

8. bagging与boost对比

1. boosting与 bagging 区别

1) ~~样本选择上~~ Bagging: 训练集在原始集中有放回地选取, 从原始集中选出的各轮训练集之间是独立的。

Boosting: 每一轮的训练集不变, 只是训练集中每个样例在分类器中的权重发生变化。而权重是根据上一轮的分类结果进行调整。

2) ~~样例权重~~ Bagging: 使用均匀抽样, 每个样例的权重相等。

Boosting: 根据错误率不断调整样本权重, 错误率越大, 权重越大

3) ~~预测函数~~ Bagging: 所有预测函数的权重相等。

Boosting: 每个弱分类器都有相等权重。
对于分类误差小的分类器会有更大的权重。

4) ~~并行性~~ Bagging: 各个预测函数可以并行

Boosting: 只能顺序进行, 因为后一个模型需要前一轮模型的结果

B. 随机森林与GBDT区别:

1) 训练集的选取: 随机森林采用 Bagging 思想, GBDT 采用 Boosting 思想;

这两种方法都是 Bootstrap 思想的应用。

(从个数的数据集中有放回的抽取 N 次, 每次抽 n 个)

但 Bagging 是有放回的均匀抽样

Boosting 根据错误率来取样

2) 组成随机森林的树可以是 分类树, 也可以是 回归树。

GBDT 只能由 回归树 组成

3) 组成随机森林的树可以并行生成; 而 GBDT 只能串行生成

4) 对于最终的输出结果, 随机森林采用投票法:

GBDT 采用加权累加,

5) RF 对异常值不敏感, GBDT 对异常值敏感

6) 随机森林对训练集一视同仁; GBDT 是基于权值的的弱器的集成

7) RF 是通过减少模型来提高性能; GBDT 是通过减少模型偏差提高性能