说明书

先导入必要的包:  
import pandas as pd  
import numpy as np  
from sklearn.linear\_model import SGDClassifier  
from sklearn.metrics import accuracy\_score  
from sklearn.preprocessing import StandardScaler  
from sklearn.model\_selection import train\_test\_split  
import os

预先将准备好的数据集放在S3存储桶中



读入citrus

dataset\_path = os.path.join("~/s3data/dataset", 'citrus.csv')  
df = pd.read\_csv(dataset\_path)

citrus里面存放着二分类数据，有五个特征来判断是橙子还是葡萄

分别得到label个features

classes=df.get('name').values  
label=[]  
features=df.drop('name', axis=1).values  
for item in classes:  
 if(item=="orange"):{  
 label.append(1)  
 }  
 else: {  
 label.append(0)  
 }

然后将数据集分为测试集与训练集，按1：2的比例进行分割

X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, label, test\_size=0.33, random\_state=0)

然后构建SGD分类器进行训练

#  构建SGD分类器进行训练    
sgdClassifier = SGDClassifier(random\_state=42)  
sgdClassifier.fit(X\_train, y\_train)  
#  y作为label已经是0,1形式，不需进一步处理    
#  使用训练好的SGD分类器对陌生数据进行分类

对y\_test数据进行预测

test\_predicted = sgdClassifier.predict(X\_test)  
y\_test=np.array(y\_test).reshape(3300,1)  
test\_predicted=np.array(test\_predicted).reshape(3300,1)  
X\_test=np.array(X\_test)

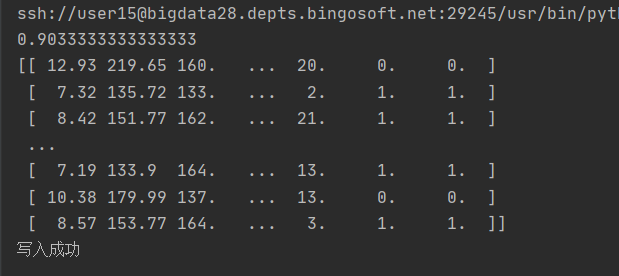
输出准确率

print(accuracy\_score(y\_test,test\_predicted))

将测试集数据以及测试结果存入S3

result=np.hstack((X\_test,y\_test))  
result=np.hstack((result,test\_predicted))  
np.savetxt('result.csv', result, delimiter=',')  
print(result)  
print("写入成功")

结果截图：



准确率有0.903