《机器学习》上机实验指导书

实验2：决策树与随机森林

**实验目的：**

1. 加深对决策树与随机森林分类思想的熟悉，基本掌握scikit-learn库中相关模型的参数、属性与内部函数的含义与使用；
2. 通过设置决策树与随机森林的不同算法参数和数据，观察分类效果，根据所学进行分析。

**实验内容：**

**（1）修改blobs数据的生成参数，熟悉和修改决策树的主要参数，观察并报告分类结果**

根据实验提供的代码文件（05.08-Random-Forests-Excercise.ipynb）开展实验：

* 1. 使用jupyter运行代码，修改make\_blobs()中的类别数（centers=2、4、6），观察决策树分类结果（可以附加修改生成的样本个数n\_samples，以及random\_state等值，进一步了解make\_blobs函数的使用）；
  2. 借助在线文档，通过print(tree.get\_params())了解决策树模型的主要参数的意义及其缺省值（需在报告中写出下一条中列出的几个主要参数的含义与作用）；
  3. 通过初始化修改诸如max\_depth，min\_samples\_split，min\_samples\_leaf，max\_leaf\_nodes等参数（合理组合不同取值），观察并报告分类结果。

**（2）修改随机森林的初始化参数，观察并报告分类结果**

根据实验提供的代码文件（05.08-Random-Forests-Excercise.ipynb）开展实验：

1. 借助在线文档，通过print(model.get\_params())了解决策树模型的主要参数的意义及其缺省值（需在报告中写出下一条中列出的几个主要参数的含义与作用）；
2. 通过初始化修改诸如n\_estimators，criterion，min\_weight\_fraction\_leaf，max\_features，bootstrap，random\_state，max\_samples等参数（合理组合不同取值），观察并报告分类结果；
3. 结合代码中的手写数字识别实验，修改上述主要参数值，观察识别准确率等结果，以及生成的树与森林的关键属性值（单个树的深度，森林中树的个数，是否随机化），写出简要分析报告；
4. 有余力的同学，可以熟悉train\_test\_split()，对训练数据与测试数据进行进一步调整，观察对识别效果的影响。

**参考资料：**

**scikit-learn user Guide, Release 0.23.1（May 19, 2020）.pdf**

**决策树：sklearn.tree.DecisionTreeClassifier**

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.get_depth>

**随机森林：sklearn.ensemble.RandomForestClassifier**

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=randomforestclassifier#sklearn.ensemble.RandomForestClassifier>

**训练与测试数据集准备：sklearn.model\_selection.train\_test\_split**

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html?highlight=train_test_split#sklearn.model_selection.train_test_split>