Decision theory behind.

"Loss of false prediction" truth.

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

$C_1, C_2 > 0$ $C_1 \neq C_2$.

C_1: when $y=1$. $\hat{y}=0$ cost of dating a fuckboy (渣男)

C_2: when $y=0$. $\hat{y}=1$. the loss of not dating a decent guy.

Given C_1. C_2 $\hat{y}=1$ or $\hat{y}=0$?

Suppose $p\%$ of men are fuckboy.

Expected loss of $\hat{y}=1$ "Risk" $\nearrow (1-p)$ $p\%$ are fuckboy

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

$= (1-p)C_2$.

} Guessing he is a fuckboy.

$$\hat{y}^* = \begin{cases} 1 & , p > \frac{C_2}{C_1+C_2} \\ 0 & \text{o.w.} \end{cases}$$

when p high enough. we predict $\hat{y}=1$

Risk of $\hat{y}=0$

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

$= P \cdot C_1$.

} Guessing he is not a fuckboy. Rmk when $C_1 \uparrow$ threshold $\frac{C_2}{C_1+C_2} \downarrow$

$\Rightarrow$ tend to predict $\hat{y}=1$.

Rmk when $C_2 \uparrow$

$\frac{C_2}{C_1+C_2} \uparrow \Rightarrow$ optimal rule. tends to predict $\hat{y}=0$.

Optimal Prediction.

$$\hat{y} = \begin{cases} 1 & , (1-p)C_2 < pC_1 \\ 0 & \text{o.w.} \end{cases}$$

$(1-p)C_2 < pC_1$

$\Leftrightarrow C_2 < (C_1+C_2)p$

$\Leftrightarrow p > \frac{C_2}{C_1+C_2}$

0-1 loss

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

in this situation

$$\hat{y}^* = \begin{cases} 1, & p > \frac{1}{2} \\ 0, & \text{o.w.} \end{cases} \quad \text{when } p > \frac{1}{2}$$

we predict $\hat{y}=1$

In practice, we don't know p.

We have to estimate it.

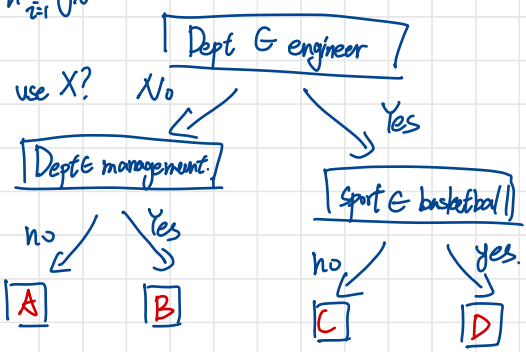
Data.

y	dept	age	sport
1	econ	21	⋮
0	EE	20	⋮
1	⋮	19	⋮
1	⋮	21	⋮
0	⋮	⋮	⋮

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

How to use X?

Decision tree:
partition out data. by X



In D

$\#\{y=1\}=80$

$\#\{y=0\}=20$

$\hat{p}(y|D)=0.8$

In C

$\#\{y=1\}=10$

$\#\{y=0\}=90$

$\hat{p}(y|C)=0.1$

decision rule

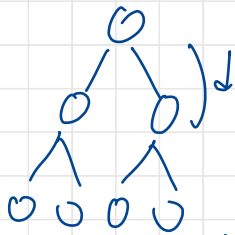
$$p(y|x\in\square) > \frac{C_2}{C_1+C_2}$$

There are many possible trees. Which one is better?

Question: Suppose depth of tree = d

Suppose there are p variables.

How many possible trees are there?



How many nodes? $2^d - 1$

Trees: $(p)^{2^d}$ combinations.

eg. $p=10$ $D=3$

$(p)^{2^D-1} = 10^7$

Again.

Combinational problem.

hard to find global optimum.

Greedy algorithm

if you can only make one step

to it. which one's the best?

Split 1

Dept.E Econ

No → $n_1 = 20$, $n_0 = 80$

Yes → $n_1 = 20$, $n_0 = 80$

Split 2

Sport & Basketball

No → $n_1 = 10$, $n_0 = 90$

Yes → $n_1 = 80$, $n_0 = 20$

Split 2 is a better split because it reduces entropy 熵(亂度)

$X_1 \sim Ber(0.5)$ ← 比較亂

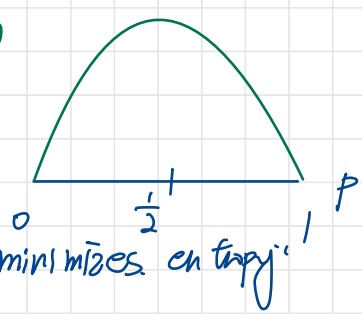
$X_2 \sim Ber(0.001)$

Def (Entropy for Ber(p))

$X \sim Ber(p)$

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

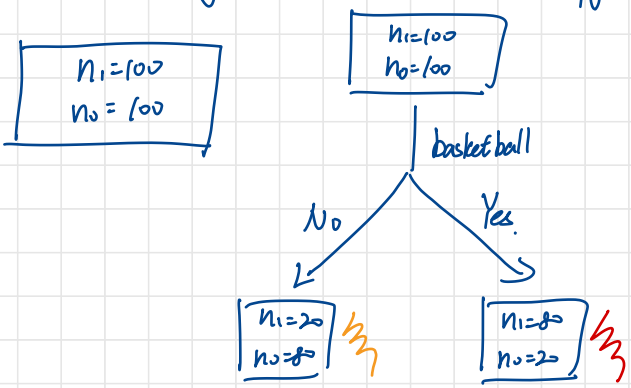
Since $p \in (0,1)$, $\log p \le 0$, so $H(X) \ge 0$



"recursively find the split that minimizes entropy"

Information gain = reduction in entropy

Original, $H = 1(P=\frac{1}{2})$



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

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

avg entropy = 0.57.

information gain = $1 - 0.57 = 0.43$.

Greedy algorithm. for decision tree.

Repeat: find the split with highest information gain.

Rmk depth = d

variable = p

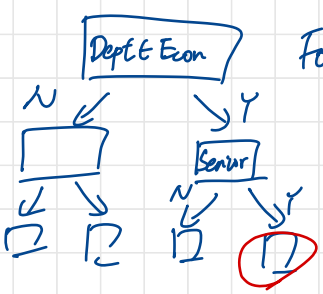
$(2^d - 1)p << p^{(2^d-1)}$

Rmk: Greedy algorithms does not guarantee to find global optimal solutions

Rmk When to stop?

"bias-variance trade off" cross-validation.

How to use a tree?



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

Plug in decision rule.

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

$\hat{y} = 1$. where $C = \frac{c_2}{c_1+c_2}$

Trade-off in C

$C\uparrow \Rightarrow$ too lenient

$C\downarrow \Rightarrow$ too strict.

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

$$\text{precision} = \frac{TP}{TP+FP} \Rightarrow \#\ \hat{y}=1$$

$$\text{recall} = \frac{TP}{TP+FN} \Rightarrow \#\ y=1$$

if $C\uparrow$, precision $\uparrow$ recall $\downarrow$

if $C\downarrow$, precision $\downarrow$. recall $\uparrow$