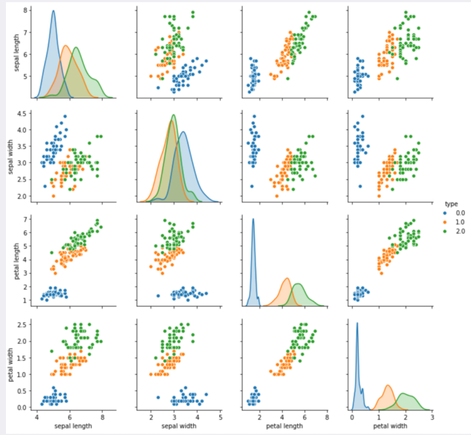
Iris文档

- 类
  - 1.数据预处理类
    - 输入
      - X特征值
      - Y目标值
    - 功能
      - 训练测试集分割
      - 数据标准化
      - label转换为one hot编码
      - one hot编码转换为label
      - 简单的可视化，观察数据情况
    - 用法
      - 输入X和Y初始化对象，通过调用对象的方法完成处理和可视化。处理后的数据变成的对象的属性，可以直接调用
  - 2.参数调优类
    - 功能
      - 画出指定超参数的验证曲线，选择合适的超参数值
      - 在训练集上进行5折GridsearchCV，返回最佳参数组合
  - 3.模型评估类
    - 功能
      - 画ROC曲线（对能够输出预测概率值的分类器）
      - 画混淆矩阵（根据预测值和真实值）
      - 画学习曲线，看模型的偏差和方差
      - .在数据集上得到交叉验证的分数
- 流程
  - 1.将iris数据集输入数据预处理类，完成预处理，与初步的可视化观察



  - 尝试不同分类模型
    - 1.SVM
      - 参数调优
        - gamma验证曲线
        - C验证曲线
        - 得到gamma, C最佳参数组合
      - 模型验证
        - 1.画模型在测试集上的ROC曲线
        - 2.画模型在测试集上的混淆矩阵
        - 3.画模型在训练集上的学习曲线
        - 4.在整个数据集上进行5折交叉验证，交叉验证分数：0.9733333333333334 (std:0.03265986323710903)
    - 2.XGB
      - 参数调优
        - learning_rate验证曲线
        - 得到n_estimators, max_depth, learning_rate最佳参数组合
      - 模型验证
        - 1.画模型在测试集上的混淆矩阵
        - 2.画模型在训练集上的学习曲线
        - 3.在整个数据集上进行5折交叉验证，交叉验证分数：0.9266666666666665 (std:0.05333333333333332)
    - 3.LogisticRegression
      - 参数调优
        - C验证曲线
        - 得到最佳参数组合
      - 模型验证
        - 1.画模型在测试集上的ROC曲线
        - 2.画模型在测试集上的混淆矩阵
        - 3.画模型在训练集上的学习曲线
        - 4.在整个数据集上进行5折交叉验证，交叉验证分数：0.9733333333333334 (std:0.03265986323710903)
    - 4.并联以上三种模型的模型
      - 把三个模型并联在一起，在整个数据集上进行5折交叉验证，交叉验证分数
        - 在整个数据集上进行5折交叉验证，交叉验证分数：0.9666666666666666 (std:0.04216370213557838)
        - avg_clf将先前训练的3个模型并联在一起，将对同一个样本做出各自判断，形成三个结果，最后少数服从多数，如 [2 1 2]说明第一个模型预测的是2，第二个模型预测的是1，第三个模型预测的是2，那么最终结果是2
          从最终结果来看，集成模型在整个训练集上的表现并没有SVM和逻辑回归好，但是比其中的成员XGB有所提升