广东东软学院

学生实验报告

**实验课程名称：机器学习**

**实验项目名称：树叶分类**

**实验类型：实践**

**指导教师：苏康**

**实验日期：2019年5月30日**

|  |  |  |  |
| --- | --- | --- | --- |
| **学生姓名** | **李光宇** | **学 号** | **16210121035** |
| **班 级** | **10班** | **专业名称** | **软件工程** |
| **实验组**  **其他成员** | **黄楚** | | |
| **实验地点** | **教室** | | |
| **实验成绩**  **（教师签名）** |  | | |

|  |
| --- |
| **实验目的与要求**  **目的:**  综合运用《机器学习》课程中的分类或聚类算法对树叶数据集进行筛选，把同种类别的树叶分开。要求学会掌握Python文件操作、特征工程、调用sklearn算法包、调整算法参数和数据可视化等过程。最终要求基本掌握所运用算法的原理，并分析比较各种算法分类（聚类）的好坏。  **实验原理与内容:**   1. 要求结合相关性矩阵进行特征工程 2. 对原始数据集的192维数据PCA降维，降成m维后可以解释95%以上的数据 3. 对train.csv进行划分（80%用于训练，20%用作测试），要求训练集上的准确率不能低于80% 4. 采用4种以上的分类算法，对分别对PCA降维和没降维数据进行训练，采用网格搜索等方法调整训练模型参数，最终使得要求训练集上的准确率不能低于80%，并分析比较模型测试集上的准确率。 5. 预测test.csv内的树叶类型，查看在filtered\_imgs下生成99个文件夹，其中4和7文件夹内容如下：   人工数出分类错误的树叶总数。  如4文件夹下有1个错误情形，7下为0，要求数出0-98文件夹下所有错误的树叶个数，并记录下它们的图片ID   1. 汇总train.csv和test.csv，采用3种以上的聚类算法，对汇总的数据进行聚类，再完成步骤5的人工数出分类错误的树叶总数过程。 |
|  |
| **实验设备与软件环境**   1. 操作系统：Windows10 2. 硬件平台：Jupyter 3. 使用语言：python |
| **实验过程与结果（可贴图）**   1. 了解数据   选定作业区域，对工作区域进行了解，导入需要用到的工具包，了解接下来进行操作数据概况。    2、数据预处理  进一步了解train.csv与test.csv两个文件的数据结构和内容，并且将数据处理为方便操作的结构。    3、特征工程      4、机器学习  调参及评价模型  C:\Users\Administrator\AppData\Roaming\Tencent\Users\876305663\QQ\WinTemp\RichOle\FPU96E)PKD)UTJ{2I8@GJ(M.png    5、预测未知树叶  C:\Users\Administrator\AppData\Roaming\Tencent\Users\876305663\QQ\WinTemp\RichOle\(HY1$(10V9404]%A3NN7$29.png |
| **操作异常问题与解决方案**   * 使用PCA对数据进行降维时，要将所有数据都进行降维 * PCA降维与归一化的先后顺序对于结果的影响。 |
| **实验总结**  通过期末设计，我进一步熟悉csv文件处理以及常用了解文件的途径。对于进行大数据机器学习前的数据预处理有了更深入得了解。通过对机器学习算法的运用及特征工程的配合使用，熟悉了机器学习算法的运用与结果的差异，了解了特征工程对于不同算法的不同影响。 |