# 模式识别作业——SVM与决策树，神经网络对比

# 实验环境

## 数据集的选取

本实验数据集采用UCI数据集中的iris（鸢尾花）数据集和wine（红酒）数据集

## 编程工具的选取

本实验使用python语言，由于本实验中模型的训练消耗时间不多，代码复杂度也较低，因此jupyter notebook在训练模型中交互式运行的优势并没有那么大，故采用pycharm编程环境。

## 库的选取

本实验采用sklearn库中的函数训练模型

## 操作系统

操作系统为Windows10操作系统，版本为19043.1645

# 主要函数

主要函数可以分为两类，第一类是应老师要求，自己编写的处理数据集的代码，我将对数据集处理思路做简单介绍。第二类是sklearn库中机器学习算法的函数接口，我将对如何fit数据集以及函数中用到的的参数做一个介绍。

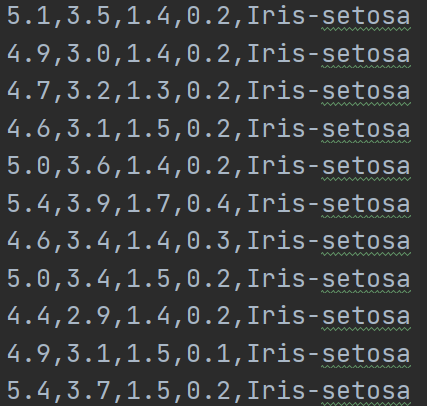
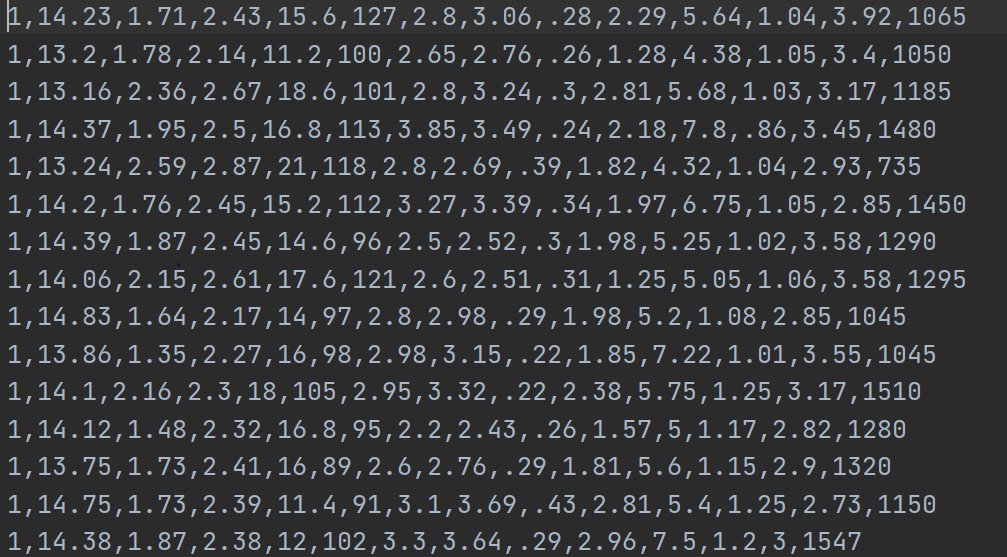
## 数据集处理部分

### 数据集简述

这部分其实不是传统意义上编程的函数，因为不同数据集的target和data顺序不同。举个例子，iris数据集中target位于每一行的尾部,而wine数据集中target位于每一行头部。因此对于二者的处理在细节上不能保证完全一致，也不能抽象成为一个代码重复度极高的函数。但虽说如此，二者的逻辑上相似度还是极高的，所以我把数据集处理部分放在主要函数部分进行说明。

### 数据集处理

在处理数据集之前，首先对数据集格式做一个介绍。UCI数据集的文件后缀为.data类型，实际上就是.txt，在pycharm中将.data与.txt关联后就可以用处理.txt文件的方式处理.data文件了。处理数据之前，我们先创建一个空字典，储存数据和标签。数据集总体可以分为两部分，数据部分和标签部分，每一组数据占一行如下图所示。

因此我们需要用文件操作中的readlines()函数读取所有行。采用.strip()来去除行首尾空格等空字符后，再进行下一步处理。由于txt文件中所有数据都是string类型，所以我们要先把这些string类型的数据数字转化成float类型，把标签转换成int类型。转化方法就是用split(‘ , ’)，把逗号作为分隔符将每个数字分开存在一个列表中，然后遍历这个列表，把每个元素都转化为float类型。

处理到这一部分基本就结束了，最后需要做的就是往一开始创建的空字典iris和wine中添加键值对了，将’data’，’target’分别作为键，值就是我们刚刚类型转化过后的列表。需要注意的是，作为字典的值的列表，还要用np.array转化为np数组，因为我们后边用到的函数处理的对象是ndarray这种类型。

## 机器学习算法函数接口

* **支持向量机**

SVC()函数，是sklearn中的支持向量机分类器。我在训练时采用参数如下

**SVC(kernel=’linear’, gamma=’auto’, C=1)**

其中kernel代表核函数，linear是线性核，除线性核之外我还用了多项式核和高斯核。

Gamma代表核函数系数，默认为auto，只对高斯核，多项式核和sigmoid核起作用

C代表软间隔分类时的惩罚系数，C=0代表硬间隔分类。

* **BP神经网络**

MLPClassifier()，是sklearn中的神经网络分类器。我在训练时采用参数如下

**MLPClassifier(random\_state=1, max\_iter=2000, alpha=0.001, activation=’tanh’)**

其中random\_state代表随机数种子，这里我选取1。max\_iter为迭代次数，alpha为L2正则化参数，activation为激活函数。

* **决策树**

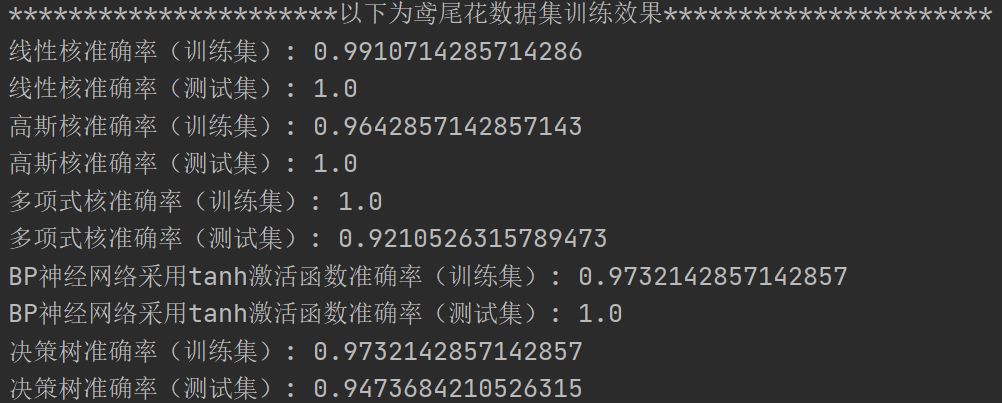
DecisionTreeClassifier()，是sklearn中的决策树分类器。我在训练时采用参数如下

**DecisionTreeClassifier(random\_state=3, max\_depth=4, criterion=’entropy’)**

其中random\_state代表随机数种子，为了与BP神经网络区别我采用3。Max\_depth为树最大深度，criterion代表采用什么作为分类标准，可选的有’gini’，’entropy’。前者是基尼系数，也是默认参数，在CART算法中采用基尼系数。后者是信息增益比，由于我们的要求是C4.5算法，因此我在这里采用信息增益比。

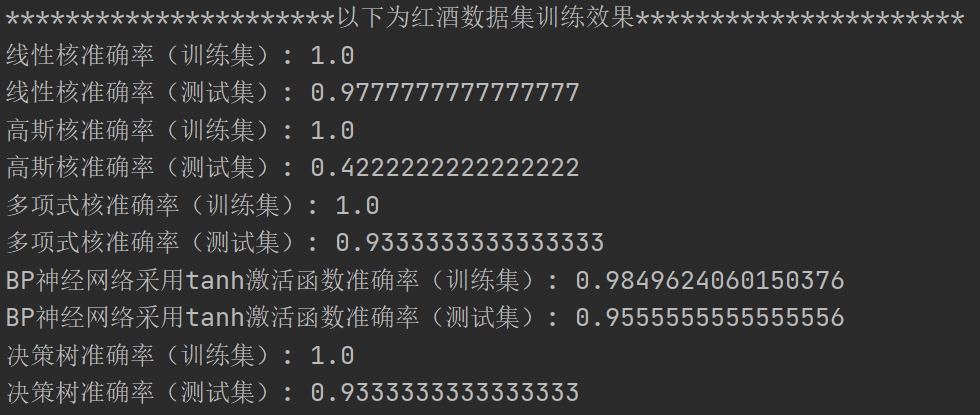
# 实验结果分析

## 鸢尾花数据集



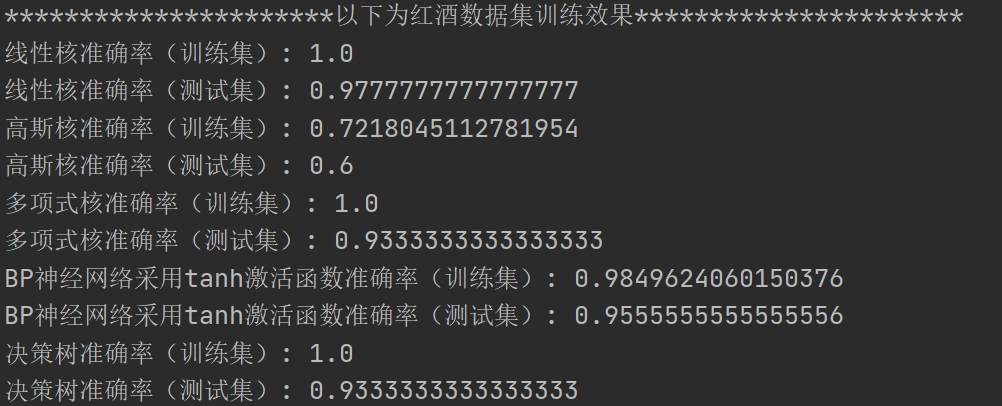
上图可以看出，采用SVM训练时，线性核表现最好，测试集达到了100%准确率，训练集也有较高准确率。多项式核可能发生了过拟合，训练集达到了100%准确率，但是测试集准确率较低。采用tanh激活函数的BP神经网络模型表现也较好，但是我在将激活函数换位relu后，表现会有所下降。C4.5决策树模型表现也还可以。

## 红酒数据集



大部分模型在红酒数据集中表现都不如鸢尾花数据集，而且可能都发生了一定过拟合情况，其中高斯核的SVM过拟合及其严重，训练集准确率100%的情况下测试集正确率只有42.2%。

其中表现最好的分类器是采用线性核的SVM。



为了改善高斯核的SVM的过拟合问题，我将gamma参数由auto改为scale。发现做出改变后准确率确有所提升。

# 总结与感悟

本次实验我有两个收获吧，首先就是自己动手下载数据集、处理数据集。虽然现在看来是比较简单的，但是一开始做的时候也确实有点无从下手，经过这次实验自己以后对数据集的处理能力应该会有所提高。

第二个收获就是将课本中的理论知识联系到了实际生活中的数据集的分析处理，通过改变不同的参数，使得模型在面对数据集时获得一个更好地表现，这也对我以后训练模型的能力应该也是有一定提升的。