第五章：决策树

涉及到的新概念理解

熵 $H(X) = -\sum p_i \ln p_i$

基尼指数 $Gini(p) = \sum_{i=1} p_i(1-p_i)$

用来衡量随机变量的离散程度（离散）

大小取值无意义（比如 抛硬币 0-1）

都是衡量 随机变量 的离散程度（混乱程度）

- 两种衡量离散程度的方法

$T = \{x_1, \cdots, x_n\}$

离散随机变量常用（变量的大小取值无意义）

① 方差 $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

② $\sum_j (x_j - \bar{x})^2 p_j$ （p_j为取值相等的变量出现的频率）

连续随机变量常用 或 变量的大小取值有意义的

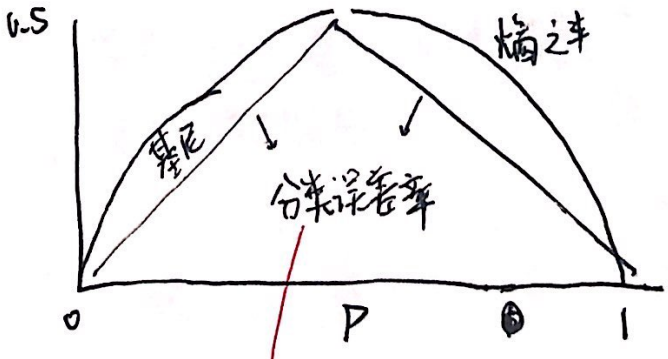
- 混乱程度的理解，如让我预测抛一枚硬币是or正反面的结果，

比如抛一枚均匀的硬币，正反面出现的可能性都为 $\frac{1}{2}$

但如果抛一枚不均匀的硬币，比如正面出现可能性为 $\frac{8}{9}$

反面出现可能性为 $\frac{1}{9}$

我就有很大地信心说它可能会是正面，没有那么混乱

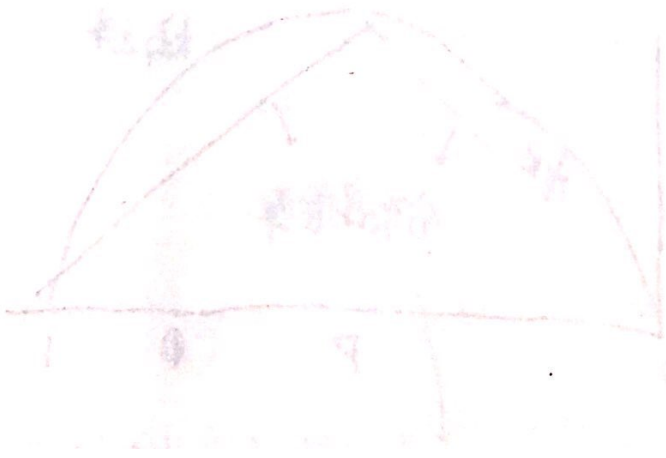


表示我有多大的可能性预测错误

第五章中比较特别的是

不仅输出的变量是分类的

而且，输入的变量也是分类的



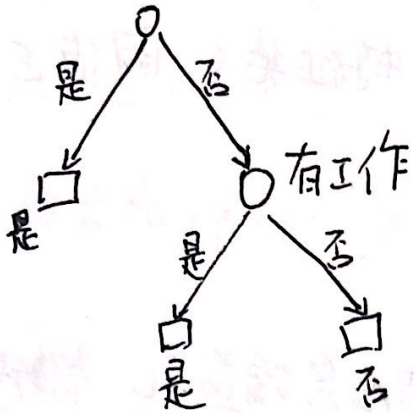
可视化好，但问题是

5.1 决策树模型与学习

① 如何选择特征作为根结点or每层节点
如何衡量分类好坏程度 信息增益 最大的

② 如何确定树状结构停止标准.
(若把每个实例都加入树状结构则会过拟合)

有自己的房子

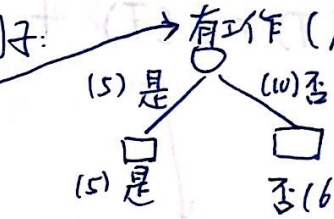


5.2 特征选择

熵: $H(X) = -\sum_{i=1}^{n} p_i \log p_i$ (共有n个类别; p_i: 每个类出现的概率; $\log p_i$ 即 $\log_2 p_i$)

- 用样本求出的熵叫经验熵
- 要对比每个特征可以把混乱度降低到什么样
- 比如选取有无工作这个特征

计算例子:



$-1 \cdot \log 1$

$-(\frac{6}{10} \cdot \log \frac{6}{10} + \frac{4}{10} \log \frac{4}{10})$

信息增益 $g(D,A) = H(D) - H(D|A)$ (条件熵)

$\frac{5}{15} \times (-1 \log 1) + \frac{10}{15} \times (-(\frac{6}{10} \cdot \log + \sim))$ (根据权重聚合熵)

信息增益比 $g_R(D,A) = \frac{g(D,A)}{H_A(D)}$

信息增益: 混乱程度的减少程度

但存在问题. 若A的类别很多. 叶子节点(每个)只有1个实例. 则信息增益会很大. 过拟合

信息增益比: 还考虑了原本样本的混乱程度

5.3 决策树的生成

算法5.2（ID3算法）—用信息增益来衡量 → 都提高泛化能力

输入：训练集D，特征集A，阈值ε → 预剪枝

算法5.3（C4.5的生成算法）-用信息增益比来衡量

算法5.4 决策树的剪枝 — 后剪枝

决策树损失函数

$$C_\alpha(T) = C(T) + \alpha|T|$$

$|T|$：叶节点的个数

α 表示对树叶节点的惩罚，因为叶节点越多，越复杂，泛化能力↓

$$C(T) = \sum_{t=1}^{|T|} N_t H_t(T)$$

N_t：权重

$H_t(T)$：熵

t 表示叶节点

$C(T)$：用熵来衡量训练集的拟合程度，熵↓，分类能力↑

5.4 CART算法（只有2个分枝，不管该特征有几类）

基尼指数

$$Gini(p)=\sum_{k=1}^{K} p_k(1-p_k)$$

（2：平方；K：类别）

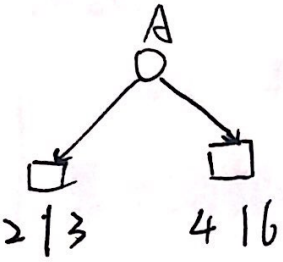
特征A的条件下，集合D的基尼指数：

goal: $\min Gini(D,A)$

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2)$$

A：特征；D_1：左树；D_2：右树

例



$$Gini(left)=\frac{2}{5}(1-\frac{2}{5})+\frac{3}{5}(1-\frac{3}{5})$$

$$Gini(right)=\frac{4}{10}(1-\frac{4}{10})+\frac{6}{10}(1-\frac{6}{10})$$

$$Gini(D,A)=\frac{2+3}{2+3+4+6}\times Gini(left)+Gini(right)\times\frac{4+6}{2+3+4+6}$$

★重要理解

在CART中不考虑特征A（本身的混乱程度）

因为CART只分2个类，而在~~信息增益~~前面里可能会分多个分枝，导致~~信息增益~~信息增益过大

算法5.5（最小二乘回归树生成算法）

算法5.6（CART生成算法）

算法5.7（CART剪枝算法）—交叉验证.