**《机器学习》上机实验报告**

**学号：10930130448 姓名：吕思勤**

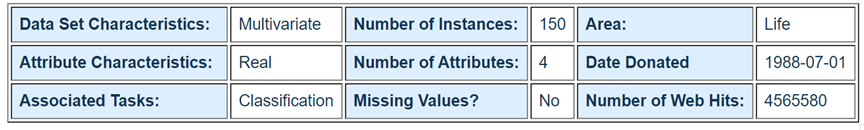
**第一次实验：鸢尾花分类**

**一、 数据探索**

1.数据来源

https://archive.ics.uci.edu/ml/datasets/Iris

2.数据描述



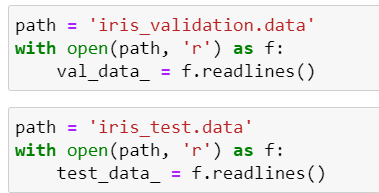
根据萼片长、萼片宽度、花瓣宽度找到鸢尾植物的类别

**二、数据预处理**

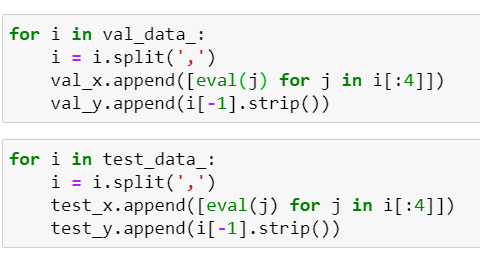
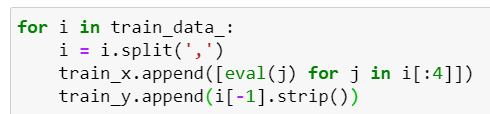
1. 数据读取

一張含有 文字 的圖片

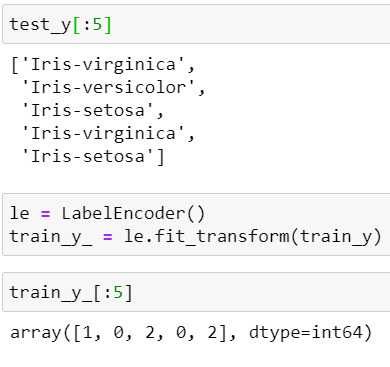
自動產生的描述



1. 特征选取



1. 建立标签



'Iris-versicolor' :0

'Iris-virginica' :1

'Iris-setosa' :2

1. 建立np矩阵

一張含有 文字 的圖片

自動產生的描述

**三、****模型训练（****SVC分类）**

1. SVC模型参数

C：C-SVC的惩罚参数C?默认值是1.0

C越大，相当于惩罚松弛变量，希望松弛变量接近0，即对误分类的惩罚增大，趋向于对训练集全分对的情况，这样对训练集测试时准确率很高，但泛化能力弱。C值小，对误分类的惩罚减小，允许容错，将他们当成噪声点，泛化能力较强。

kernel ：核函数，默认是rbf，可以是‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’

– 线性：u’v

– 多项式：(gamma\*u’v + coef0)^degree

– RBF函数：exp(-gamma|u-v|^2)

–sigmoid：tanh(gammau’\*v + coef0)

degree ：多项式poly函数的维度，默认是3，选择其他核函数时会被忽略。

gamma ： ‘rbf’,‘poly’ 和‘sigmoid’的核函数参数。默认是’auto’，则会选择1/n\_features

coef0 ：核函数的常数项。对于‘poly’和 ‘sigmoid’有用。

probability ：是否采用概率估计.默认为False

布尔类型，可选，默认为False

决定是否启用概率估计。需要在训练fit()模型时加上这个参数，之后才能用相关的方法：predict\_proba和predict\_log\_proba

shrinking ：是否采用shrinking heuristic方法，默认为true

tol ：停止训练的误差值大小，默认为1e-3

cache\_size ：核函数cache缓存大小，默认为200

class\_weight ：类别的权重，字典形式传递。设置第几类的参数C为weight\*C(C-SVC中的C)

verbose ：允许冗余输出？

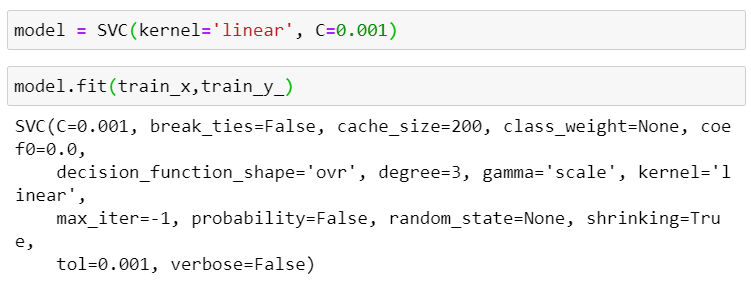
max\_iter ：最大迭代次数。-1为无限制。

decision\_function\_shape ：‘ovo’, ‘ovr’ or None, default=None3

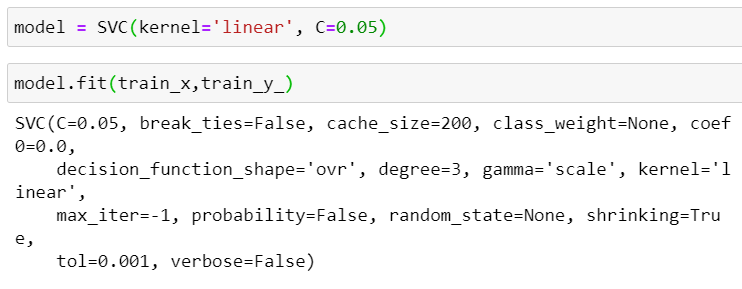
random\_state ：数据洗牌时的种子值，int值

2. SVC模型

**C=0.001**

****

**C=0.05**

****

**四、分析与总结**

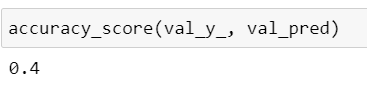
1. 评价指标Accuracy

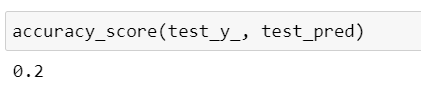
Accuracy（准确率）

分类正确的样本数与样本总数之比。即：(TP + TN) / ( ALL )

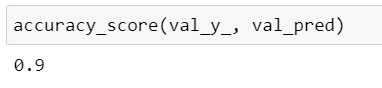
1. 结果分析

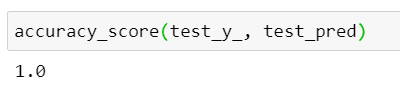
C=0.001时:





C=0.05时:





显然，由实验结果可知，当C值越大时，分类预测的准确率越高。