WEEK 12 TA

DECISION TREE

- ID3（僅能分類、僅適用離散型變數、無法剪枝）
- C4.5（僅能分類）
- CART（可做分類與迴歸，且改善兩者的缺點）

GINI IMPURITY

- 用於衡量資料的亂度
- 數字越大代表資料越混亂（不純）；數字越小代表資料越不混亂（純）
- 公式：

$$\text{Gini} = 1 - \sum p_j^2$$

P_j表示欲分類的第j個類別的佔比

EXAMPLE

是否有房	婚姻狀況	年收入	是否拖欠貸款
是	單身	125k	否
否	已婚	100k	否
否	單身	70k	否
是	已婚	120k	否
否	離婚	95k	是
否	已婚	60k	否
是	離婚	220k	否
否	單身	85k	是
否	已婚	75k	否
否	單身	90k	是

是否有房 VS 是否拖欠貸款

	有房	無房
否	3	4
是	0	3

$$\text{Gini}(t_1)=1-(3/3)^2-(0/3)^2=0$$

$$\text{Gini}(t_2)=1-(4/7)^2-(3/7)^2=0.4849$$

$$\text{Gini}=0.3 \times 0 + 0.7 \times 0.4898 = 0.343$$

婚姻狀況 VS 是否拖欠貸款－I

	單身或已婚	離婚
否	6	1
是	2	1

$\text{Gini}(t_1)=1-(6/8)^2-(2/8)^2=0.375$

$\text{Gini}(t_2)=1-(1/2)^2-(1/2)^2=0.5$

$\text{Gini}=(8/10)\times 0.375+(2/10)\times 0.5=0.4$

婚姻狀況 VS 是否拖欠貸款—2

	單身或離婚	已婚
否	3	4
是	3	0

$\text{Gini}(t_1)=1-(3/6)^2-(3/6)^2=0.5$

$\text{Gini}(t_2)=1-(4/4)^2-(0/4)^2=0$

$\text{Gini}=(6/10) \times 0.5 + (4/10) \times 0 = 0.3$

婚姻狀況 VS 是否拖欠貸款—3

	離婚或已婚	單身
否	5	2
是	1	2

$\text{Gini}(t_1)=1-(5/6)^2-(1/6)^2=0.2778$

$\text{Gini}(t_2)=1-(2/4)^2-(2/4)^2=0.5$

$\text{Gini}=(6/10)\times 0.2778+(4/10)\times 0.5=0.3667$

年收入 VS 是否拖欠貸款：連續型採用切割點進行分類

	60		70		75		85		90		95		100		120		125		220
	65		72		80		87		92		97		110		122		172		
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	
是	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	
否	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	
Gini	0.400		0.375		0.343		0.417		0.400		0.300		0.343		0.375		0.400		

婚姻狀況 VS 年收入

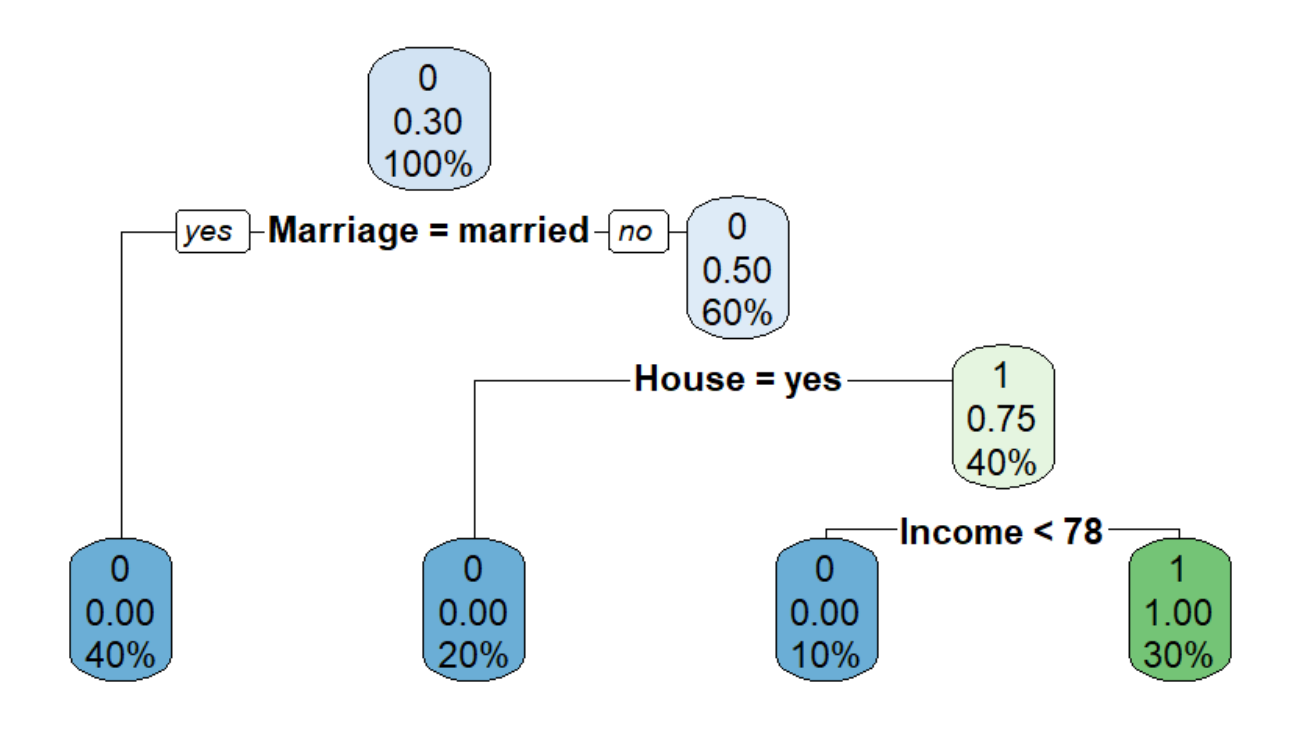
	單身或離婚	已婚
否	3	4
是	3	0

Gini=0.3

	年收<=97	年收>97
否	3	4
是	3	0

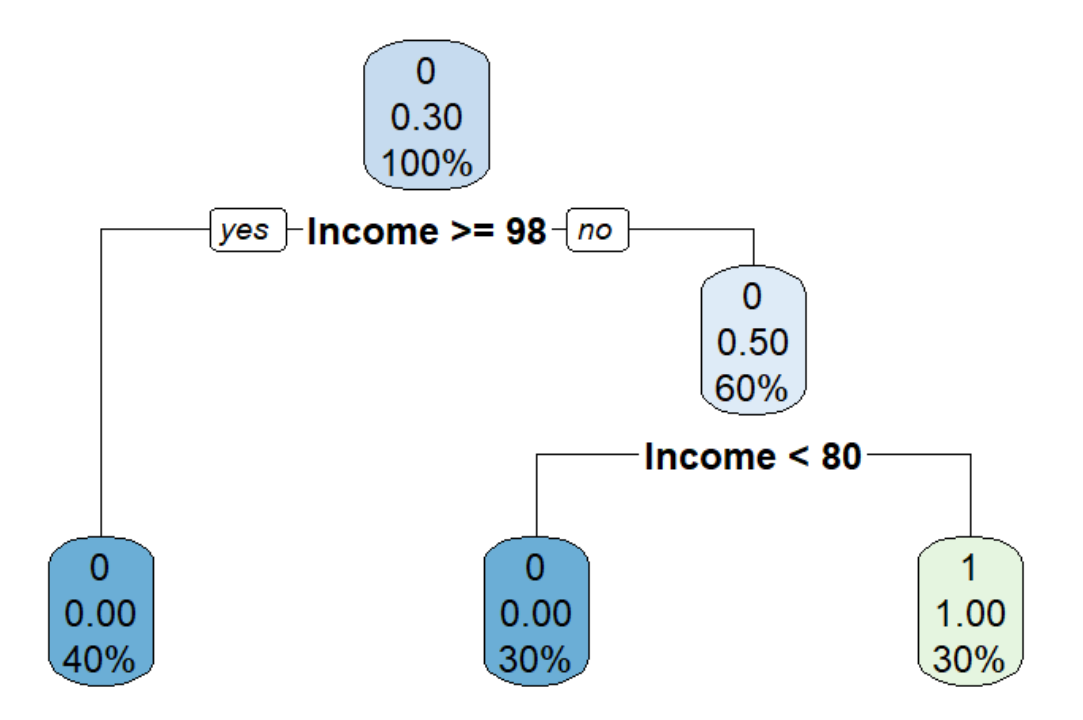
Gini=0.3

MODEL 1 : IN_DEBT ~ HOUSE + MARRIAGE + INCOME



```
In_Debt                                                                  cover
   0.00 when Marriage is             married                             40%
   0.00 when Marriage is divorced or single & House is yes               20%
   0.00 when Marriage is divorced or single & House is  no & Income <  78 10%
   1.00 when Marriage is divorced or single & House is  no & Income >= 78 30%
```

MODEL 2 : IN_DEBT ~ INCOME + HOUSE + MARRIAGE



```
In_Debt                                          cover
   0.00 when Income >=          98                40%
   0.00 when Income <  80                         30%
   1.00 when Income is 80 to 98                   30%
```

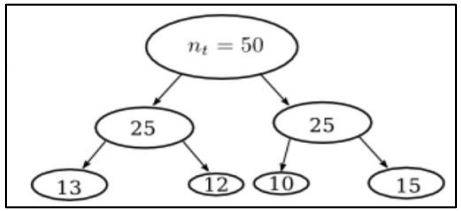
Random Forest

Introduction

- 以隨機抽取的訓練資料，產生多顆CART決策樹，以多數決預測分類
- 建立一棵樹的步驟：
  1. 隨機抽一定數量的樣本去訓練樹，可選擇取後放回或取後不放回
  2. 針對樹各個要分支的節點，每次隨機抽取K個變數當作候選變數
  3. 決定樹的深度，可用nodesize或maxnode決定
- 決定要種幾棵樹

Hyperparameter (In R)

Hyperparameter	Explanation	Default Setting For Classification
ntree	要種幾棵樹	500
mtry	分裂每個節點時要隨機選多少個變數當候選	√p 無條件捨去 p：解釋變數數量
replace	建每顆樹抽的樣本是否要取後放回	True
samplesize	建每棵樹時要抽多少樣本	if (replace) nrow(x) else ceiling(.632*nrow(x))
nodesize	建每棵樹時，末端節點最少要有多少樣本	1 for classification
maxnode	每棵樹內部最多可有幾個節點	Null



nodesize=10

參考資料

決策樹方法的比較：https://zhuanlan.zhihu.com/p/85731206

CART決策樹範例：https://zhuanlan.zhihu.com/p/139523931

交換變數的範例（Model 1& Model 2）：https://stackoverflow.com/questions/47715677/inconsistent-results-from-r-part-package