

SVM实验报告

**江辉--171250506**



**2020-5-2**

**南京大学**

**软件学院**

目录

[一、 实验内容描述 1](#_Toc39339375)

[1、 项目目录描述 1](#_Toc39339376)

[二、 实验结果图 1](#_Toc39339377)

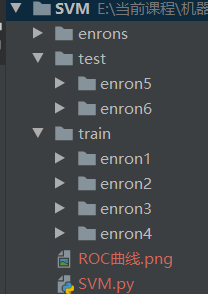
[三、 算法原理阐述 2](#_Toc39339378)

[四、 代码简述 3](#_Toc39339379)

# 实验内容描述

本程序用python编写，使用scikit-learn构建SVM模型。评价指标采用混淆矩阵、精准率、召回率、RUC曲线、AUC值

## 项目目录描述



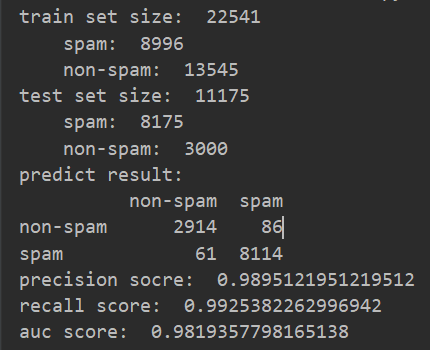
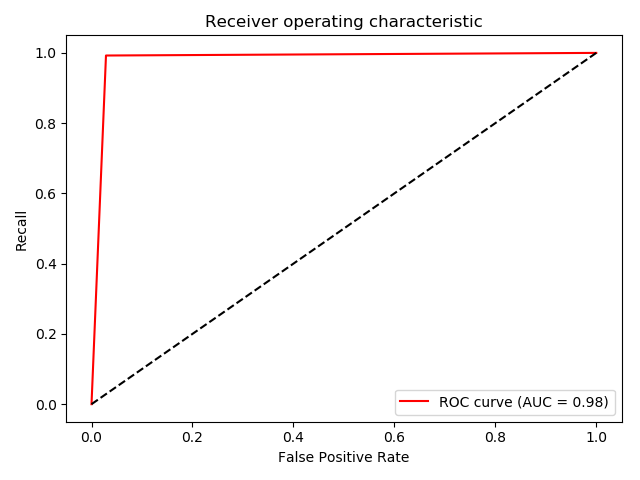
本项目推荐使用pycharm打开

**SVM.py**里写是代码

**train**和**test**分别是训练集和测试集，我将enron1-4作为训练集，enron5-6作为测试集

# 实验结果图

下面是实验的结果图，左边打印了采用的训练集和测试集的大小、混淆矩阵、精准率、召回率和AUC值。右图是打印的ROC曲线。



# 算法原理阐述

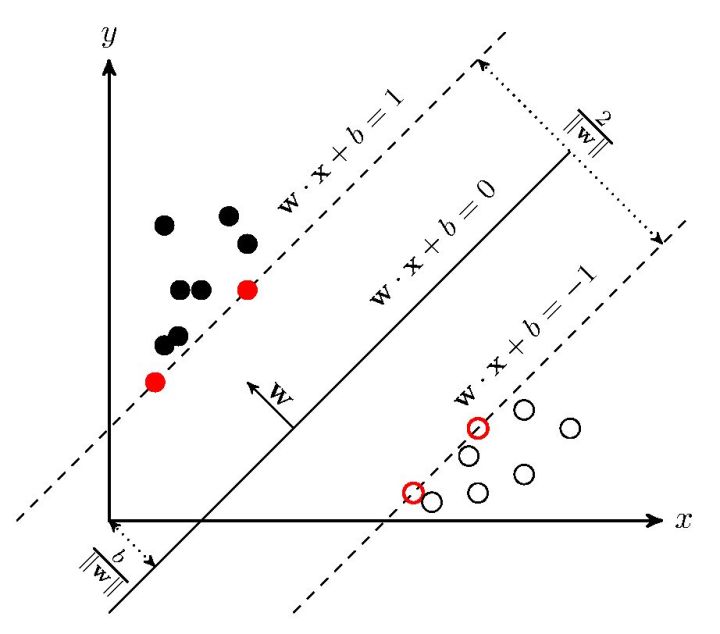
SVM是一种二分类模型，通过计算特征空间上的最大间隔来进行线性分类。SVM的学习策略是通过求解凸二次规划的最优化问题来求解一个能够正确划分训练数据集并且几何间隔最大的分离超平面。

最大化间隔的原理：

定义每个样本x的n个特征(x1，x2……xn)和一个类型标记y∈{-1,+1}

假设所求的决策边界是w·x+b=0，对于线性可分的数据集，w·x+b的值就是预测值

定义样本点到决策边界的距离为γ = y·(w/||w|| · x + b/||w||)

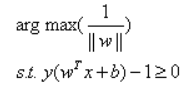
因为希望求得一个能将所有样本点的间隔最明显(最大)的边界，所以以间隔最小的点作为支持向量，最大化决策边界到支持向量的最小间隔γ即可，如下图中红色点为支持向量，希望找到最大化间隔γ的边界。即求下面的最优化函数

上述定义中γ = min γi = min yi·(w/||w|| · xi + b/||w||)

同除γ得yi·(w/||w||γ · xi + b/||w||γ)>=1

此时||w||和γ都是标量，可以简化为yi·(w · xi + b)>=1

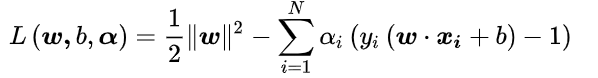
因此上面的问题可以简化为

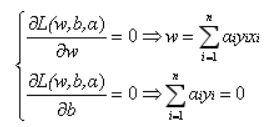


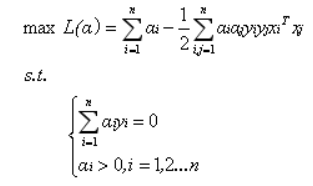
在xi给定的情况下，最大化 γ 等价于最大化1/||w||，也就等价于1/2||w||^2 ，即

约束条件是yi·(w · xi + b) >= 1

对于含不等式约束的凸二次规划问题，使用拉格朗日乘子法来处理

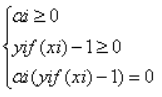
定义目标函数

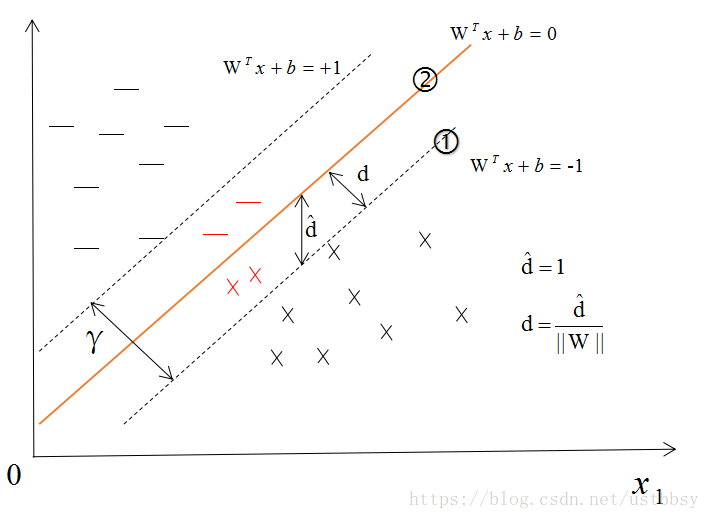
求L关于求偏导数得



带入前面的L(w,b,α)化简后得到右边最终的优化目标函数

该问题对应的KKT条件为



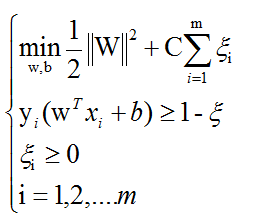
SVM引入了松弛变量解决无法完全线性分割的噪点问题

如右图中分割边界中间有一些散落的点，这些点不满足γ>=1的限制条件。因此需要加入一个松弛变量 ξ

这时约束条件不再是1，而是1 - ξ

这里写图片描述

数学模型中加入一个惩罚因子C，C值越大，对分类的惩罚越大



目标函数推导过程同前面没有加入松弛变量的一样，结果也一样，不过修改惩罚参数C更能解决噪点数据的问题

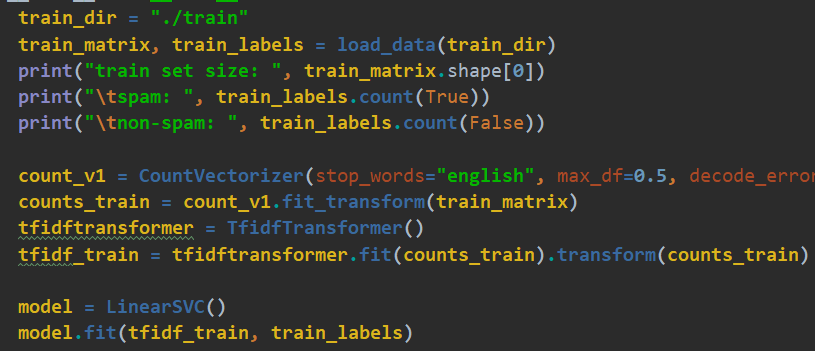
# 代码简述

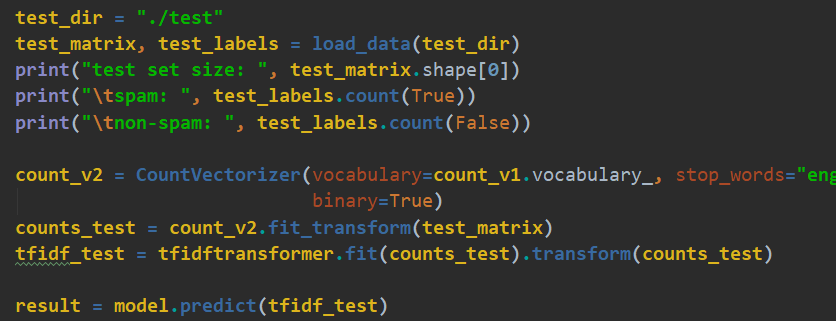
先切分了train和test的数据集并根据ham和spam文件夹来打上对应标签(0和1)





整个代码实现处理过程参考了作业pdf里给出的样例代码，使用TF-IDF特征，用sklearn的默认的SVM模型。





在测试评估时采用了精确率、召回率、ROC曲线、AUC值作为评估指标

