Contents

1 决策树

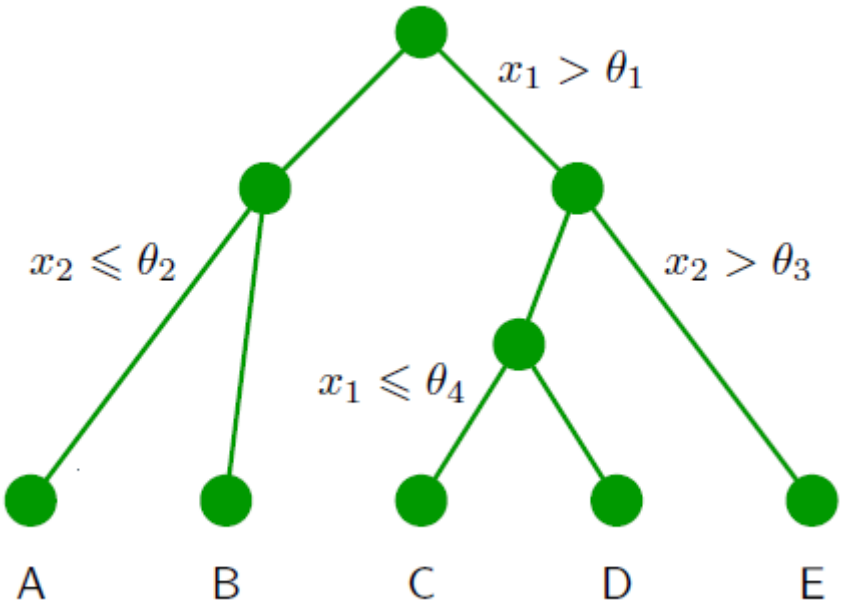

Figure 1: 图 1

1.1 思想

如上图所示，决策树每次选取单个输入（特征），将整个数据集划分成两个子集，生成一个二叉树。然后递归地对两个子集进行划分，直到达到停止条件。

停止条件选择比较多，一般是在模型的复杂度（叶子结点）和整体损失取折中的方案。

1.2 4 个问题

1. 每个结点选哪个特征？
2. 分隔阈值的选择？
3. 分隔终止条件？
4. 每个区域的预测值（回归问题）?

1.3 回归树

问题：对于 N 个输入样本 $\{x_i, y_i\}, i = 1, 2...N$，每个 x_i 包含 K 个特征，最后 N 个样本划分成 M 个子集 sub，每个子集预测值为 $p_m, m = 1, 2...M$。不妨取损失函数为均方误差，则有 （问题 **4.**）

$$p_m = avg(y_i), \quad y_i \epsilon sub_m$$

取均值时，均方误差最小。目标是确定 m 个结点所用特征 k 及其阈值 thr_m，直接根据均方误差最小化来优化，复杂度太高，无法实现。采用贪心算法，一个一个确定，每个都取最优。每一步的优化可以用下面的公式来表示，（问题 **1.**，**2.**）

$$cost_m = \min_{k,thr} \left[\min_{p_1} \sum (y_i - p_1)^2 + \min_{p_2} \sum (y_i - p_2)^2 \right]$$

具体来说就是枚举，比如说，每个特征 k 都对当前数据集确定最优的 thr，再取其中最大的 k。（不太聪明的方法 ~）

最直接的终止条件就是二叉树的叶子结点，也就是深度，本问题就是 M 的个数。

$$\min_{\alpha,M} C = \sum_{m}^{M} cost_m + \alpha|M|$$

上式极为整体损失，其中 α 为超参数，平衡模型的复杂度（叶子结点）和整体损失。（问题 **4.**）

1.4 分类树

分类树只需更换损失函数

$$\text{Misclassification error:} \quad \frac{1}{N_m} \sum_{i \in R_m} I\left(y_i \neq k(m)\right) = 1 - \hat{p}_{mk(m)}$$

$$\text{Gini index:} \quad \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}\left(1 - \hat{p}_{mk}\right)$$

$$\text{Cross-entropy or deviance:} \quad -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

一般地，Gini 指数和交叉熵对分类更敏感，用于 cost，在剪枝时候，一般用分类误差。