决策树

知识点

熵：表示随机变量的不确定性。

条件熵：在一个条件下，随机变量的不确定性。

信息增益：熵 - 条件熵。表示在一个条件下，信息不确定性减少的程度。

通俗地讲，X(明天下雨)是一个随机变量，X的熵可以算出来， Y(明天阴天)也是随机变量，在阴天情况下下雨的信息熵我们如果也知道的话（此处需要知道其联合概率分布或是通过数据估计）即是条件熵。

X的熵减去Y条件下X的熵，就是信息增益。具体解释：原本明天下雨的信息熵是2，条件熵是0.01（因为如果知道明天是阴天，那么下雨的概率很大，信息量少），这样相减后为1.99。在获得阴天这个信息后，下雨信息不确定性减少了1.99，不确定减少了很多，所以信息增益大。也就是说，阴天这个信息对明天下午这一推断来说非常重要。

所以在特征选择的时候常常用信息增益，如果IG（信息增益大）的话那么这个特征对于分类来说很关键，决策树就是这样来找特征的。

- 信息熵
- 条件熵
- 信息增益
- 决策树认识
- 基于信息与信息增益的ID3及C4.5决策树
- 基尼指数（基尼不纯度）

  基尼指数是信息熵的1阶泰勒展开；
- CART树
- bagging模型集成与随机森岭
- 随机森林参数介绍
- 拓展阅读：模型融合
- 拓展阅读：预测偏差、方差与模型融合

QA

1. 采用信息增益、信息增益率作为决策树生长策略，有什么区别；
2. 其他条件一致，对样本某变量进行单调非线性变化，是否会影响决策树生长，为什么；
3. 随机森林参数有哪些重要的参数，分别的作用试什么？
4. 多个模型预测结果做Average融合，模型间具备怎样的特点会取得更好的效果？

项目

Titanic: Machine Learning from Disaster

- 尝试用随机森林提交预测结果，并调整模型参数提升成绩；
- 尝试利用Average融合思路设计随机森林&逻辑回归融合；参考文章