110年衛生福利資料科學中心

R軟體推廣課程《進階篇》

預測模型

講師：劉品崧　統計分析師

花蓮慈濟醫院

課程大綱

- 前言
- 迴歸模型預測(linear)
- 迴歸模型預測(logistic)
- 分類預測KNN
- 如何調整模型

前言

- 預測 & 模型
- 模型的信度與效度
- 建立預測模型之流程

預測

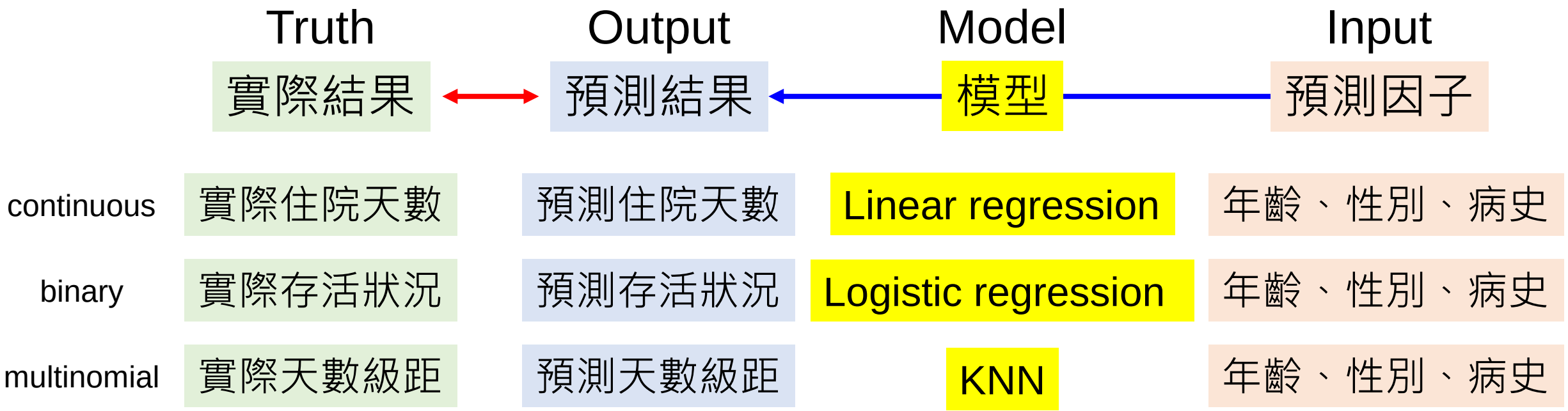
- 人類自古以來就一直嘗試想要預測
  - 巫術、觀天象、卜卦、塔羅牌、擲筊 etc
- 目的
  - 如果能夠得知結果，可以提早應對

    e.g. 重複入院患者的住院天數較長，提早開始安排控床
  - 如果能夠影響結果，可以提早介入

    e.g. 電商發現在FB投放的廣告比起IG有更多觸及率，加碼FB

模型

- 使用將預測因子與實際結果建立模型
- 模型的鏈結函數會因為實際結果的資料類型而有所不同
- 利用預測因子套用模型可以獲得預測結果
- 預測結果可以和實際結果比較，得知模型的準確度

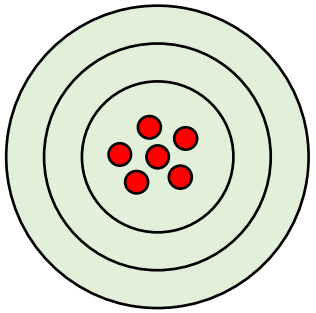


模型的信度與效度

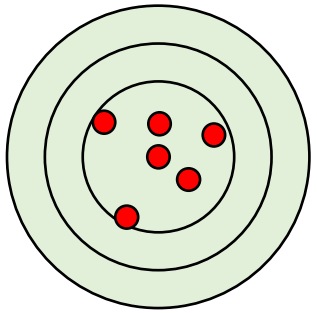
(地圖製造 MakeWorld.tw)

- 效度(validity)：執行動作與實際目標的準確性
- 信度(reliability)：準確性是否能穩定、一致地、重複地產生

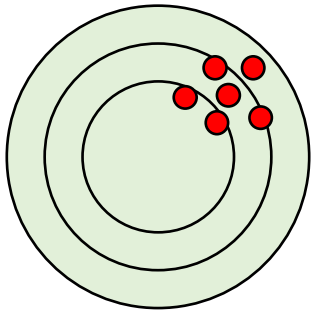
效度很好
信度很好



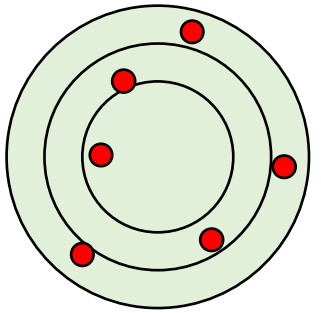
效度普通
信度不好



效度不好
信度很好



效度不好
信度不好



內部驗證與外部驗證(validation)

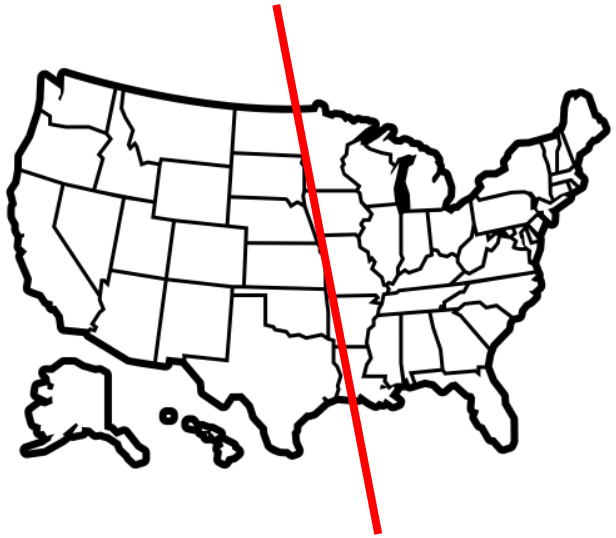
內部驗證
Internal validation

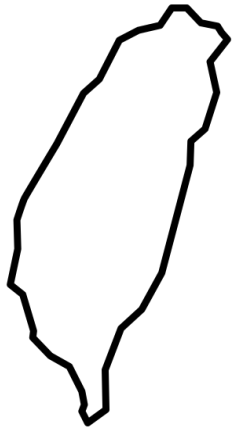
外部驗證
External validation

以美國西半邊人口
驗證中風後1年內死亡的模型
預測準確度為83%

以美國東半邊人口
建立中風後1年內死亡的模型
預測準確度為85%

以台灣人口
驗證中風後1年內死亡的模型
預測準確度為70%





Cross validation

- 自己的模型拿來預測自己的結果 → 自說自話，沒有說服力
  - 缺乏內部 & 外部的信效度
- 交叉驗證 Cross validation
  - 將資料分割為訓練資料集(train)以及驗證資料集(valid)
  - 常見的分割比例如7：3或8：2或2：1
  - 以train產生模型
  - 以valid套用模型而產生預測
  - 以valid預測的結果與實際值比較模型的準確度

Cross validation
內部驗證流程解析

(1)整理資料集

Full dataset

ID	住院天數	年齡
BB	16	59
OO	12	84
LL	17	63
FF	2	60
SS	6	1
QQ	24	38
DD	9	84
MM	9	36
NN	3	3
AA	8	72

(2)分割資料集

Train dataset(70%)

ID	住院天數	年齡
BB	16	59
OO	12	84
LL	17	63
FF	2	60
SS	6	1
QQ	24	38
DD	9	84

(3)訓練模型

(4)確認模型

預測住院天數 = 3.8668 + 0.0863 * 年齡

(5)放入預測因子

(6)獲得預測結果

Valid dataset(30%)

ID	住院天數	年齡
MM	9	36
NN	3	3
AA	8	72

預測住院天數
6.9736
4.1257
10.0804

(7)比較實際值與預測結果

Summary：建立預測模型之流程

1. 瞭解資料：預測結果的種類、預測因子的轉換
2. 分割資料：訓練(train)資料集、驗證(valid)資料集
3. 建立模型：模型公式(formula)、決定鏈結(link)函數
4. 訓練模型：以訓練資料集配適(fit)模型，取得參數(parameter)
5. 驗證模型：以驗證資料集套用模型，評估準確性(accuracy)
6. 調整模型：模型架構，因子選擇

迴歸模型預測(linear)

- 瞭解資料
- 分割資料
- 建立模型
- 訓練模型
- 驗證模型

線性模型架構

- $Y = \underbrace{\alpha + \beta_1 X_1 + \beta_2 X_2}_{=\hat{Y}} \ldots + \varepsilon$

  $= \hat{Y} + \varepsilon$

  - Y：實際值；$\hat{Y}$ ：預測值；ε：殘差，是實際值與預測值的誤差
  - α：截距，當所有input為0時$\hat{Y}$的基礎值
  - X_i：第i個因子的input；β_1 ：第i個因子的權重

- 實際範例
  - 預測住院天數 = 3.85124 + (0.08344 * 年齡)
  - 預測住院天數 = 4.25318 + (0.08288 * 年齡) + (-0.92894 * 性別為女性)

預測準確度評估指標

- RMSE，root mean square error
  - 殘差(residual) = 實際值 – 預測值
  - $\mathrm{RMSE} = \sqrt{mean(residual^2)}$

R語言指令

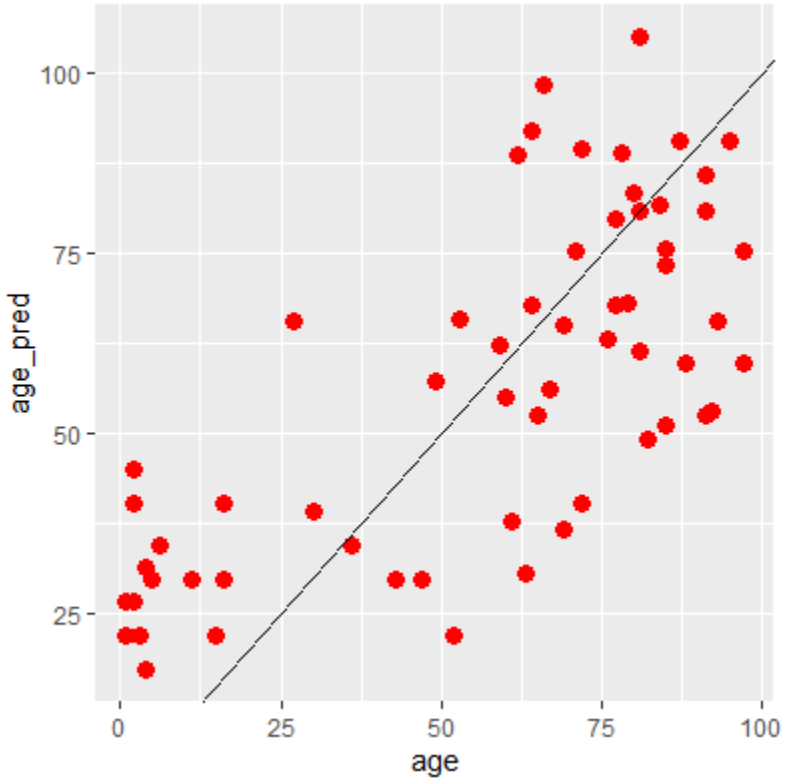
- 建構模型
  - `glm(formula = 模型公式, data = 訓練資料集)`
- 預測結果
  - `predict(object = 模型物件, newdata = 要放進去的資料)`

模型驗證與比較

- 結果來說，訓練資料集訓練產生的模型，其預測能力不算太好，R^2=53.0%。
- 但可以在驗證資料集被重複再現，具有一定的信度。

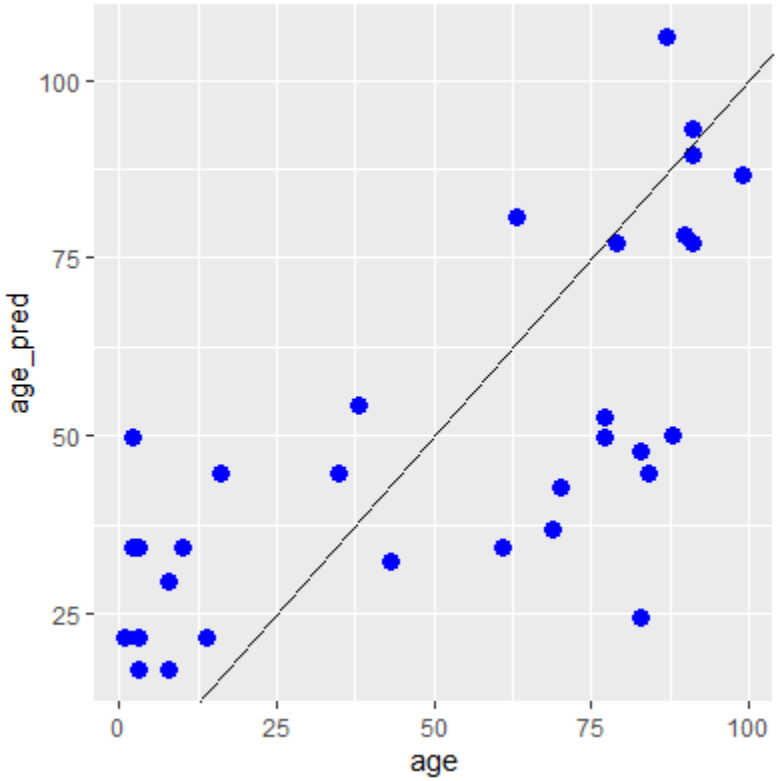
訓練資料集(n=63)

- RMSE：21.94
- *r*：0.7280



驗證資料集(n=32)

- RMSE：25.87
- *r*：0.7045



迴歸模型預測(logistic)

- 瞭解資料
- 分割資料
- 建立模型
- 訓練模型
- 驗證模型

Logistic regression model架構

- 當outcome是二元變數時，$ln(\frac{p}{1-p}) = \alpha + \beta_1 X_1 + \beta_2 X_2 \ ... + \varepsilon$
- $\hat{p} = \frac{1}{1+e^{(-(\alpha+\beta_1 X_1+\beta_2 X_2))}}$
- 實際範例

β ↓　　X(input) ↓

  - $ln(\frac{\text{出院後半年內死亡}}{\text{出院後半年內存活}}) = -1.4404 + 0.2273 * \text{年齡為65歲以上}$
  - 若年齡為65歲以上，則出院後半年內死亡的機率為0.229153
    - $\hat{p} = \frac{1}{1+e^{(-(-1.4404+0.2273*1))}} = 0.229153$
  - 若年齡為65歲以下，則出院後半年內死亡的機率為0.191483
    - $\hat{p} = \frac{1}{1+e^{(-(-1.4404+0.2273*0))}} = 0.191483$

預測準確度評估指標

- 比較實際結果與預測結果的交叉表
  - 要先從predict probability去找出最佳切點
  - 依據切點來預測事件是否會發生

	預測存活	預測死亡
實際存活	21	11
實際死亡	1	9

  - 準確率 $= \frac{\text{預測正確數量}}{\text{總樣本數量}} = \left(\frac{21+9}{42}\right) * 100\% = 71.4\%$

- AUROC，ROC曲線下面積
  - 越接近1越好，代表可以被完美預測

R語言指令

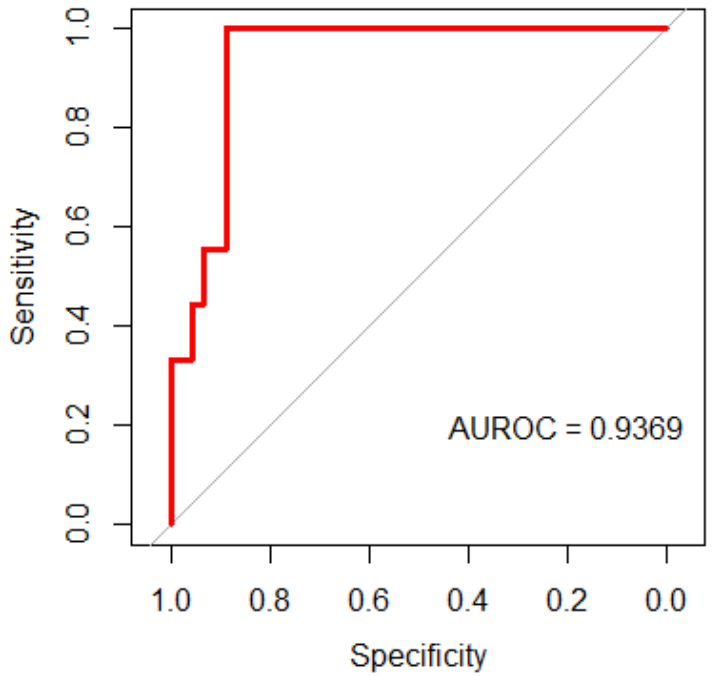
- 修改機率函數
  - `glm(formula = 模型公式, data = 訓練資料集, family = "binomial")`
- 製作ROC物件
  - `roc(response = 實際事件向量, predictor = 預測機率向量)`
- 找尋預測機率的最佳切點
  - `coords(roc = roc物件, x = "best", ret = "threshold")`

模型驗證與比較

- 結果來說，訓練資料集訓練產生的模型，其預測能力很好，c=93.69%。
- 再現的過程並沒有展現相同的預測力，需要再調整。

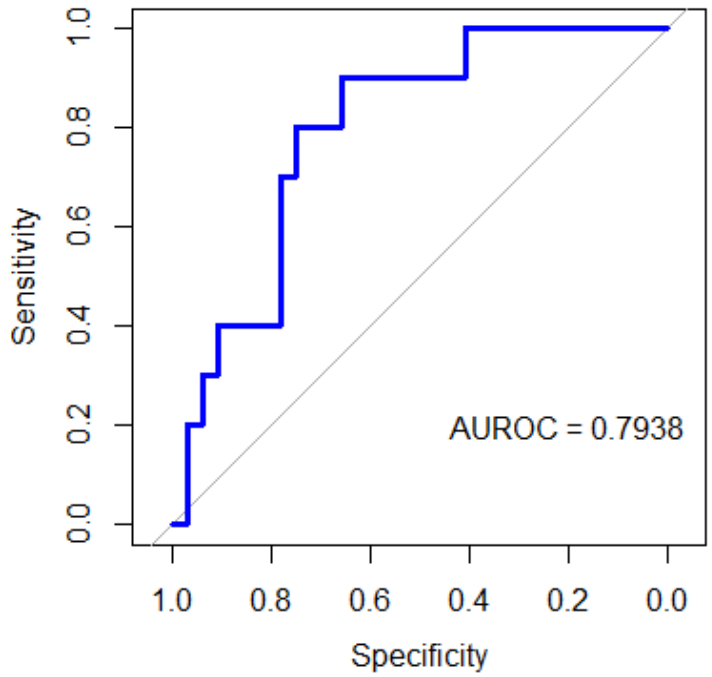
驗證資料集(n=53)

- *c*：0.9369



驗證資料集(n=42)

- *c*：0.7938



分類預測KNN

- 瞭解資料
- 分割資料
- 建立模型
- 訓練模型
- 驗證模型

KNN的概念與用途

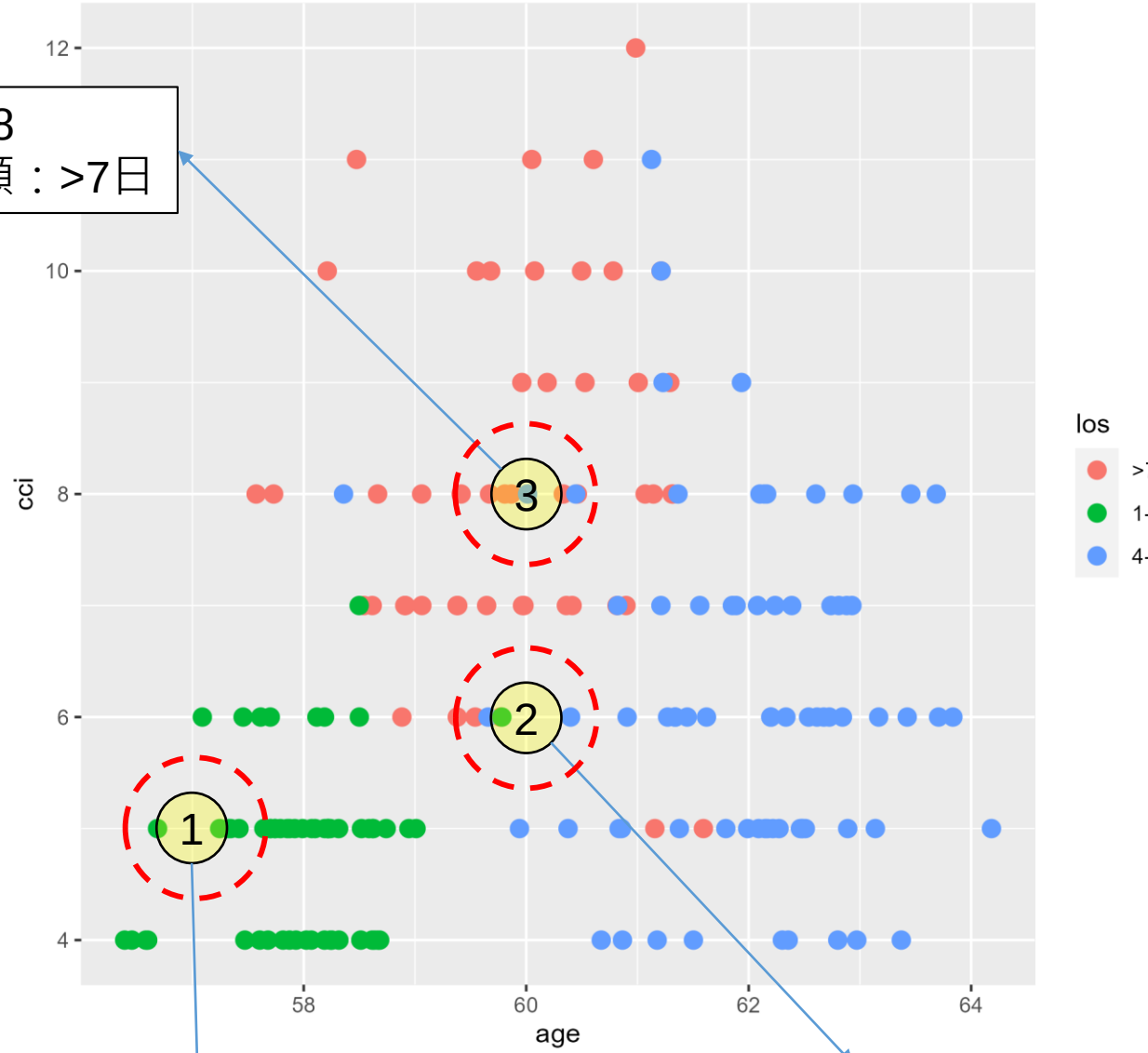
- 當想要預測的變項是多分類的時候
  - 中午吃什麼 → 雞腿便當、豬排丼、咖哩烏龍麵？
    - 多個分類之間不具備序位、可量測距離、連續等特性
- 此時不能再使用linear或是logistic regression
- 最簡易而常見的分類模型就是KNN模型
  - k-nearest neighbors
  - 利用歐式距離作為計算方式
  - 把最像你的K個其他人抓出來，投票決定你最有可能是誰！

歐式距離Euclidean distance

- 模型若有m個因子，將會對應出m維度的空間
- 二維空間：年齡、CCI
  - 總共i個valid樣本會與全部j個train樣本計算歐式距離
  - 距離$_{valid_i\ vs\ train_j} = \sqrt{(age_{valid_i} - age_{train_j})^2 + (CCI_{valid_i} - CCI_{train_j})^2}$
  - 進而找出最k個最鄰近(k-nearest neighbors)的train樣本進行投票
    - 三維空間(或以上)
    - $d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$

KNN的流程

- Train dataset
  - DV：住院天數分類
    - 分為3組，1-3天、4-6天、7天以上
  - IV：年齡、CCI
- Test dataset
  - 把你的sample標在這個地圖上
  - If K = 5則找尋離你最近的的5個人
  - 5個人裡面哪一個類型最多
  - 你就會被預測為那個類型



預測準確度評估指標

- 比較實際結果與預測結果的交叉表
  - 準確率
  - $= \frac{\text{預測正確數量}}{\text{總樣本數量}}$
  - $= \left(\frac{17+5+2}{41}\right) * 100\% = 58.5\%$

		預測結果		
		1-3天	4-6天	7天以上
實際結果	1-3天	17	2	0
	4-6天	8	5	1
	7天以上	5	1	2

R語言指令

- 建構指令可以容許的matrix物件
  - as.matrix(data[, c("var1", "var2", "var3")])
- 執行knn
  - knn(train = train的因子, test = test的因子, cl = train的答案, k = 3)

模型驗證與比較

- 結果來說，訓練資料集訓練產生的模型，其預測能力普通，準確度只有58.5%。

		預測結果		
		1-3天	4-6天	7天以上
實際結果	1-3天	17	2	0
	4-6天	8	5	1
	7天以上	5	1	2

如何調整模型

- 模型因子的變化
- 模型因子的選擇

模型因子的變化(1)連續變項的轉化

- 使用情境
  - 連續數值可能具有級距或類別的意義
  - 將年齡切為65歲上下
- 範例
  - 死亡 $= \alpha + \beta_{age} \cdot age + e$
  - 死亡 $= \alpha + \beta_{age(\geq 65歲)} \cdot age_{\geq 65歲} + e$

年齡以連續數值形式放入

	預測存活	預測死亡
實際存活	21	11
實際死亡	1	9

年齡以切點類別形式放入

	預測存活	預測死亡
實際存活	22	10
實際死亡	1	9

模型因子的變化(2)變化因子的次方

- 使用情境
  - 連續型結果可能不只是線性發展
  - 增加因子的不同次方項目有機會擬合曲線型的部分
- 範例
  - 住院天數 $= \alpha + \beta_{age} \cdot age + e$
  - 住院天數 $= \alpha + \beta_{age(1次方)} \cdot age^{1} + \beta_{age(2次方)} \cdot age^{2} + \beta_{age(3次方)} \cdot age^{3} + \beta_{age(1/2次方)} \cdot age^{1/2} + e$

模型因子的變化(3)因子之間的交互作用項

- 使用情境
  - 因子之間
    可能具有交互作用
  - 造成1+1 > 2的狀況
- 範例

模型沒有交互作用項

	預測 不再入院	預測 再入院
實際 不再入院	24	8
實際 再入院	9	1

模型加入交互作用項

	預測 不再入院	預測 再入院
實際 不再入院	22	10
實際 再入院	3	7

- 再次入院 $= \alpha + \beta_{DM} \cdot DM + \beta_{hyperlipidemia} \cdot hyperlipidemia + e$
- 再次入院 $= \alpha + \beta_{DM} \cdot DM$

$+\beta_{hyperlipidemia} \cdot hyperlipidemia$

$+\beta_{DM\&hyperlipidemia} \cdot DM\&hyperlipidemia + e$

模型因子的選擇(1)不同判斷標準

- 模型擬合
  - 逐一將可能的因子放入(forward)/剔除(backward)/雙向(both)調整模型
  - 以AIC為判斷標準，越小越好，直到AIC不再降低為止
    - Akaike information criterion，$AIC = 2k + n\ln(RSS/n)$
    - n為樣本數，k為因子數量，RSS為殘差平方合
- 統計顯著性選擇
  - 逐一將可能的因子分別對想預測的結果作迴歸分析
  - 以P value為判斷標準，單變項分析顯著的因子，才納入模型
    - 顯著性可以自由設定，0.1, 0.05, 0.01, etc.

模型因子的選擇(2)模型擬合的報表判斷

(a) 起始的AIC

- forward

```
Start:  AIC=43.47
death ~ age + sex_2

                 Df Deviance    AIC
+ pneumonia_pre   1   33.522 41.522
<none>                37.470 43.470
+ dm              1   36.354 44.354
+ incgp_2         1   36.618 44.618
+ incgp_1         1   36.815 44.815
+ cci             1   37.228 45.228
+ htn             1   37.382 45.382
+ hyperlipidemia  1   37.428 45.428
+ med_center      1   37.461 45.461
```

- backward

```
Start:  AIC=47.5
death ~ age + sex_2 + incgp_1 + incgp_2 + med_center + cci +
    pneumonia_pre + htn + dm + hyperlipidemia

                 Df Deviance    AIC
- hyperlipidemia  1   25.578 45.578
- incgp_1         1   25.696 45.696
- med_center      1   25.832 45.832
- cci             1   26.297 46.297
- incgp_2         1   26.407 46.407
- htn             1   27.304 47.304
<none>                25.499 47.499
- dm              1   27.998 47.998
- pneumonia_pre   1   33.936 53.936
```

(b) 調整後會讓AIC降至最低的一個因子
報表已經依照AIC進行排序，所以看最上面的一個就可以了
在forward裡面，變項以+表示應該放入這個因子
在backward裡面，變項以-表示應該剔除這個因子

模型因子的選擇(2)結果比較

模型因子 Factors	基本模型 Base model	全部放入 Full model	向前選擇 Forward	向後剔除 Backward	統計顯著 P-value
年齡	v	v	v	v	v
性別(女性 vs 男性)	v	v	v	v	v
保費級距(1)		v			
保費級距(2)		v			
醫學中心治療		v			
CCI		v	v		
肺炎病史		v	v	v	v
糖尿病		v		v	
高血壓		v		v	
高血脂		v			
預測準確度(%)	71.43	71.43	64.29	85.71	80.95

課程討論 & Final remark

- Summary
  - 瞭解資料
  - 分割資料
  - 建立模型
  - 訓練模型
  - 驗證模型
  - 調整模型

- See more
  - Stack Overflow
  - STHDA

- 劉品崧
  - **Email**：psliu520@gmail.com
  - **PubMed**：Peter Pin-Sung Liu

使用圖片版權來源

Created by P Thanga Vignesh
from Noun Project

Created by Setyo Ari Wibowo
from Noun Project