**实验二 基于SVM的数据分类**

**一、实验目的**

通过编程调试实现SVM算法并在给定数据集上进行测试，实现数据分类。

**二、实验设备和环境**

**windows系统下的vscode，python3.9**

**三、实验内容**

自己编写或从网上下载SVM的源代码，调试运行SVM算法，研究其参数设定，并在给定数据集上完成测试，实现数据分类，并计算准确率（Precision）和查全率（Recall）。

**四、实验过程、结果及分析**

1. 实验过程，关键代码，实验结果及分析。

在本次实验中，首先通过网络搜索了解到python的scikit-learn外部库可以直接使用现成

的svm，因此本次实验的代码仍然选择使用python进行研究开发

本次实验所有代码如下

import numpy as np

import sklearn.svm as svm

xx,yy=open('train\_sample.txt','r'),open('train\_sample\_label.txt','r')

s\_data=np.array(list(map(lambda x:list(map(lambda x:float(x),x.strip().split())),xx.readlines())))

s\_label=np.array(list(map(lambda x:int(x.strip()),yy.readlines())))

xx.close();yy.close()

bord=svm.LinearSVC()

bord.fit(s\_data,s\_label)

xx,yy=open('test\_sample.txt','r'),open('test\_sample\_label.txt','r')

t\_data=np.array(list(map(lambda x:list(map(lambda x:float(x),x.strip().split())),xx.readlines())))

t\_label=np.array(list(map(lambda x:int(x.strip()),yy.readlines())))

xx.close();yy.close()

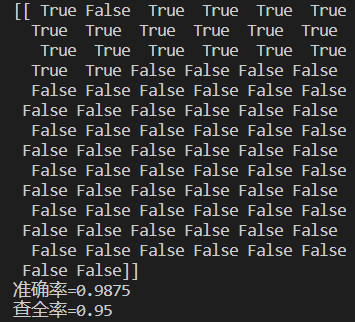
t\_otc=((t\_data.dot(bord.coef\_.T)+bord.intercept\_)>0).T

print(t\_otc)

print('准确率=',(t\_otc==t\_label).sum()/len(t\_label),sep='')

print('查全率=',(t\_otc&t\_label).sum()/(t\_label==1).sum(),sep='')

其中大部分代码实际上是读取文件的过程，而LinearSVC和.fit()是核心的模型构建过程，通过fit函数投入训练样本，使用svm的机器学习算法计算得到模型的各变量的系数和常量，再通过该模型来分类测试样本，测试结果如下图



从结果来看，程序很好地完成了对所有样本的测试和结果的输出，准确率有98.75%，在本例中就是有1次预测错误和79次预测正确，查全率有95%，在本例中就是有19个正样本成功找出，1个正样本判断错误，准确率和查全率是比较高的，这说明sklearn库自带的线性svc的学习效果对于二元分类有不错的效果

1. 实验中遇到的问题及解决办法。

首先是对于如何进行svm算法实现的问题，在查询了网络上的相关资料并仔细查看了书

上的相关内容后，我认为自主使用基础逻辑实现svm存在一定困难，而恰好在寻找时发现python有一个外部库scikit-learn，其中有一个svm模块可以直接实现svm相关的机器学习算法模型，于是研究使用了其中的LinearSVC模型进行一个边界平面的拟合

接着是关于如何对测试用例使用已拟合的模型进行测试的问题，经过一定的考虑发现读入的向量数据是一个矩阵形态，因此使用了矩阵乘法，将向量数据和模型系数向量相乘再加上常数项，得到一个中间结果矩阵，这个中间矩阵中的每个值为点和平面的某一种距离，只要通过其正负来判断每个点在平面的哪一边即可，因此再使用一个np数组的整体操作将其转化为01矩阵，true表示是正样本，false表示是负样本

最后是计算准确率和查全率时，如果对矩阵的元素一个一个统计将会在循环中浪费很多时间，于是计算准确率时，将测试结果t\_otc和读入的实际情况t\_label使用等号对比并对结果求和，如此来求得总共有多少个测试结果与实际情况相符，而计算查全率时，我发现需要计入的样本需要在测试结果和实际情况中都呈现true，因此对t\_otc与t\_label作逻辑与操作再对矩阵求和，如此就可以求出两个矩阵中均为true的项数，并用它来计算查全率

1. 实验还存在哪些问题。

在实验中，没有对LinearSVC进行参数的调整而全部使用默认值，在结果中准确率没有

达到100%，对于正样本有缺漏的情况，因此应该考虑对模型进行精度的强调或方法的转化，如使用非线性或高斯法进行模型拟合以获取更优秀的分类边界

本次实验中，训练样本可能仍然较少，测试样本量也较少，得到的学习结果和测试结果可能存在较高的随机性，可能需要更多的数据来支持

本次实验是二元分类，在多元分类中可能SVC的作用较为有限

1. 本实验有哪些收获和心得体会。

在本次实验中充分地学习了svm算法的原理，并用一定量的学习和实操巩固了对这个

算法的实现的基础，进一步加深了对数据分类的理解

1. 其它需要补充的问题。

暂无