April 22nd 2023.

决策树

树的组成：

根节点：第一个选择点

非叶子节点与分支：中间过程

叶子节点：最终的决策结果.

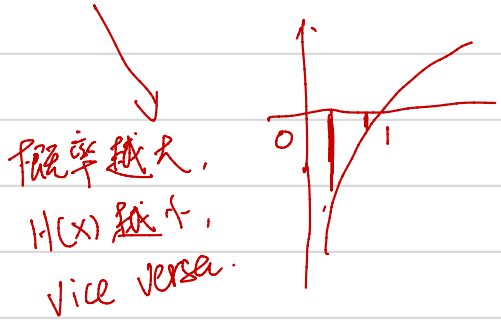
怎么要选择特征？根节点很重要. → 找到一个衡量标准.

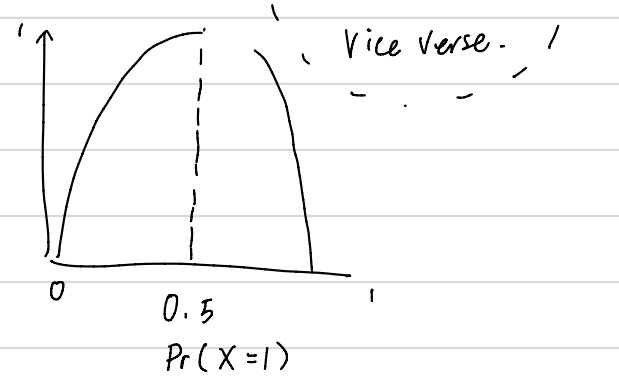
衡量标准：熵

随机变量不确定性的度量

$$H(X) = -\sum p_i \cdot \log p_i, \quad i = 1, 2, \dots, n$$

↑ 概率





9天打球，5天不打球

$$-\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

Outlook:

sunny: 0.971, 5天　　4天 Overcast: 0

rainy: 0.971, 5天

$$H(X) = \frac{5}{14}\cdot 0.971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.971 = 0.693.$$

$0.940 - 0.693 = 0.247$ ⇐ 信息增益.

ID3, C4.5　　CART (用 GINI)

$$GINI(p) = \sum_{k=1}^{k} p_k(1-p_k) = 1 - \sum_{k=1}^{k} p_k^2.$$

剪枝策略：

预剪枝：

限制深度

叶子节点个数

叶子节点样本数

信息增益量

后剪枝：

做完决策树再剪枝

$C_\alpha(T) = C(T) + \alpha \cdot |T_{leaf}|$

$0.44 \times 6 + 1 \cdot \alpha$

$0.3 + 0.4444 \times 3 + 2 \cdot \alpha$

Compare

越大越不好

回归问题：看方差