《机器学习》第三次作业

黄磊 计702 2022E8013282156

1. 分析随机森林为何比决策树Bagging集成的训练速度更快

Bagging在选择划分属性时需要考察结点所有属性，而随机森林只随机考察一个属性子集。随机森林可以看作Bagging算法的一个扩展变体。比Bagging速度更快，泛化能力更强。

2. 试比较Gradient Boosting和Adaboost的异同

异：	GB	Ada
更新方式：	学习数据中的噪声和异常值更新参数	学习模型输出和标签之间误差更新参数。
模型结构：	包括多层感知机和自编码器，可以使用多种不同特征学习不同问题	通常是决策树和随机森林。
数据增强：	通过随机平移，旋转，缩放和翻转等操作来扩充数据集，以提升泛化性能。	通过添加异常值来扩充数据集，以帮助模型识别异常和错误预测。

优点：对噪声和异常值的鲁棒性、模型泛化能力、复杂问题处理能力　模型简单，易于实现，能处理多种特征表示。

缺点：训练时间复杂度高，数据要求较高　可解释性差，需要大量数据增强。

同：都是基于树形结构的分类模型，采用了数据增强、迭代计算、梯度计算、能处理多种特征表示。

3. 试比较包裹式选择、过滤式选择与嵌入式选择的异同:

异：

	包裹式	过滤式	嵌入式
构建方式：	树状结构分为多个小独立子集	在树状基础上加入特征过滤低置信度特征。	将树状结构嵌入到具体任务中。
输出：	所有子集最大值。	子集中概率最大值	当前节点推荐值。

同：都基于决策树模型，输出是决策树或随森林，采用了数据增强、迭代计算、都关注当前节点置信概率和整个树形结构。

4. 试述直接求解 L0 范数正则化会遇到的困难

① L_0通常为指数函数，数值稳定性欠佳；

② 收敛很慢.

③ 正则化效果差，可能过拟合.

④ 在参数空间不单调，可能导致泛化性能差.

5. 试述为什么基于 L1 范数可以进行特征选择

L_1相当于将模型空间限制在L_1球上，目标函数L等高线有很大概率与坐标轴边相交，解具有稀疏性。

基于L_1的特征选择不能直接设置最终选择特征个数K，通过正则化系数隐式控制λ，λ越大，越具有稀疏性，得到非零系数个数越少.

6. 试比较 K-SVD 与 K-means 方法的异同;

同：都是聚类算法.

异：

K-SVD

基于奇异值

按保留的奇异值的权重排序，构建聚类中心

K-means ← 更简单，更易实现，

基于距离.

根据每个数据点与聚类中心的距离计算聚类中心点.