进行垃圾邮件分类的分析

摘　要

信息技术的飞速发展推动了数字信息时代的到来在数字信息时代，由于信息技术的不断进步，它有力地推动了计算机技术的发展，使计算机技术日益呈现出新的特点，极大地提高了人类的生产效率，正在深刻地改变着人们的生活和工作方式。其中电子邮件使用户可以以非常低廉的价格、非常快速的方式，与世界上任何一个角落的网络用户联系。 电子邮件形式多样，包括图片、文字、音乐，同时，用户可以得到大量免费的新闻、专题邮件，并实现轻松的信息搜索，这极大地方便了人与人之间的沟通与交流，促进了社会的发展。电子邮件的快速发展和相关业务的开发使得电子邮件用户数达到惊人的数量，随之而来的就是垃圾邮件在网络上的泛滥。鉴于此，利用 sklearn 进行垃圾邮件的分类分析。研究内容分为三个部分：第一部分进行网络数据的收集以及清洗；第二部分利用词袋模型、朴素贝叶斯等方法进行模型的构建以及训练；第三部分对此模型进行评估，计算准确率 acc 值和 auc 值以及 roc 曲线的绘制。综上所述，根据分析研究结果综合得出合理的结论，对该模型进行补充和总结。

关键词　垃圾邮件分类；sklearn；词袋模型；朴素贝叶斯

Analysis of spam classification using sklearn

ABSTRACT

The rapid development of information technology has promoted the arrival of the digital information age. In the digital information age, due to the continuous progress of information technology, it has effectively promoted the development of computer technology, making computer technology increasingly show new characteristics, greatly improving the efficiency of human production, and is profoundly changing the way people live and work. E-mail has become an important application of transmitting communication data in the Internet environment, and has been paid more and more attention. Nowadays, e-mail has become an indispensable tool in people's daily life, mainly used for a variety of information exchanges. The rapid development of e-mail and the development of related business has made the number of e-mail users reach an alarming number, followed by the proliferation of spam on the network. In view of this, we use sklearn to classify and analyze spam. The research content is divided into three parts: The first part is to collect and clean the network data; The second part uses the word bag model, naive Bayes and other methods to build and train the model; The third part evaluates the model, calculates the accuracy of acc value and auc value, and draws roc curve. In conclusion, reasonable conclusions are drawn based on the analysis and research results to supplement and summarize the model.

KEYWORDS Spam classification; sklearn； Word bag model; Naive Bayes

目录

图目录

一 、绪论

（一）研究背景与意义

1. 研究背景

信息技术的飞速发展推动了数字信息时代的到来。在数字信息时代，由于信息技术的不断进步，它有力地推动了计算机技术的发展，使计算机技术日益呈现出新的特点，极大地提高了人类的生产效率，正在深刻地改变着人们的生活和工作方式。电子邮件已成为 Internet 环境中传送通讯数据的一个重要应用，越来越受到重视 2020 年，全球有 36.7%的公司增加对电子邮件的总体投资其中有 7%的公司增加超过 15%的投资预算，有 43.1%的公司对电子邮件预算支出保持不变，只有 1.3% 的公司减少电子邮件预算支出。

2. 研究意义

现今，电子邮件已经成为人们日常生活中不可或缺的一种重要工具，主要用于多种信息交互。电子邮件的快速发展和相关业务的开发使得电子邮件用户数达到惊人的数量，随之而来的就是垃圾邮件在网络上的泛滥。垃圾邮件主要来自匿名的转发服务器、一次性账户和僵尸主机等。一些人为了推销自己的产品或宣传网站等，通过乱发电子邮件来达到自己的目的，使得垃圾邮件远远超过正常邮件数量，占用大量用户邮箱空间和网络带宽，影响用户使用的同时，损耗广大用户的合法权益，给邮件服务提供商带来沉重的经济负担和社会压力。这已经不仅仅是技术性问题，也不仅仅是政策法律问题，而是一个全球性、综合性的问题。因此，如何快速有效的解决垃圾邮件问题具有比较大的现实意义。

二、问题描述

（一）分析目标

首先利用词袋模型抽取特征数据，并利用 sklearn 库构建垃圾邮件分类的模型并对已有的邮件进行训练集和验证集的划分，然后利用朴素贝叶斯进行模型训练，最后进行模型评估，评判该模型是否可用。

（二）现有的研究方法

目前有很多垃圾邮件分类识别的方法，如基于 BERT_DPCNN 文本分类算法的垃圾邮件分类过滤系统、基于改进的卷积神经网络的垃圾邮件过滤方法、基于深度学习的图像型垃圾邮件分类等等。本文通过利用 sklearn 构建模型，并用朴素贝叶斯方法训练模型达到垃圾邮件分类的目的。

三、数据准备

从网络上下载一些垃圾邮件和正常邮件，并将其放在一个文件这种，标明两种类型，其中 ham 标签代表正常邮件，spam 代表垃圾邮件。

四、数据探索

（一）导入函数库

```
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_curve,roc_auc_score
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import learning_curve

from matplotlib import pyplot as plt
```

图 1 函数库的导入

（二）邮件分类标记

```
def get_data():
    """
        prepare mail data and labels
    """
    #open file
    f = open("mail_data.txt","r",encoding="utf-8")

    data = []
    labels = []

    while True:
        #read a line
        l = f.readline()

        #got data
        if l:
            #ignore lines that is  not label + data
            if not l.startswith(("ham","spam")):
                continue

            #normal data
            label,mail = l.split(",",maxsplit=1)
            data.append(mail)
            if label == "ham":
                #正常邮件表示为0
                labels.append(0)
            else:
                #垃圾邮件表示为1
                labels.append(1)

        #no data any more
        else:
            break

    return data,labels
```

图 2 邮件分类标记

由图 2 可以看出其中代码设置将正常邮件的类别 ham 表示为 0，垃圾邮件的类别 spam 表示为 1。

五、建模分析

（一）抽取特征数据

导入数据并用词袋模型提取特征数据，其代码及结果如下：

代码实现：

```python
if __name__=="__main__":

    #组织数据
    x,y = get_data()

    print("邮件列表:\n",x,"\n","标记列表:\n",y)

    #词袋模型，抽取特征
    vector = CountVectorizer()
    x_ = vector.fit_transform(x)

    #x_ is a sparse matrix
    x_data = x_.toarray()
```

图 3 词袋模型

部分运行结果：

```
邮件列表:
 ['"SPJanuary Male Sale! Hot Gay chat now cheaper, call 08709222922. National rate from 1.5p/min cheap to 7.8
p/min peak! To stop texts call 08712460324 (10p/min)"\n', "Yeah you should. I think you can use your gt atm n
ow to register. Not sure but if there's anyway i can help let me know. But when you do be sure you are read
y.\n", 'Nationwide auto centre (or something like that) on Newport road. I liked them there\n', 'He is there.
You call and meet him\n', "Yeah sure I'll leave in a min\n", 'URGENT! Your Mobile number has been awarded wit
h a £2000 prize GUARANTEED. Call 09061790121 from land line. Claim 3030. Valid 12hrs only 150ppm\n', '"Mah b,
I\'ll pick it up tomorrow"\n', 'Then she dun believe wat?\n', "I've sent u my part..\n", 'Hey so whats the pl
an this sat? \n', "Can you just come in for a sec? There's somebody here I want you to see\n", "Yup. Izzit st
ill raining heavily cos i'm in e mrt i can't c outside.\n", '"I think we\'re going to finn\'s now, come"\n',
'Our Prasanth ettans mother passed away last night. Just pray for her and family.\n', '"Enjoy the showers of
possessiveness poured on u by ur loved ones, bcoz in this world of lies, it is a golden gift to be loved trul
y.."\n', "He's really into skateboarding now despite the fact that he gets thrown off of it and winds up with
bandages all over his arms every five minutes\n", 'You please give us connection today itself before  DECIMAL
or refund the bill\n', 'Congratulations ur awarded 500 of CD vouchers or 125gift guaranteed & Free entry 2 10
0 wkly draw txt MUSIC to 87066 TnCs www.Ldew.com1win150ppmx3age16\n', 'Mum not going robinson already.\n', 'G
ot meh... When?\n', 'We are both fine. Thanks\n', '"awesome, how do I deal with the gate? Charles told me las
t night but, uh, yeah"\n', '"I thought i\'d get him a watch, just cos thats the kind of thing u get4an18th. A
```

图 4 邮件列表

邮件标记列表：

```
标记列表:
[1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0]
```

图 5 邮件标记结果

从图 5 可以看出，导入的邮件大多数为正常邮件，只有小部分是垃圾邮件，这与收集的邮件数据息息相关。

（二）训练集和验证集的划分

利用 train_test_split 函数进行训练集和验证集的划分，设置划分比例为 0.3，其代码如下：

```python
#划分训练集，验证集
x_train, x_validate, y_train, y_validate = train_test_split(x_data, y, test_size=0.3, random_state=3)
```

图 6 训练集和验证集的划分

（三）训练模型

利用朴素贝叶斯进行模型的训练：

```
#实例化分类模型
clf = MultinomialNB(alpha=1)

#训练模型
clf.fit(x_train, y_train)
```

图 7 训练模型

（四）模型评估

计算准确率进行模型的评估：

```
#准确率
acc = accuracy_score(y_validate, y_pred)
print(acc)
```

```
0.95
```

图 8 模型评估

由图 8 可以看出该模型的准确率达到 95%，可见该模型对垃圾邮件的分类的正确率是很高的，可以有效地进行垃圾邮件的分类。

六、可视化

计算 auc 并利用 matplotlib.pyplot 函数进行 roc 曲线的绘制，将 x 轴设置标签为“FPR”，表示为“错误的预测为正的数量/原本为负的数量”，将 y 轴设置标签为“TPR”，表示为“正确的预测为正的数量/原本为正的数量”。：

```
#roc 曲线
fpr, tpr, threshold = roc_curve(y_validate, y_pred_proba, pos_label=1)
auc = roc_auc_score(y_validate, y_pred_proba)
#可视化
plt.figure()
#plt.rcParams["font"]
plt.plot(fpr, tpr, "ro-", label="roc curve")
plt.title("auc %.1f, acc %f.1"%(auc, acc))
plt.xlabel("FPR")
plt.ylabel("TPR")
plt.grid()
plt.legend(loc="best")
plt.show()
```

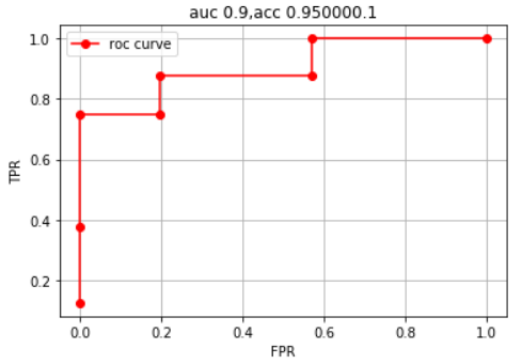

图 9 roc 曲线

由图 9 可以看出 auc=0.9，表明该模型是一个很好地分类器模型，可以有效

地将垃圾邮件和正常邮件划分开。

七、结论

利用词袋模型抽取特征数据，并利用 sklearn 库构建垃圾邮件分类的模型并对已有的邮件进行训练集和验证集的划分，然后利用朴素贝叶斯进行模型训练，最后进行模型评估，计算其准确率 acc 值和 auc 值并绘制 roc 曲线，即可得出此模型可有效地分辨出垃圾邮件，但也因样本数据过小，正常邮件和垃圾邮件数量相差过大可能会导致模型出现错误，因此需要加大样本数据来尽可能减少此类错误。