实验报告

东北林业大学

信息与计算机科学技术实验中心

|  |
| --- |
| 一、实验目的  （1）理解KNN算法的数学原理，能够利用Sklearn中KNN分类器解决现实世界中的分类问题；  （2）掌握利用matplotlib对分类结果进行可视化显示的方法，并分析模型的优劣；  （3）掌握利用Sklearn进行分类模型模型评价的方法。 |
| 二、实验环境  （1）硬件：PC机；  （2）软件：Anaconda Jupyter Notebook，Spyder，Pandas |
| 三、实验内容及结果  （一）基于KNN和随机森林分类算法的森林类型分类  数据集说明：  1-Spruce/Fir：云杉/冷杉  2-Lodgepole Pine：黑松，海滩松  3-Ponderosa Pine：美国黄松  4-Cottonwood/Willow：杨木/杨柳  5-Aspen:(欧洲)山杨;大齿杨  6-Douglas-fir：花旗松,北美西北部的主要树种之一。  7-Krummholz：高山矮曲林，矮盘灌丛  1、导入相关库    2、打开数据集，分割X和y数据集，并进行数据探索。    3、分割测试集和训练集  4、利用KNN模型进行森林类型预测。  5、评估模型精度，从中可以看出测试集上k=5时KNN模型的精度为82.3%，从分类报告中可以看出，1和2类型分类精度相对较低，对应云杉/冷杉、Lodgepole Pine和Ponderosa Pine的分类精度相对较低。  6、显示分类结果的混淆矩阵。从中可以看出1、2之间错分的情况最多，查看数据说明，云杉/冷杉和黑松区分度相对较差。  （二）利用水色图像数据集进行水质分类。  1、导入相关库。  2、获取图像文件列表  3、查看部分图像  （部分）  4、查看某一个图像的形状和直方图。  5、定义一个函数，求一个波段数据的一阶矩，二阶矩和三阶矩，方便使用。  6、提取所有图像的信息。  7、保存df数据到csv文件中，方便下次使用。  8、如果实验中途打断，那么之前的图像处理部分将不需要重复进行，只需要打开题7中保存的文件即可。如果没有中途打断，则直接使用上面的df对象即可，跳过本题。  9、分割X、y数据集。  10、分割训练集和测试集，测试集比例用0.2。  11、利用KNN建立模型，并利用X测试集预测结果。  12、利用SVM建立模型，并利用X测试集预测结果。  13、利用测试集评估每个模型的准确率。从中可以看出KNN模型的准确率为76.75%，SVM模型的准确率为67.82%。  14、输出两个模型的混淆矩阵，并分析模型的优劣。    15、可以再选择其他的分类模型，如随机森林来测试模型的优劣。  从中可以看出，随机森林模型的准确率为84.15%，效果较KNN和SVM都要好。当前应该最广泛的分类算法就是随机森林，它对于常规问题的分类效率通常要优于其他分类。 |

|  |
| --- |
| 四、实验过程分析与讨论  1． 实验过程分析与讨论：  （1）使用train\_test\_split函数将数据集划分为训练集和测试集。K最近邻分类器：使用sklearn库的KNeighborsClassifier类创建K最近邻分类器，并使用fit函数将训练集拟合到分类器上。  （2）使用score函数计算分类器在测试集上的准确率。使用classification\_report和confusion\_matrix函数生成分类报告和混淆矩阵，用于评估分类器的性能。使用PIL库的Image类读取图像文件，并使用matplotlib库的imshow函数显示图像。  （3）图像特征提取定义了一个名为info的函数，用于计算图像的一阶矩、二阶矩和三阶矩。使用pandas库的DataFrame类创建一个空的数据框，并使用loc函数将图像特征添加到数据框中。使用to\_csv函数将数据框保存为CSV文件。  （5）支持向量机分类器使用sklearn库的SVC类创建支持向量机分类器，并使用fit函数将训练集拟合到分类器上。随机森林分类器使用sklearn库的RandomForestClassifier类创建随机森林分类器，并使用fit函数将训练集拟合到分类器上。  （6）使用accuracy\_score函数计算分类器在测试集上的准确率。使用confusion\_matrix函数生成混淆矩阵，用于评估分类器的性能。  （7）在plt中可以利用plt.show()函数来显示图像。 |
| 五、指导教师意见    指导教师签字：  年 月 日 |