東吳大學 資料探勘導論 期中考題目

一、假設 I 是所有購物籃 (或車) 中曾經賣過的所有商品項目 (Items) 的集合，我們以 $I = \{I_1, I_2, I_3, \ldots, I_m\}$ 來表示，而資料庫 D 為所有交易 (Transaction) 資料 T 的集合，$D = \{T_1, T_2, T_3, \ldots, T_n\}$，如下表 14-1 所示：

✦ 表 14-1　交易資料庫 D

交易代號 (TID)	商品項目 (Items)
T_1	飯糰、豆漿、尿布
T_2	飯糰、尿布、啤酒、麥片
T_3	豆漿、尿布、啤酒、綠茶
T_4	飯糰、豆漿、尿布、啤酒
T_5	飯糰、豆漿、尿布、綠茶

(定義 1)每一個項目集 X 的支持數量可以寫成 $\sigma(X)$，

$\sigma(X) = |\{T_i \mid X \subseteq T_i, T_i \in D\}|$，其中 $|\cdot|$ 表示集合中元素的數量

(定義 2)每一個項目集 X 的支持度可以寫成 s(X)，其定義為

$$s(X) = \frac{\sigma(X)}{|D|}$$

，其中 $|D|$ 為交易資料庫中的總交易資料筆數

(定義 3)令 X 與 Y 為 I 的兩個子項目集，假如 X 與 Y 產生關聯規

則，則關聯規則可以表示為

$X \Rightarrow Y[s, c]$ 的型式 **$X, Y \subseteq I$**，其中 **$X, Y \neq \phi$** 且 **$X \cap Y = \phi$**

根據上述三個定義描述內容，請使用”購物車或者購物籃的白話觀念”回答下列問題

(A)根據上述定義 **1** 中的描述請舉例說明 **$\sigma(X)$**的白話意義為何**?**

(B)根據上述定義 **3** 中的描述請舉例說明 **$X, Y \neq \phi$** 的白話意義為何**?**

(C)根據上述定義 **3** 中的描述請舉例說明 **$X \cap Y = \phi$** 的白話意義為何**?**

二、請透過表 **2** 中的五筆交易資料內容，並利用關聯規則**(Association Rules)**中的 **Apriori** 方法回答下列問題：

(A)找出滿足最小支持度**(Minimum Support Count)**為 **3** 的頻繁項目集**(Frequent Itemsets)**，

(B)以及滿足最小信賴度**(Minimum Confidence)**為 **50%**的關聯規則，

(C)如果**(B)**中的答案內有提升度**(lift)>1** 的規則，請標示星號特別註明，

請注意：解題時，頻繁項目集、關聯規則的產生過程全部都需要有計算過程，並保持演算法中計算的邏輯與特性。

表 **2**

TID	Items Bought

100	f, a, c, d, g, i, m, p
200	a, b, c, f, l, m, o
300	b, f, h, j, o
400	b, c, k, s, p
500	a, f, c, e, l, p, m, n

三、請回答下列問題：

(A)簡單貝氏分類法**(Naïve Bayes Classifier)**可以應用在資料分類的假設為何**?**請以表 **3** 簡易舉例說明，不用證明。

(B)商業資料經過各式各樣的分類技術確認其類別後的主要目的為何**?**

(C)承上題**(B)**的答案，請使用 **Naïve Bayes Classifier** 並回答表 **3** 中的兩個問號答案是**?**一定要有計算過程否則不給分。

表 **3**

ID	婚姻	年齡層	收入	是否購買不動產
1	已婚	青年	低	有
2	已婚	中年	高	無
3	單身	中年	高	無

4	單身	青年	高	有
5	已婚	中年	中	有
6	單身	中年	低	有
7	單身	青年	高	無
8	已婚	青年	高	無
9	已婚	中年	高	有
10	已婚	青年	高	有
11	已婚	中年	高	?
12	單身	青年	中	?

四、請簡單說明圖 1 的意涵

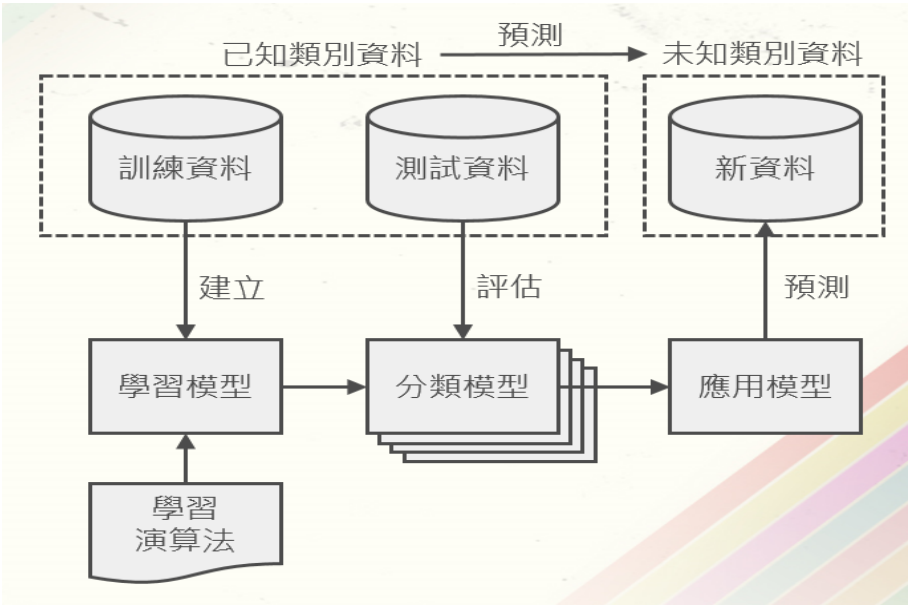

圖 1

五、當目標變數只有兩個類別資料值的時候，請根據圖 **2** 以及下列三條公式，請問該如何描述乾淨**(**混亂**)**的程度？

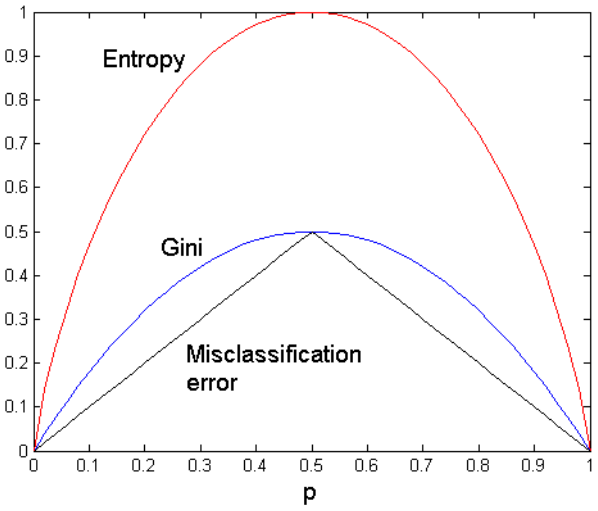

圖 2

三個亂度衡量指標的定義如下

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

$$Misclassification\ error(t) = 1 - \max_i [p(i|t)]$$

六、請使用第五題的 misclassification error 亂度公式以及表 4 資料集分別建立決策樹分類模型以及產出測試階段的混淆矩陣、Accuracy、Precision、Recall 以及 F1-Measure。

七、請利用 K 最鄰近法(KNN)、表 4 資料集以及下列距離公式回答下列問題：

$$距離 = \frac{屬性值相異個數}{全部屬性個數}$$

(A)當 K=1 時的混淆矩陣(Confusion Matrix)

(B)當 K=3 時的混淆矩陣(Confusion Matrix)

(C)請計算說明(A)與(B)答案中不同 K 值條件下，Precision、Recall 以及 F1-Measure 在"是否還款"上的差異為何?

表 4

客戶編號	是否負債	性別	婚姻狀況	收入	是否還款
1	是	男	單身	低	否
2	否	女	單身	低	否
3	是	男	單身	高	是
4	否	女	結婚	低	是
5	否	男	單身	高	是

6	是	女	單身	高	否
7	否	女	結婚	低	是
8	是	男	結婚	高	否
9	否	男	單身	低	是
10	是	女	結婚	低	否
11	否	女	結婚	高	是
12	是	男	結婚	高	否
13	是	男	單身	高	是

八、請利用表 **6-1**、圖 **6-6**、圖 **6-7** 以及下列新客戶**(**顧客編號 **15)**說明何謂**”**過度配適**(Overfitting)”**的觀念。新顧客資料如下：顧客編號 **15**，年紀 **19** 歲，女性，已婚，居住於鄉鎮中，忠誠度目前未知 **(unknown)**。

表 6-1　分類分析之範例資料

顧客編號	顧客屬性				類別
	居住區域	年紀	婚姻狀況	性別	（忠誠度）
1	市區	小於 21	已婚	女	低
2	市區	小於 21	已婚	男	低
3	市郊	小於 21	已婚	女	高
4	鄉鎮	21 至 30	已婚	女	高
5	鄉鎮	大於 30	未婚	女	高
6	鄉鎮	大於 30	未婚	男	低
7	市郊	大於 30	未婚	男	高
8	市區	21 至 30	已婚	女	低
9	市區	大於 30	未婚	女	高
10	鄉鎮	21 至 30	未婚	女	高
11	市區	21 至 30	未婚	男	高
12	市郊	21 至 30	已婚	男	高
13	市郊	小於 21	未婚	女	高
14	鄉鎮	21 至 30	已婚	男	低

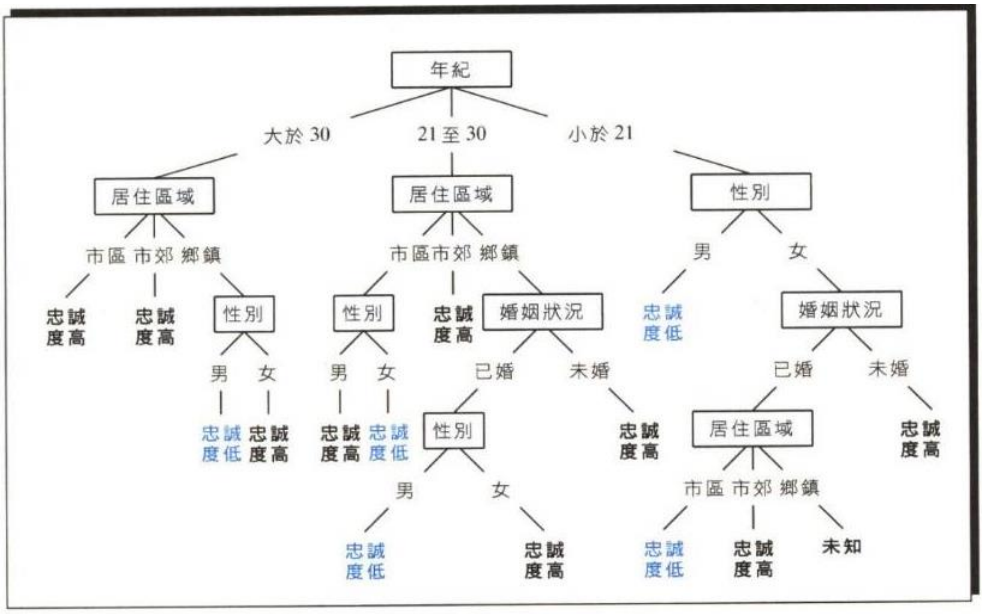

圖 6-6　可正確描述表 6-1 中資料的複雜決策樹

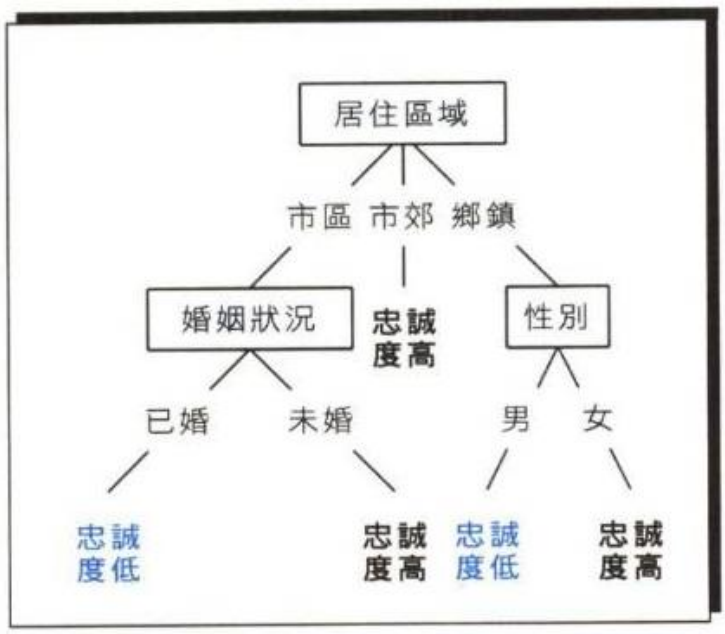

圖 6-7　可正確描述表 6-1 中資料的簡單決策樹

九、**ANN** 技術基本上為深度學習方法的基礎，請回答下列問題

(A)請簡單說明 **ANN** 技術處理資料常見被詬病的缺點為何**?**

(B)請簡單回答或者繪圖說明感知機**(Perceptron)**可以解決的資料分類問題有哪些**? (**可以繪圖說明**)**

(C)請簡單回答或者繪圖說明感知機**(Perceptron)**不能解決的問題為何**?** **(**可以繪圖說明**)**

十、假設講義中的這一題，各個參數值都不變，僅將學習速率分別改成 **0.5** 與 **1** 之後，請重新解題，說明差異。

範例 8.1 倒傳遞學習演算法的計算範例

圖 8.5 為一個多層前饋式類神經網路的範例，假設學習速率為 0.9，初始的權重值及偏移量顯示在表格 8.1 中。第一個餵入的訓練資料值組為 $X=(1,0,1)$，它對應的類別標籤為 1。

此範例說明當此值組餵入網路後，倒傳遞演算法詳細的計算步驟，首先，計算網路中每一個單元的輸入值與輸出值，這些輸入 / 輸出數值的計算顯示在表格 8.2 中，接著，計算每一個單元的錯誤值，並將錯誤值向後傳遞，這些錯誤值的計算顯示在表格 8.3 中，最後，進行權重值與偏移量的更新，其計算過程顯示在表格 8.4 中。

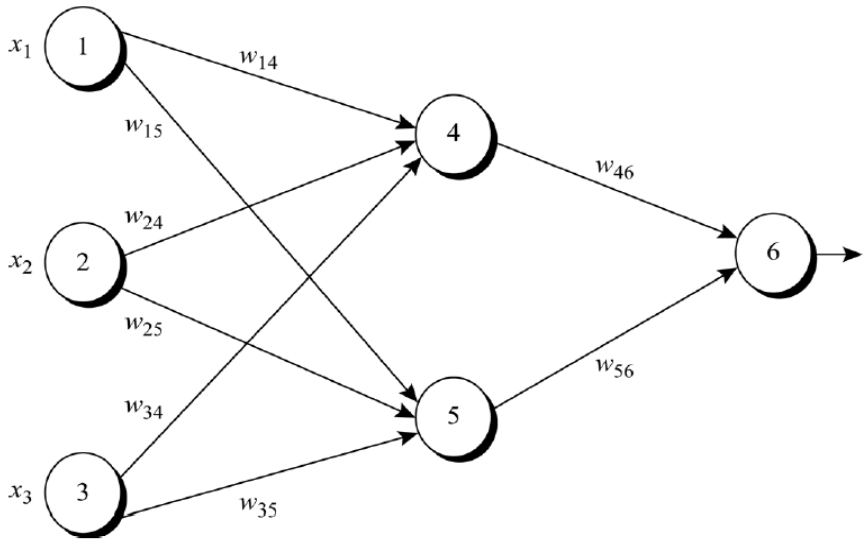

圖 8.5 多層前饋式類神經網路的範例。

表 8.1 輸入、初始權重值與偏移量

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	–0.3	0.4	0.1	–0.5	0.2	–0.3	–0.2	–0.4	0.2	0.1