**实验1 机器学习实验**

**一、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。

**二、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。

1. 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。

**三、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。

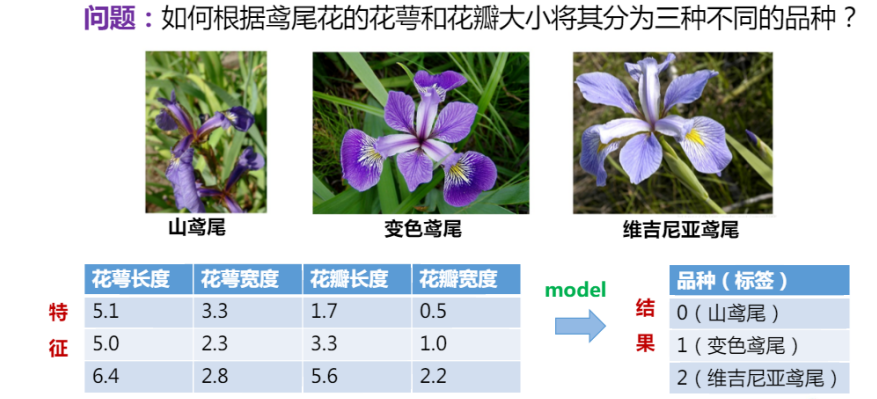


图1鸢尾花的特征和结果示例

　　在三个类别中，其中有一个类别和其他两个类别是线性可分的。另外。在sklearn中已内置了此数据集，总共包含150行数据，每一行数据由 4 个特征值及一个目标值组成，如下图2所示。4 个特征值分别为：萼片长度、萼片宽度、花瓣长度、花瓣宽度；目标值为三种不同类别的鸢尾花，分别为： Iris Setosa、Iris Versicolour、Iris Virginica。

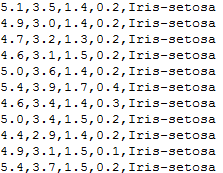


图2鸢尾花的数据集格式示例

**四、实验内容：**

1.基于逻辑回归的鸢尾花分类

参考核心代码，利用sklearn库（https://www.showmeai.tech/article-detail/203）等完成基于逻辑回归的鸢尾花分类，要求按照准备数据（训练集：测试集=6:4）、配置模型、训练模型、模型评估和模型预测5个步骤完成分类，如下图3所示，然后把不同输入特征个数下逻辑回归的分类结果填入表1，并给出只输入萼片长度、萼片宽度2个特征的分类区域图（包括训练数据点、测试数据点和分类边界）。

备注：

1）sklearn库中逻辑回归LogisticRegression()的参数见<https://blog.csdn.net/weixin_46411214/article/details/110090552>

2）有关F1分数f1\_macro,精确率/查准率precision\_macro ,召回率/查全率recall\_macro的说明见https://www.jianshu.com/p/42d5d1121b01

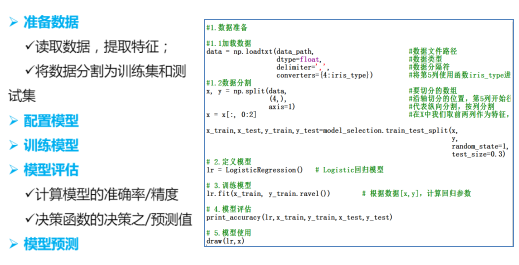


图3基于逻辑回归的鸢尾花分类步骤

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 |

2.基于支持向量机（SVM）的鸢尾花分类

参考核心代码，利用sklearn库等完成基于支持向量机的鸢尾花分类，要求按照准备数据、配置模型、训练模型、模型评估和模型预测5个步骤完成分类如下图4所示，然后把不同输入特征个数下支持向量机的分类结果填入表1，并给出只输入萼片长度、萼片宽度2个特征的分类区域图（包括训练数据点、测试数据点和分类边界）。

分别对线性核和高斯核的SVM，采用网格搜索确定最佳参数，结果填入表2.

备注：

1. sklearn.svm.SVC 参数说明：<https://blog.csdn.net/szlcw1/article/details/52336824>
2. 网格搜索详见https://zhuanlan.zhihu.com/p/642060346.

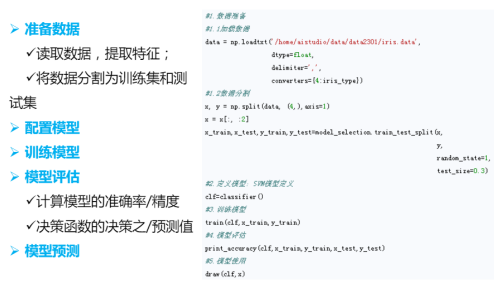


图4基于支持向量机的鸢尾花分类步骤

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 |  | 0.774 | 0.8 | 0.85 | 0.8533, 0.8725, 0.8655 |
| 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 |  |  |  |  |  |
| 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 |  |  |  |  |  |

3、基于K-means的鸢尾花聚类

参考下图5和标签映射算法Kuhn-Munkre，利用sklearn库应用K-means聚类算法实现对Iris数据集进行聚类，要求按照图5所示准备数据、配置模型、训练模型、评估模型和输出聚类结果5个步骤完成聚类，然后把不同输入特征个数和不同初始聚类数下的聚类结果填入表3，要求给出原数据集的截图、不同聚类结果的截图。

备注：

1）sklearn库中K-means聚类算法的参数见

<https://blog.csdn.net/qq_41645987/article/details/100045906>

2）有关F1分数f1\_macro,精确率precision\_macro ,召回率recall\_macro见https://www.jianshu.com/p/42d5d1121b01

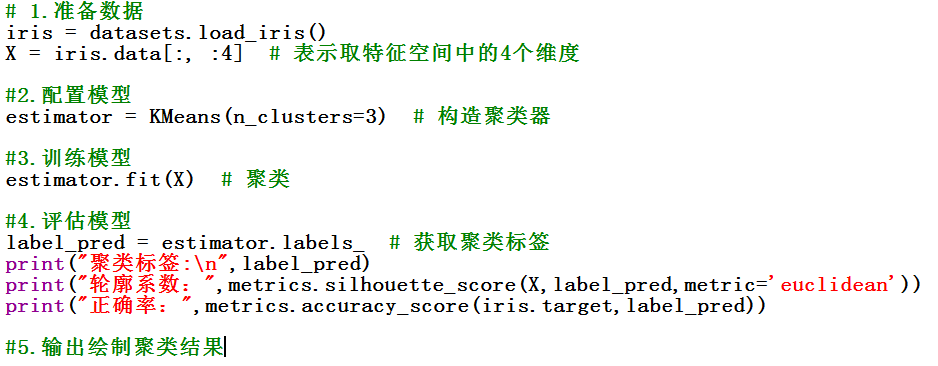


图5基于k-均值的鸢尾花聚类步骤

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 |  |  |  |  |  |
| 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 |  |  |  |  |  |
| 2 | 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 |  |  |  |  |  |
| 4 | 4 | 萼片长度萼片宽度花瓣长度花瓣宽度 |  |  |  |  |  |

4、提交实验报告和源程序。

**五、实验报告要求：**

1、根据表1和表3鸢尾花实验的分类和聚类结果，分析输入特征个数对分类算法、聚类算法的性能影响。

2、根据表2鸢尾花实验的分类结果，分析输入相同特征时不同分类模型的性能。

3、根据表2鸢尾花分类实验结果，包括线性核和高斯核的支持向量机、逻辑回归的不同超参交叉验证，请分析网格搜索的作用，以及不同超参对分类模型性能的影响。

4、根据表3不同聚类数目的聚类结果，分析聚类数目对K-means算法性能的影响。

5、结合“形色”花卉识别等一些APP，谈谈基于机器学习的花卉分类或聚类算法如何进行推广应用。

下面是实验报告的基本内容和书写格式。

实验名称

班级： 学号： 姓名：

一、实验目的

二、实验原理

三、实验条件

四、实验内容

按照实验内容，记录实验结果。

五、实验总结

首先，根据实验报告要求，给出相应的总结分析。最后，给出本次实验的心得体会。

——————————————————————————————————