K折交叉验证和网格搜索

K折交叉验证

- 将数据集平均分成不相交的K个子集
- 留一份作为测试集，剩下的为训练集
- 重复上面步骤K次
- 将K次的结果取均值，作为验证结果
- 交叉验证带来一定的计算代价，尤其是当数据集很大的时候，导致计算过程会变得很慢

网格搜索

对模型预设几种超参数组合，每组超参数排列组合采用交叉验证来进行评估。最后选出最优参数组合建立模型。

注意

gridsearch很慢时，可以试试RandomizedSearchCV，但它只适合于服从连续分布的参数

sklearn.model_selection.GridSearchCV参数

- estimator:选择使用的分类器
- param_grid:需要最优化的参数的取值，值为字典或者列表
- scoring :准确度评价标准，默认None,这时需要使用score函数
- cv :交叉验证参数，默认None，使用三折交叉验证
- refit :默认为True,程序将会以交叉验证训练集得到的最佳参数，重新对所有可用的训练集与开发集进行，作为最终用于性能评估的最佳模型参数。
- iid:默认True,为True时，默认为各个样本fold概率分布一致，误差估计为所有样本之和，而非各个fold的平均。
- verbose：日志冗长度，0：不输出训练过程，1：偶尔输出，>1：对每个子模型都输出。
- n_jobs: 并行数。-1：跟CPU核数一致
- pre_dispatch：指定总共分发的并行任务数。当n_jobs大于1时，数据将在每个运行点进行复制，这可能导致OOM，而设置pre_dispatch参数，则可以预先划分总共的job数量，使数据最多被复制pre_dispatch次

sklearn.model_selection.KFold参数

- n_splits:表示样本集划分几等份
- shuffle：false每次运行得到的结果一样，true每次都随机选，默认false