python机器学习笔记

方法

1. one-hot encoding：用于将离散特征转换为无差别的标记，非重复个数为向量维数，某位为1，其余为0
2. 分类的准确率、精度、召回率和F1得分
    - 准确率：对于给定的测试数据集，分类器正确分类的样本数与总样本数之比，(TP+TN)/Total
    - 精度：预测正类预测正确的样本数，占预测是正类的样本数的比例 TP/(TP+FP)
    - 召回率：预测正类预测正确的样本数，占实际是正类的样本数的比例 TP/(TP+FN)
    - F1得分：精度和召回率的调和均值
3. 混淆矩阵

程序

1. sys.arg[]是一个从程序外部获取参数的桥梁，0代表本程序本身，1，2，及其他代表在命令行中输入的参数的第1，2个
2. pickle包保存训练后的模型为.pkl文件，记得用'wb'写入和'rb'读入
3. 在交叉验证时选取不同的评分scoring
4. 随机森林在ensemble里面
5. 验证曲线优化超参
6. 学习曲线理解训练集大小的影响
7. numpy中axis使用0值表示沿着每一列或行标签\索引值向下执行方法,使用1值表示沿着每一行或者列标签模向执行对应的方法
8. pdb模块可以用于python调试
9. SVM中RBF核函数对非线性很好
10. SVM中解决不同类数据不平衡问题：class_weight='auto'