**实验1 机器学习实验**

班级：软件工程2102 学号： 姓名：Mcrivers

## 实验目的

掌握逻辑回归、支持向量机和K-means聚类等常用机器学习方法的原理，熟悉分类问题和聚类问题的求解流程，理解输入特征个数和不同超参数对机器学习算法的训练和测试结果的影响。

### 实验原理

逻辑回归（Logistic Regression）是一种有监督学习方法，属于分类算法，可以处理二元分类以及多元分类，预测结果是离散的分类，例如判断这封邮件是否是垃圾邮件，以及用户是否会点击此广告等。逻辑回归只是对线性回归的计算结果加上了一个Sigmoid函数（Logistic函数），将数值结果转化为0到1之间的概率，根据这个概率可以做预测，例如概率大于0.5，则这封邮件就是垃圾邮件，或者肿瘤是否是恶性的等等。

支持向量机(Support Vector Machine)也是一种有监督学习方法，它尝试寻找一个最优决策边界，使距离两个类别最近的样本最远，从而对分类问题提供良好的泛化能力。支持向量机是由Cortes和Vapnik于1995年首先提出的，在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

K-means（又称k-均值或k-平均）聚类是一种无监督学习方法。K-means聚类算法的思想是首先随机确定k个中心点作为聚类中心，然后把每个数据点分配给最邻近的中心点，分配完成后形成k个聚类，计算各个聚类的平均中心点，将其作为该聚类新的类中心点，然后重复迭代上述步骤直到分配过程不再产生变化。

## 实验条件

1. python 3.7或3.8，sklearn（python的重要机器学习库，其中封装了大量的机器学习算法，如：分类、回归、降维以及聚类），numpy（python第三方库，用于科学计算），matplotlib（python第三方库，主要用于进行可视化）等。
2. 鸢尾花数据集

鸢尾花（Iris）数据集是常用的分类实验数据集，由Fisher, 1936收集整理。Iris也称鸢尾花卉数据集，是一类多重变量分析的数据集。数据集包含150个数据集，分为3类，每类50个数据，每个数据包含4个属性。可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性预测鸢尾花卉属于（Setosa，Versicolour，Virginica）三个种类中的哪一类。

## 实验内容

1. 基于逻辑回归的鸢尾花分类

表1 鸢尾花分类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 分类算法 | 算法参数 | 输入特征个数 | 输入特征名称 | 训练集的准确率acc | 测试集的准确率acc | 测试集的F1分数f1\_macro,精确率precision\_macro ,召回率recall\_macro |
| 逻辑回归 | C=1.0,multi\_class='ovr',solver='liblinear',tol=0.0001,penalty='l2'  备注：C为正则化系数λ的倒数，默认值为1 | 2 | 萼片长度、萼片宽度 | 0.678 | 0.633 | F1macro 0.5680 , Pmacromacro 0.8167 , Rmacro 0.6812 |
| 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.933 | 0.950 | F1macro 0.9511 , Pmacromacro 0.9524 , Rmacro 0.9565 |
| 支持向量机  （线性核） | C=0.5,decision\_function\_shape='ovr'  备注：C为惩罚参数，默认值为1.0 | 2 | 萼片长度、萼片宽度 | 0.811 | 0.833 | F1macro 0.8365 , Pmacromacro 0.8624 , Rmacro 0.8510 |
| 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.989 | 1.000 | F1macro 1.0000 , Pmacromacro 1.0000 , Rmacro 1.0000 |
| 支持向量机  （高斯核） | C=0.5,gamma=1.0,decision\_function\_shape='ovr'  备注：C为惩罚参数，默认值为1.0；gamma为高斯核的宽度 | 2 | 萼片长度、萼片宽度 | 0.789 | 0.783 | F1macro 0.7807 , Pmacromacro 0.8246 , Rmacro 0.8076 |
| 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 | 0.989 | 1.000 | F1macro 1.0000 , Pmacromacro 1.0000 , Rmacro 1.0000 |

1. 基于支持向量机（SVM）的鸢尾花分类

表2 支持向量机（线性核与高斯核）的分类结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 分类算法 | 算法参数 | 输入特征个数 | 输入特征名称 | 交叉验证结果的平均值 | 最佳结果 | 最佳参数 | 测试集的准确率acc | 测试集的F1分数f1\_macro,精确率precision\_macro ,召回率recall\_macro |
| 支持向量机（线性核） | "C":[0.1,0.5,0.8, 1.0,5.0,10.0] | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 |  | 0.988 | 'C':1.0 | 1.000 | F1macro 1.0000 , Pmacromacro 1.0000 , Rmacro 1.0000 |
| 支持向量机（高斯核） | "C":[0.1,0.5,0.8, 1.0,5.0,10.0], "gamma":[0.01,0.1,1,10] | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 |  | 0.977 | 'C': 5.0, 'gamma': 0.1 | 1.000 | F1macro 1.0000 , Pmacromacro 1.0000 , Rmacro 1.0000 |
| 逻辑回归 | "tol": [1e-4, 1e-3, 1e-2,1e-1],"C": [0.01,0.1,0.5,0.8,1.0,5.0],'penalty':['l1','l2'] | 4 | 萼片长度、萼片宽度、花瓣长度、花瓣宽度 |  | 0.952 | 'C': 5.0, 'penalty': 'l1', 'tol': 0.0001 | 0.967 | F1macro 0.9670 , Pmacromacro 0.9670 , Rmacro 0.9670 |

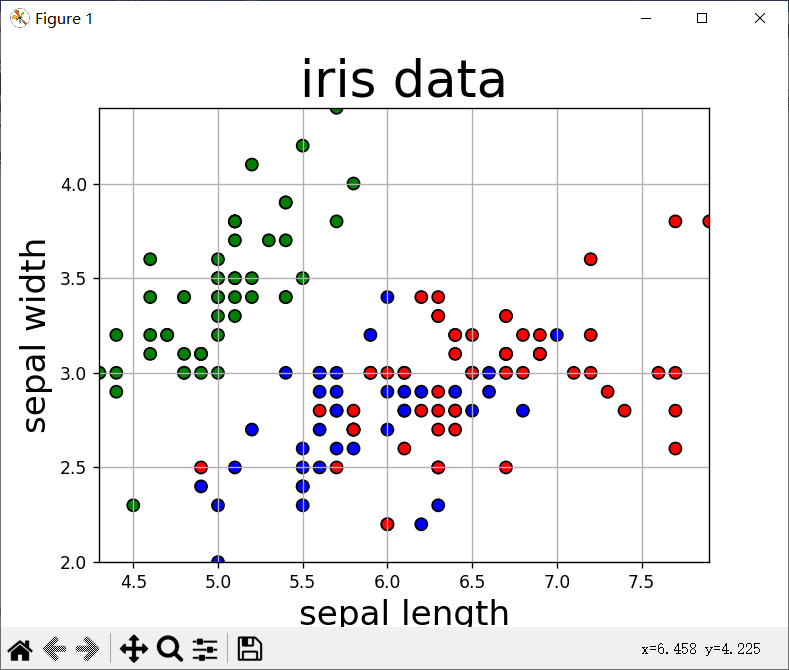
1. 基于K-means的鸢尾花聚类

表3 基于k-均值的鸢尾花聚类结果

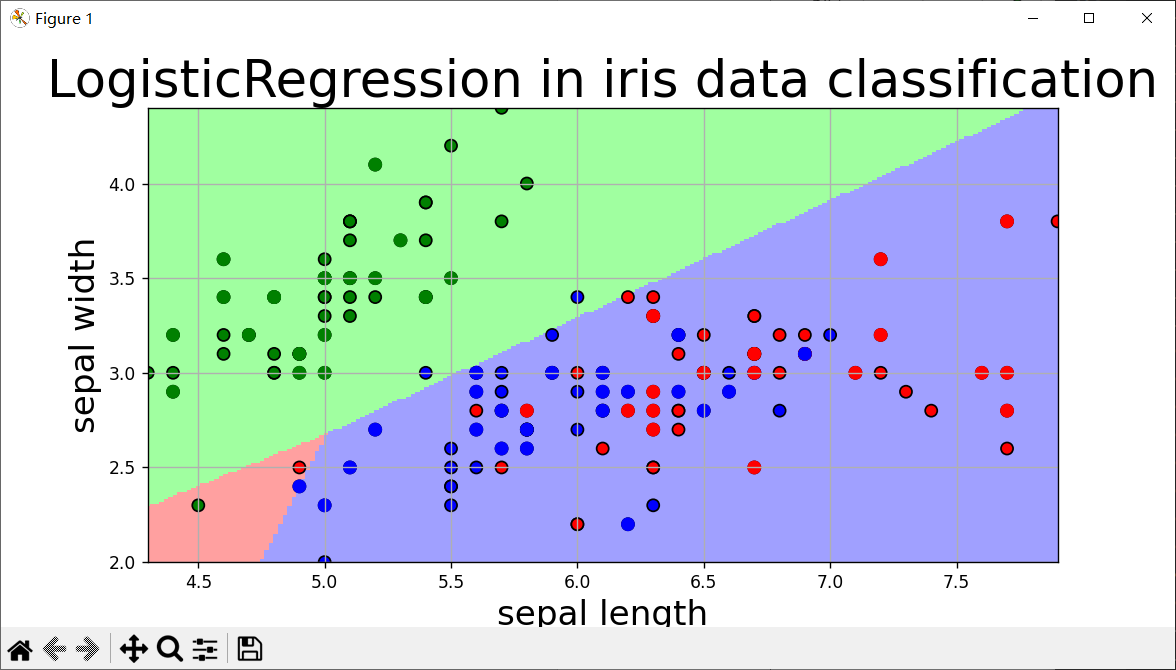
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 聚类数目 | 输入特征个数 | 输入特征名称 | 轮廓系数 | 正确率Acc | F1分数  f1\_macro | 查准率  precision\_macro | 查全率recall\_macro |
| 3 | 2 | 萼片长度萼片宽度 | 0.445 | 0.820 | 0.8198 | 0.513 | 0.513 |
| 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.553 | 0.893 | 0.892 | 0.075 | 0.093 |
| 2 | 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.681 | 0.660 | 0.547 | 0.476 | 0.64 |
| 4 | 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 | 0.498 |  |  |  |  |

附：

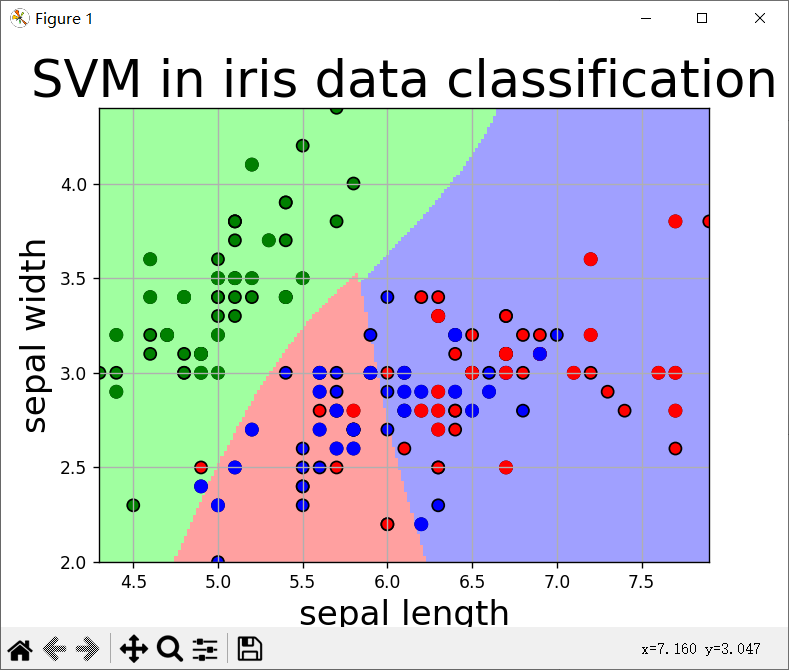
1. 原始数据图



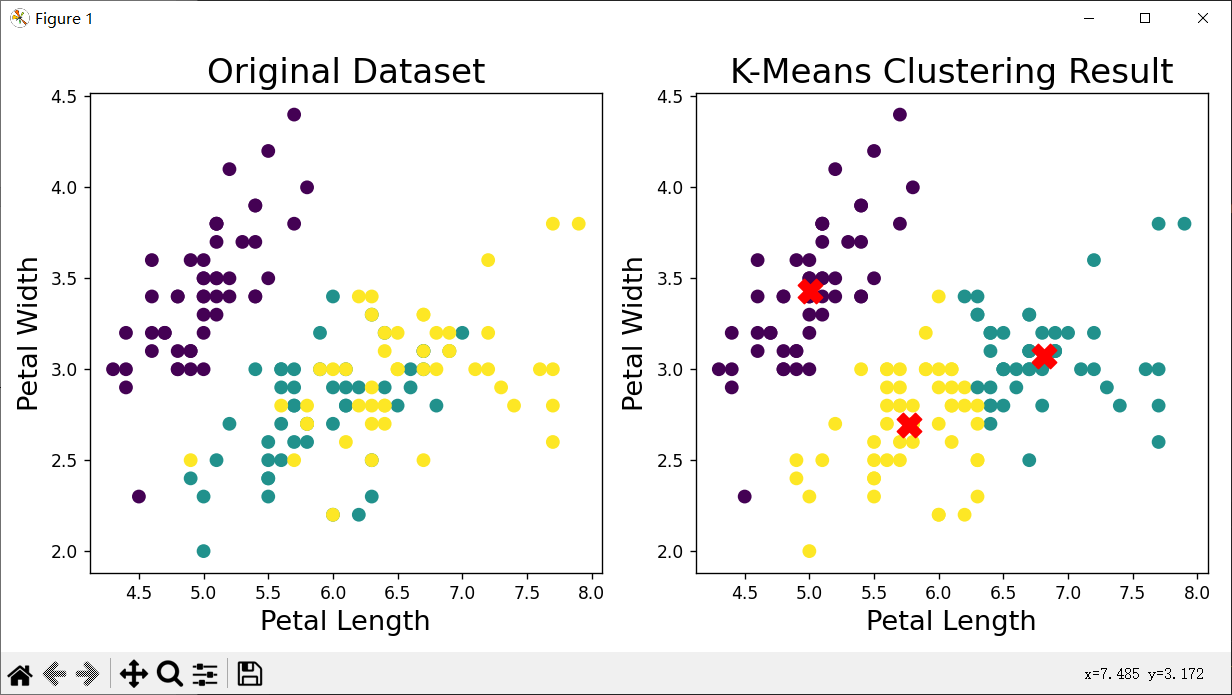
2. 逻辑回归分类图



1. SVM分类图



1. K-Means分类图



## 实验总结

1. 增加输入特征数量能有效提高模型的准确率以及F1分数、精确率、召回率
2. 在输入特征相同时，使用线性核的支持向量机模型的性能要略好于使用高斯核的支持向量机模型，使用高斯核的支持向量机模型的性能要略好于逻辑回归模型
3. 网格搜索可以有效缩短寻找最优超参数耗费的时间，可以有效提高模型的训练效率；超参数C控制惩罚力度，如果过大容易导致过拟合，过小容易导致欠拟合；gamma定义高斯核的宽度，过大容易导致过拟合，过小容易导致欠拟合；tol定义模型的停止标准，小的tol会让模型精细地迭代，但是过小会使训练时间变长，大的tol会让模型快速迭代，但是过大就会使得模型精确度较低；penalty定义正则化类型，L1类型会起到特征选择的作用，L2类型会使权重均匀分布
4. 越小的聚类数目一般会得到性能更加优秀的模型
5. 使用SVM算法进行多类别花卉分类。SVM在高维空间中找到最优的决策边界，适用于处理多特征的花卉图像数据。将训练好的SVM模型集成到APP中，可以让用户拍摄花卉照片并进行种类识别。

我的感想：

通过这次实验，我深入理解了逻辑回归、支持向量机和K-means聚类等常用机器学习方法的原理和应用。逻辑回归是一种处理分类问题的有效工具，实现了二元分类和多元分类。支持向量机是一种分类算法，也是一种寻找最优决策边界的策略。同时，K-means聚类算法则帮助我理解了无监督学习的概念。