梯度提升树

回顾：随机森林

Random Forest（随机森林）是 Bagging 的扩展变体，它在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入了随机特征选择，因此可以概括 RF 包括四个部分：

1. 随机选择样本（放回抽样）；
2. 随机选择特征；
3. 构建决策树；
4. 随机森林投票（平均）。

随机选择样本和 Bagging 相同，采用的是 Bootstrap 自助采样法；**随机选择特征是指在每个节点在分裂过程中都是随机选择特征的**（区别与每棵树随机选择一批特征）。

这种随机性导致随机森林的偏差会有稍微的增加（相比于单棵不随机树），但是由于随机森林的“平均”特性，会使得它的方差减小，而且方差的减小补偿了偏差的增大，因此总体而言是更好的模型。

随机采样由于引入了两种采样方法保证了随机性，所以每棵树都是最大可能的进行生长就算不剪枝也不会出现过拟合。

优点

1. 在数据集上表现良好，相对于其他算法有较大的优势
2. 易于并行化，在大数据集上有很大的优势；
3. 能够处理高维度数据，不用做特征选择。

知识点

- 通俗易懂理解——Adaboost算法原理

  AdaBoost（Adaptive Boosting，自适应增强），其自适应在于：**前一个基本分类器分错的样本会得到加强，加权后的全体样本再次被用来训练下一个基本分类器。同时，在每一轮中加入一个新的弱分类器，直到达到某个预定的足够小的错误率或达到预先指定的最大迭代次数。**

  Adaboost 迭代算法有三步：

  1. 初始化训练样本的权值分布，每个样本具有相同权重；
  2. 训练弱分类器，如果样本分类正确，则在构造下一个训练集中，它的权值就会被降低；反之提高。用更新过的样本集去训练下一个分类器；
  3. 将所有弱分类组合成强分类器，各个弱分类器的训练过程结束后，加大分类误差率小的弱分类器的权重，降低分类误差率大的弱分类器的权重。
- GBDT原理
- [拓展阅读:Random Forest、Adaboost、GBDT]
- XGBoost、LightGBM原理
- XGBoost、LightGBM对比

QA

1. 介绍一下GBDT；
2. xgboost有哪些改进？
3. GBDT与随机森林的异同点？
4. xgb防止过拟合有什么方法，如何调参？
5. xgb为什么对缺失值不敏感，如何处理缺失值的？
6. 解释一下GBDT沿着梯度下降方向提升，如何实现的？

项目

Titanic: Machine Learning from Disaster

采用lightGBM\XGBoost\GBDT进行建模，参考案例Microsoft LightGBM with parameter tuning (~0.823)；

尝试在案例基础上新增特征，提高score。

RefS

- Adaboost 论文
- Xgboost 论文
- Boosting 与 AdaBoost
- XGBoost
- [校招-基础算法]GBDT/XGBoost常见问题