§4 决策树

1. 基本流程：divide-and-conquer 分治策略。使各分支结点所包含的样本尽可能属于同一类别

递归返回：(1) 当前结点包含的样本全属同一类别，无需划分 (2) 当前属性集为空，即所有样本在所有属性上取值相同，无法划分 (3) 当前结点包含的样本集为空

2. 划分选择

信息熵 (information entropy)：度量样本纯度

$$Ent(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

信息增益 (information gain)：$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$ — ID3

信息增益率 (gain ratio)：$Gain\_ratio(D,a) = \frac{Gain(D,a)}{IV(a)}$ — C4.5

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

基尼系数 (Gini index)：$Gini(D) = \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2$ — CART

$Gini\_index(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$　　$Gini(D)$ 越小，D 纯度越高

3 剪枝处理 pruning

预剪枝 (pre-pruning)：多数情况　　后剪枝 (post pruning)

决策树：ID3

优点	缺点	特点	衡量指标
速度大 简单	单变量决策树 只适用小样本 倾向于特征取值 数目较多的特征 构造非线性	只允许离散变量 构造多叉树	信息增益

C4.5

优点	缺点	特点	衡量指标
规则易于理解 准确率较高 实现简单	需进行多次扫描 和排序，效率较低 只适合小样本	多分支决策树 选信息增益率最大的 属性	信息增益率

CART

优点	特点	衡量指标
可用于分类和回归	构建的二叉树	Gini

算法总结

算法	支持模型	树结构	特征选择	连续值处理	缺失值处理	剪枝	特征属性多次使用
ID3	分类	多叉树	信息增益	✓	✗	✗	✗
C4.5	分类	多叉树	信息增益率	✓	✓	✓	✗
CART	分类、回归	二叉树	Gini、均方差	✓	✓	✓	✓

4. 多变量决策树

主要算法：OC1 —贪心地寻找每个属性上最优权值，在局部优化的基础上再对分类边界进行随机扰动以试图找到更好的边界

第i个神经元的连接权重

w_1 w_2 w_n → θ → y

n个输入信号

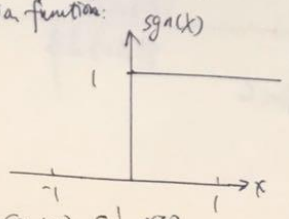
§5 神经网络

输出 $y=f(\sum_{i=1}^{n} w_i x_i - \theta)$

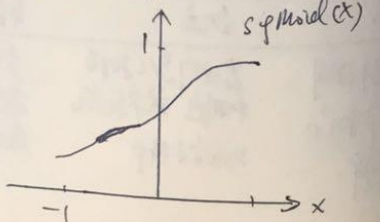
threshold logic unit

1. 感知机(perceptron)：输入层接收信号 → 输出层(M-P神经元，也称阈值逻辑单元)

activation function:



$sgn(x)=\begin{cases}1 & x\geq 0\\ 0 & x<0\end{cases}$



$sigmoid(x)=\frac{1}{1+e^{-x}}$