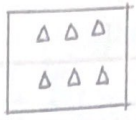
分錯的誤差（算=凌亂程度）

① Gini 0 ~ 0.5 （越接近 0.5 越=混亂）

② Entropy : 0 ~ 1 （ " 1 " ）

③ Misclassification : 0 ~ 0.5 （ " 0.5 " ）



↓ error = $1 - \max\left[\frac{0}{6}, \frac{6}{6}\right] = 0$

others 比例 / △ 比例 亂度 0.

ex1 解釋各變數：客戶編号不管，忠誠度 y（應變數），others 自變數

ex2：是否还款 => 是(P)⑤ / 否(N)④ => I(5,4)

* 計算 Entropy，2分變項 => 1分加号
* 以 I(D) 表分割前亂度 / $I_A(D)$ 表分割後亂度
* If I(0,0) = 0
* 前一階段的分割條件，下一步驟可不用，但下下步未必
* 如果算出的增量都一樣，可以 choose any 來当分枝樹
* 前9項用來建模（建決策樹），後3項用來測試

Entropy :

$$I(D) = I(5,4) = -\left[\frac{5}{9}\cdot\log_2\frac{5}{9} + \frac{4}{9}\cdot\log_2\frac{4}{9}\right]$$

$$= -\left[\frac{5}{9}\times(-0.847) + \frac{4}{9}\times(-1.169)\right] = 0.989$$

Gini :

$$I(D) = I(5,4) = 1 - \left(\frac{5}{9}\right)^2 - \left(\frac{4}{9}\right)^2 = 0.494$$

①

引入"是否負債"做分割條件

是否負債	p (會還款)	n (不會還款)	total
是	1	3	4
否	4	1	5
total	5	4	9

Entropy :

$$I_A(D): I_A(1,3) = \frac{4}{9}I(1,3) + \frac{5}{9}I(4,1)$$

0.811

$$= \frac{4}{9}\cdot\underbrace{-\left[\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right]}_{0.722} +$$

$$\frac{5}{9}\cdot\underbrace{-\left[\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right]}$$

$$= 0.36 + 0.4$$

$$= 0.76 \#$$

增益量 $Gain(A) = I(D) - I_A(D) = 0.989 - 0.76 = 0.229$

Gini:

$I_A(D) = \frac{4}{9} \cdot I(1,3) + \frac{5}{9} \cdot I(4,1)$

$= \frac{4}{9} \cdot \underbrace{[1-(\frac{1}{4})^2-(\frac{3}{4})^2]}_{6/16} + \frac{5}{9} \underbrace{[1-(\frac{4}{5})^2-(\frac{1}{5})^2]}_{8/25}$

$= \frac{1}{6} + \frac{8}{45} = 0.345$ #

$Gain(A) = I(D) - I_A(D) = 0.494 - 0.345 = 0.149$

② 引入"性別"做為分割條件

Entropy:

$I_A(D) = \frac{5}{9} \times I(3,2) + \frac{4}{9} \times \boxed{I(2,2) = 1}$

$= \frac{5}{9} \times \underbrace{-[\frac{3}{5} \times \log_2 \frac{3}{5} + \frac{2}{5} \times \log_2 \frac{2}{5}]}_{} +$

$\frac{4}{9} \times \underbrace{-[\frac{2}{4} \times \log_2 \frac{1}{2} + \frac{2}{4} \times \log_2 \frac{1}{2}]}_{1}$

$= \frac{5}{9} \times (0.971) + \frac{4}{9} \times (1) = 0.983$ #

性別	p	n	total
男	3	2	5
女	2	2	4
total	5	4	9

$Gain(A) = 0.989 - 0.983 = 0.006$

Gini:

$I_A(D) = \frac{5}{9} \times I(3,2) + \frac{4}{9} \times \boxed{I(2,2) = \frac{1}{2}}$

$= \frac{5}{9} \times \underbrace{[1-(\frac{3}{5})^2-(\frac{2}{5})^2]}_{12/25} + \frac{4}{9} \times \underbrace{[1-(\frac{1}{2})^2-(\frac{1}{2})^2]}_{1/2}$

$= \frac{4}{15} + \frac{2}{9} = 0.487$ #

$Gain(A) = 0.494 - 0.487 = 0.007$

③ 引入"婚姻狀況"做為分割條件

Entropy:

$I_A(D) = \frac{6}{9} \times I(3,3) + \frac{3}{9} \times I(2,1)$

$= \frac{6}{9} \times (1) + \frac{3}{9} \times \underbrace{-[\frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3}]}_{0.917}$

$= 0.972$ #

婚姻狀況	p	n	total
單身	3	3	6
結婚	2	1	3
total	5	4	9

$Gain(A) = 0.989 - 0.972 = 0.017$

Gini：

$\frac{6}{9} \times \boxed{I(3,3)} + \frac{3}{9} \times I(2,1) = \frac{6}{9} \times (0.5) + \frac{3}{9} \times [1 - (\frac{2}{3})^2 - (\frac{1}{3})^2] = 0.481$

0.5 = $I(3,3)$；$[1-(\frac{2}{3})^2-(\frac{1}{3})^2] = 4/9$

Gain(A) = 0.494 - 0.481 = 0.013

④

引入"收入"做分割條件

收入	p（含還款）	n（不還款）	total
高	2	2	4
低	3	2	5
total	5	4	9

Entropy：

$I_A(D) = \frac{4}{9}\boxed{I(2,2)} + \frac{5}{9} I(3,2)$

1 = $I(2,2)$

$= \frac{4}{9} \times -[\frac{1}{2}\times\log_2\frac{1}{2} + \frac{1}{2}\times\log_2\frac{1}{2}] +$

$\frac{5}{9} \times -[\frac{3}{5}\times\log_2\frac{3}{5} + \frac{2}{5}\times\log_2\frac{2}{5}]$

$= 0.444 + 0.539$

$= 0.983$

Gain(A) = 0.989 - 0.983 = 0.006

Gini：

$I_A(D) = \frac{4}{9} I(2,2) + \frac{5}{9} I(3,2)$

$= 0.222 + \frac{5}{9} \times [1 - (\frac{3}{5})^2 - (\frac{2}{5})^2]$ 0.267

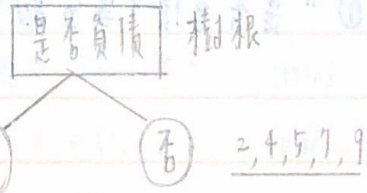
$= 0.489$

Gain(A) = 0.494 - 0.489 = 0.005

結論：選擇增益量最大者

	Entropy	Gini
是否負債	0.229	0.149
性別	0.006	0.007
婚姻狀況	0.017	0.013
收入	0.006	0.005

⇒ 以"是否負債"為分割條件的增益量最大，故選擇"是否負債"為決策樹之分支條件



樹根

* 下階段可不考慮重覆的分割條件

左分枝開始重新計算評估

只觀察 1, 3, 6, 8 (否)(否)(否) ⇒ I(1,3)

① "性別"當分割條件

起初 Entropy : I(1,3) = 0.811

Gini : I(1,3) = 0.375

性別	p	n	total
男	1	2	3
女	0	1	1
total	1	3	4

now Entropy : $\frac{3}{4} I(1,2) + \frac{1}{4} I(0,1) = 0.688$

Gini : $\frac{3}{4} I(1,2) + \frac{1}{4} I(0,1) = 0.333$

⇒ 值越大，增益量越小

② "婚姻"當分割

Entropy : $\frac{3}{4} \times I(1,2) + \frac{1}{4} \times I(0,1) = 0.688$

Gini : $\frac{3}{4} I(1,2) + \frac{1}{4} I(0,1) = 0.333$

婚姻	p	n	total
單	1	2	3
結	0	1	1

③ "收入"當分割

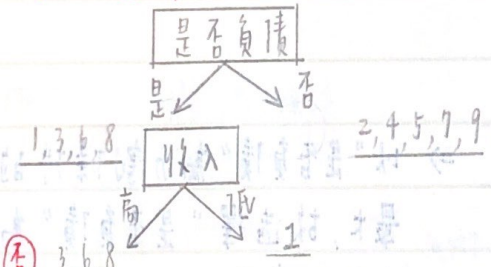
Entropy : $\frac{1}{4} I(0,1) + \frac{3}{4} I(1,2) = 0.688$

Gini : 0.333

收入	p	n	total
低	0	1	1
高	1	2	3

* 上述，不論選哪個當分枝，增益量皆相同

故"可自由選擇"



只觀察 3, 6, 8 (是)(否) ⇒ I(1,2)

① "是否負債"當分割

Entropy : $\frac{3}{3} I(1,2) + 0\, I(0,0) = 0.917$

Gini : $\frac{3}{3} I(1,2) + 0\, I(0,0) = 0.444$

負債	p	n	total
是	1	2	3
否	0	0	0

② "性別"

Entropy : $\frac{2}{3} I(1,1) + \frac{1}{3} I(0,1) = 0.667$

Gini : $\frac{2}{3} I(1,1) + \frac{1}{3} I(0,1) = 0.334$

性別	p	n	total
男	1	1	2
女	0	1	1

③ "婚姻"

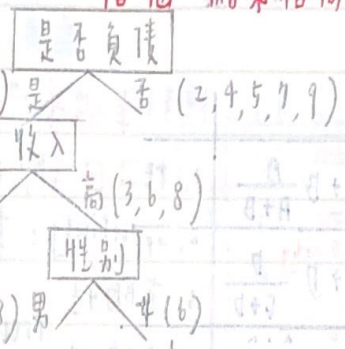
結果同上

婚姻	p	n	total
Single	1	1	2
married	0	1	1

1) 當結果值越大，增益量越小

所以不必算出原本之 Entropy & Gini

* 上述，故選"性別" or "婚姻"結果相同



是 否

只觀察 3,8 ⇒ I(1,1)

①"是否負債"

Entropy = $\frac{2}{2} I(1,1) + \frac{0}{2} I(0,0) = 1$

Gini = $\frac{2}{2} \times [1 - (\frac{1}{2})^2 - (\frac{1}{2})^2] = 0.5$

亂度 Max

負債	p	n	total
是	1	1	2
否	0	0	0

②"婚姻"

Ⓔ : $\frac{1}{2} I(1,0) + \frac{1}{2} I(0,1) = 1$

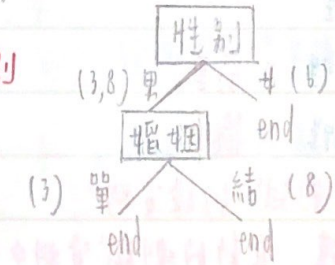
Ⓖ : $\frac{1}{2} \times [1 - 1 - 0] = 0$

婚姻	p	n	total
single	1	0	1
married	0	1	1

③ "收入"

結果同"是否負債"

* 以"婚姻"做分割



收入	p	n	total
高	1	1	2
低	0	0	0