第5章 决策树

Exercise 5.15

决策树的父节点和子节点的熵的大小关系是什么?

A、两者相等
B、子节点的熵更大
C、父节点的熵更大
D、不确定

Exercise 5.16

考虑表中二元分类问题的训练样本集

实例	a1	a2	a3	目标类
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

整个训练样本集关于类属性的熵是多少?关于这些训练集，a1，a2，a3的信息增益分别是多少?

Exercise 5.17

关于集成学习以下说法正确的是?

A、Adaboost相对于单个弱分类器而言通过Boosting增大了模型的Bias
B、随机森林相对于单个决策树而言通过Bagging增大了模型的Variance
C、我们可以借鉴类似Bagging的思想对GBDT模型进行一定的改进，例如每个分裂节点只考虑某个随机的特征子集或者每棵树只考虑某个随机的样本子集这两个方案都是可行的
D、GBDT模型无法在树维度通过并行提速，因为基于残差的训练方式导致第i棵树的训练依赖于前$i-1$棵树的结果，故树与树之间只能串行

实践题

Exercise 5.1

客户离网率预测

由于互联网企业OTT业务的兴起，运营商在传统业务上正在逐渐衰退，用户价值逐渐下降，如何利用大量用户数据进行离网分析及预测，成为了各大运营商的共识。

现我们有一个用户历史使用数据集，

1、请使用以决策树作为基学习器的随机森林模型来预测客户是否离网（流失）。

2、请使用boosting的算法来预测客户是否流失

3、对比随机森林模型和boosting模型，在训练集上学习的效果。（可以从模型的偏差和方差对比）

作业提交要求： 代码、在测试集上的结果和3）问中提到的结论。

数据集属性说明：

State：状态

Account length: 账号存在时间

Area code： 客户所属区域

International plan： 国际长途

Voice mail plan： 语音邮箱

Number vmail messages： 语音邮箱信息数

Total day minutes：日间通话总时长

Total day calls： 日间通话总次数

Total day charge： 日间服务总收费

Total eve minutes： 夜间通话总时长

Total eve calls: 夜间通话总次数

Total eve charge： 夜间服务费合计

Total night minutes：夜间通话总持续时间

Total night calls： 夜间呼叫总数

Total night charge： 夜间服务费合计

Total intl minutes : 国际长途总时长

Total intl calls : 国际长途电话总次数

Total intl charge : 国际长途总费用

Customer service calls : 客户服务电话数

Churn : 客户是否流失

数据集地址：

https://github.com/jjw12345/machine_learning/tree/main/%E9%9B%86%E6%88%90%E5%AD%A6%E4%B9%A0