Decision Theory behind Classification

觀察不到 y

prediction : $\hat{y}$

" Loss of false prediction"

	truth $y=1$	truth $y=0$
$\hat{y}=1$	0	C_2
$\hat{y}=0$	C_1	0

C_1 : when $y=1$, $\hat{y}=0$

$C_1 \neq C_2$ (不對稱)

C_2 : when $y=0$, $\hat{y}=1$

給定 C_1、C_2

$\hat{y}=0$ or $\hat{y}=1$?　　犯錯不可避免，但能計算、壓底犯錯機率

e.g. 假設男人中有 P% 是渣男，P已知

Expected Loss (Risk) of $\hat{y}=1$

$$P(y=1)\cdot 0 + P(y=0)\cdot C_2 = (1-P)\,C_2$$

Expected Loss (Risk) of $\hat{y}=0$

$$P(y=1)\cdot C_1 + P(y=0)\cdot 0 = P\,C_1$$

⇒ Optimal Prediction

$$\begin{cases} \hat{y}^* = 1, & \text{when } (1-p)C_2 < pC_1 \Rightarrow p > \dfrac{C_2}{C_1+C_2} \\ \hat{y}^* = 0, & \text{otherwise} \end{cases}$$

主觀衡量

∴ 當 p 足夠高 ⇒ 預測 $\hat{y}=1$

Remark:

- when $C_1 \uparrow$

  $\Rightarrow$ threshold $\frac{C_2}{C_1+C_2} \downarrow$ $\Rightarrow$ P更容易 > threshold $\Rightarrow$ 越容易作出 $\hat{y}=1$

- when $C_2 \uparrow$

  $\Rightarrow$ threshold $\frac{C_2}{C_1+C_2} \uparrow$ $\Rightarrow$ 越容易作出 $\hat{y}=0$

0-1 loss (loss : $C_1 = C_2 = 1$)

	$y=1$	$y=0$
$\hat{y}=1$	0	1
$\hat{y}=0$	1	0

$$\therefore \quad \hat{y}^* = \begin{cases} 1 & , \ p > \frac{1}{1+1} = \frac{1}{2} \\ 0 & , \ \text{otherwise} \end{cases}$$

然而，實際上我們並不清楚 p

$\Rightarrow$ 只能估計 $\hat{p}$

Data

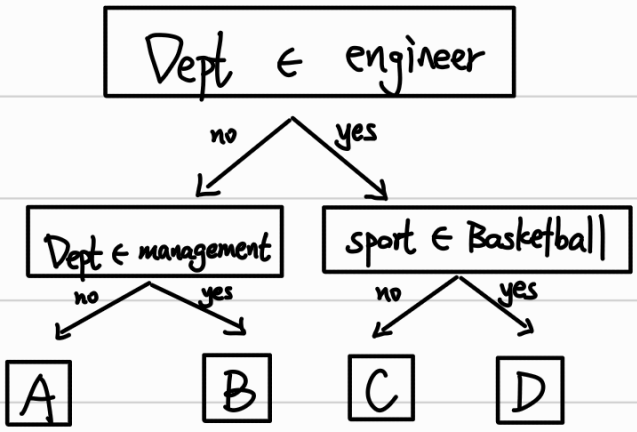
y , X (dept. , age, sport)

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

But how to use X ?

Decision Tree

先切開資料 X



⇒ Decision Tree

in D

$$\#\{y=1\} = 80$$
$$\#\{y=0\} = 20$$
$$\Rightarrow \quad \hat{p}(y|D) = 0.8$$

in C

$$\#\{y=1\} = 10$$
$$\#\{y=0\} = 90$$
$$\Rightarrow \quad \hat{p}(y|C) = 0.1$$

Decision rule

$$\underbrace{P(y \mid x \in ?)}_{\text{better}} > \frac{C_2}{C_1+C_2}$$

比 $P(y) > \frac{C_2}{C_1+C_2}$ 還好（因為用了 X 的資料）

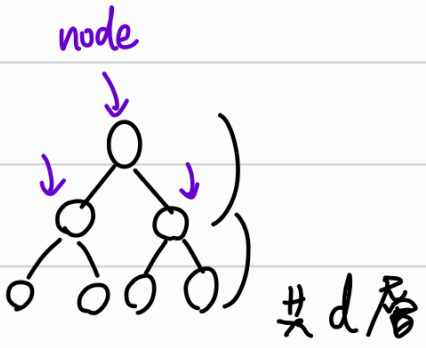
Challenge

有很多種可能的樹 → 哪種切法比較好？

Question

假設： depth of tree = d

有 p 個 variables



⇒ 共有多少種 tree？

$$2^0 + 2^1 + 2^2 + 2^3 + \cdots$$

How many nodes？

$$\frac{1(1-2^{d+1})}{-1}$$

$2^D - 1$

trees ： $p^{(2^D-1)}$ 種 組合

若 $p=10, D=3$ ⇒ 10^7 (一千萬種)

⇒ 組合問題複雜

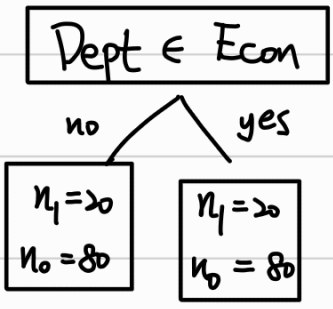
⇒ hard to find global optimum

Greedy Algorithm

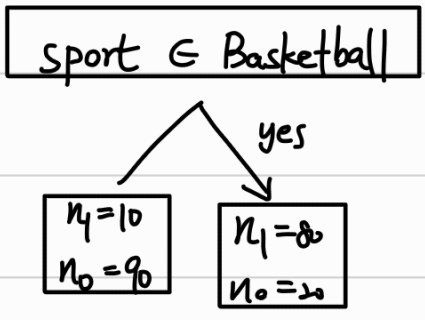
if you can only make one split

which one is the best ?

split 1



split 2



⇓

split 2 比 split 1 好

⇒ 告訴我們比較多資訊

(reduce entropy)

熵 (亂度)

entropy :

$X_1 \sim$ Bernoulli (0.5) ⇒ 抽出來可能是 0, 1, 0, 0, 1, ⋯

$X_2 \sim$ Bernoulli (0.001) ⇒ 抽出來幾乎都 0, 0, 0, ---- (entropy 低)

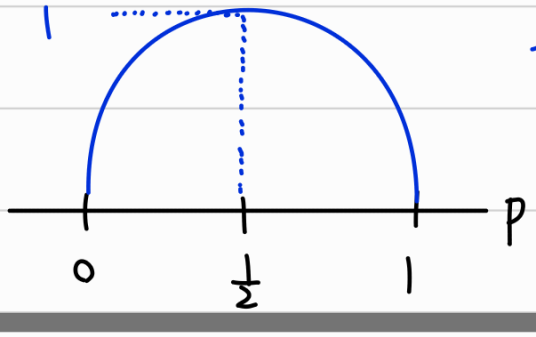
Def (Entropy for Ber(p))

$X \sim Ber(p)$

$\Rightarrow$ Entropy : $H(X) = -[p \log_2 p + (1-p) \log_2 (1-p)]$

since $p \in (0,1)$, $\log p \leq 0$

$\therefore H(X) \geq 0$



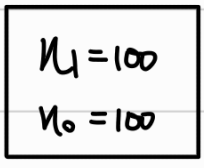
越靠近中間越亂

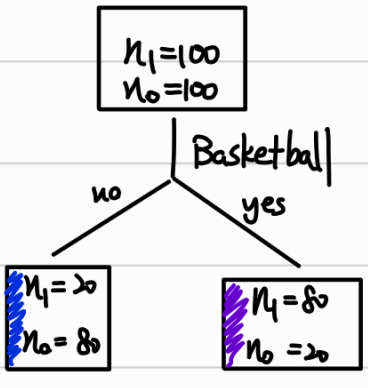
Greedy Algorithm for decision tree

遞迴地 repeatedly 尋找能最小化 entropy 的 split

(= 最大化 information gain)

information gain





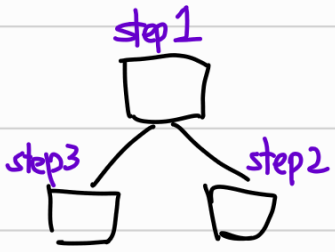
originally : $H = 1 \ (p = \frac{1}{2})$

■ $= 0.2 \log_2 0.2 + 0.8 \log_2 0.8$
$\cong 0.57$

■ $= 0.8 \log_2 0.8 + 0.2 \log_2 0.2$
$\cong 0.57$

average entropy : 0.57

information gain : $1 - 0.57 = 0.43$



Remark :

depth = d

variable = p

$(2^d - 1)p$ <<<< $p^{(2^d-1)}$

greedy

Remark :

Greedy algorithm 不保證能找到 global optimal solution

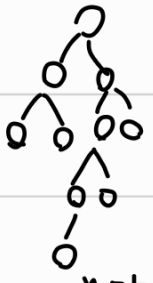
Remark :

when to stop ? (cross-validation)

bias - variance tradeoff

如果 depth 太深 ⟷ depth 太淺

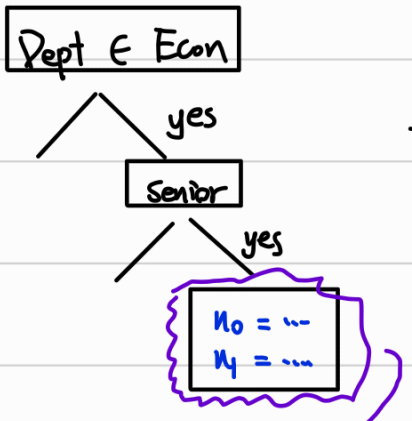
underutilize the information of X



$n_1 = 5$, $n_0 = 2$ $\Rightarrow \hat{p} = \frac{5}{7}$ (樣本數太少)

noisy

How to use a tree ?



form $\hat{p}(y=1 \mid x \in \square)$

if $\hat{p}(y=1 \mid x \in D) > C$

then $\hat{y}=1$, where $C = \frac{C_2}{C_1+C_2}$

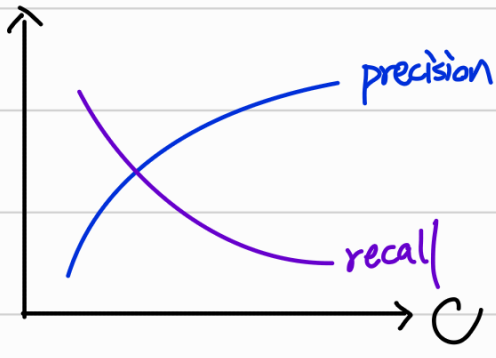
Trade-off in C

C太大 ⇒ 太寬鬆

C太小 ⇒ 太嚴格

	$y=1$	$y=0$
$\hat{y}=1$	TP	FP
$\hat{y}=0$	FN	TN

precision : $\frac{TP}{TP+FP}$ $\hat{y}=1$ 中 $y=1$ 的次數 / $\hat{y}=1$ 的次數



recall : $\frac{TP}{TP+FN}$ $y=1$ 之中，被你認為 $\hat{y}=1$ 次數 / $y=1$ 的人數