亞洲大學 生物與醫學資訊學系

99 學年度第 2 學期 (期中考)

科目：資料探勘 (Data mining)

考試日期：

地點：

一、選擇與填充 (40%)

1. 資料探勘的目的是在資料中發現前所未知的趨勢與_**樣式**_
2. 彙總統計屬於量化的資料包括資料的次數，資料的_**落點**_和資料的分佈
3. 當相關係數爲 0 表不具_**線性**_關係
4. We are drowning in data, but starving for knowledge. 其中 drowning 爲 A) 丟棄, B) 拖曳, **C) 淹沒**, D) 飢餓
5. _**眾**_數是指具有最高次數的屬性值
6. 變數轉換時的正規化目的是將變數轉換成_**常態**_分配，以使整個值的單位一致
7. 當距離公式滿足正向性、對稱性及_**三角不等式**_時，則稱爲 metrics (度量)
8. Analysis 是 A)安裝, **B) 分析**, C) 分享, D) 設定
9. Euclidean distance 指的是 A) 敏可夫斯基距離, **B) 歐氏距離**, C) 漢明距離, D) 曼哈頓距離
10. 特徵的產生有三種方法，包括將資料映射到新的空間、特徵的建構與_**特徵的萃取**_
11. Tukey 所發展的 EDA 將分群分析與_**異常偵測**_皆視爲資料探索技術
12. 關係係數值介於-1 到 1，當值爲 1 時表_**完全正**_相關
13. Diaper 是 A) 化妝棉, **B) 尿布**, C) 溼紙巾, D) 面紙
14. 屬性型態可大致分爲定性與_**定量**_
15. 一個屬性是指物件的_**特性**_，而其特性可能會隨時間而變動
16. 將資料視爲將資料視爲多維度資料的步驟有二個，包括維度的識別與_**屬性**_識別
17. 線上分析處理強調_**互動**_性的分析資料，並且提供視覺化資料的能力及產生彙總統計的資訊，因此 OLAP 系統適合做爲多維度資料分析的主要分析方法
18. 大部分的資料都可視爲一個表單，而表單中的每一列爲一個_**物件**_，同時每一行爲一個屬性
19. Slicing 指的是 A) 滑球, B) 投影片, **C) 切片**, D) 切割
20. 所有可能的多維度聚集總和，稱爲_**資料立方體**_，其名稱、每個維度大小並示相等

二、簡答題(20%)

1. 資料探索的主要目的爲何? (5%)
   甲、幫助選擇適當的前處理方式以及資料探勘的技術
   乙、協助解決資料探勘的問題，人們可以利用視覺化的方式來發現樣式以及進行結果的解釋
2. 資料特性裡說的「稀疏性」指的是什麼? (5%)
   對一些非對稱屬性資料而言，也許僅 1%的資料是不爲 0；可是實際上，因爲只有非 0 的數值需要被儲存和運算，因此節省很多時間和儲存空間，所以也算是稀疏資料的一項優點

3. 何謂「離群值」? (5%)
   離群值可能是因爲資料物件的某些特性和其他物件不一樣，或者是其屬性值較不常出現在其他物件中

4. 試舉兩點縮減資料維度的目的(5%)
   甲、避免維度的問題
   乙、降低資料探勘演算法所需的時間和記憶體
   丙、讓資料更易於用視覺化方式呈現出來
   丁、協助刪除掉一些無關的特徵或是雜訊值

三、問答與計算(40%)

1. 請解釋以下所敘述的內容是否爲資料探勘的分析工具。
   a. 將公司的顧客依性別來區分
      不是，屬資料整理
   b. 將學生依學號排序
      不是，屬資料整理
   c. 從歷史記錄中預測公司未來股價
      是，從分析資料找出未知的趨勢或樣式
   d. 預測擲骰子的結果
      不是，因爲結果已知 (骰子機率固定)
   e. 計算公司的總銷售
      不是，屬於資料整理

2. 令 p1 =(2, 3), p2 = (3, 1), p3=(1, 2), p4=(0, 1)，請分別算出歐幾里德距離矩陣與閔可夫斯基之 L1, 與 L∞之距離矩陣

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (1)$$

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{1/r} \quad (2)$$

$$\lim_{r\to\infty}\left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{1/r} = \max_{i=1}^{n}|x_k - y_k| \quad (3)$$

$$d(p1,p2) = \sqrt{(2-3)^2 + (3-1)^2} = 2.24$$
$$d(p1,p3) = \sqrt{(2-1)^2 + (3-2)^2} = 1.41$$
$$d(p1,p4) = \sqrt{(2-0)^2 + (3-1)^2} = 2.83$$
$$d(p2,p3) = \sqrt{(3-1)^2 + (1-2)^2} = 2.24$$
$$d(p2,p4) = \sqrt{(3-0)^2 + (1-1)^2} = 3$$
$$d(p3,p4) = \sqrt{(1-0)^2 + (2-1)^2} = 1.41$$

	P1	P2	P3	P4
P1	0	2.24	1.41	2.83
P2	2.24	0	2.24	3
P3	1.41	2.24	0	1.41
P4	2.83	3	1.41	0

L1

d(p1, p2) = |2-3| + |3-1| =3

d(p1, p3) = |2-1| + |3-2| =2

d(p1, p4) = |2-0| + |3-1| =4

d(p2, p3) = |3-1| + |1-2| =3

d(p2, p4) = |3-0| + |1-1| =3

d(p3, p4) = |1-0| + |2-1| =2

	P1	P2	P3	P4
P1	0	3	2	4
P2	3	0	3	3
P3	2	3	0	2
P4	4	3	2	0

L∞

d(p1, p2) = max{|2-3|, |3-1|} =2

d(p1, p3) = max{|2-1|, |3-2|} =1

d(p1, p4) = max{|2-0|, |3-1|} =2

d(p2, p3) = max{|3-1|, |1-2|} =2

d(p2, p4) = max{|3-0|, |1-1|} =3

d(p3, p4) = max{|1-0|, |2-1|} =1

	P1	P2	P3	P4
P1	0	2	1	2
P2	2	0	2	3
P3	1	2	0	1
P4	2	3	1	0

3. 令 x = {1, 3, 1, 2, 1, 1}, y = {1, 0, 1, 2, 0, 4}, cos(x, y), EJ

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \tag{4}$$

$$EJ(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x}\cdot\mathbf{y}} \tag{5}$$

$x \cdot y = 1\times1 + 3\times0 + 1\times1 + 2\times2 + 1\times0 + 1\times4 = 10$

$\|x\| = \sqrt{1\times1 + 3\times3 + 1\times1 + 2\times2 + 1\times1 + 1\times1} = 4.12$

$\|y\| = \sqrt{1\times1 + 0\times0 + 1\times1 + 2\times2 + 0\times0 + 4\times4} = 4.69$

Cos(x,y) = 10/(4.12*4.69) = 10/19.32 = 0.52

EJ(x,y) = 10 / (4.12^2 + 4.69^2 - 10) = 10 / (16.97 + 22 – 10)=0.35

4. 一組資料 {4, 3, 2, 3, 3, 4, 5, 2}下求列資料
   a. 全距
   b. 平均數
   c. 中位數
   d. 眾數
   e. 絕對平均離差 (AAD)
   f. 中位數絕對離差 (MAD)
   g. 四分位差 (IQR)

a. max{4, 3, 2, 3, 3, 4, 5, 2} - min{4, 3, 2, 3, 3, 4, 5, 2} = 5 – 2 = 3

b. (4 + 3 + 2+ 3 + 3 + 4 + 5 + 2) / 8 = 26 / 8 = 3.25

c. 排序後 {2, 2, 3, 3, 3, 4, 4, 5} 中位數爲 (3+3)/2 = 3

d. 眾數

   2 出現 2 次

   3 出現 3 次

   4 出現 2 次

   5 出現 1 次

   眾數爲 3

e. (|2-3.25|+|2-3.25|+|3-3.25|+|3-3.25|+|3-3.25|+|4-3.25|+|4-3.25|+|5-3.25|)/8

   = (1.25+1.25+0.25+0.25+0.25+0.75+0.75+1.75)/8 = 0.8125

f. {|2-3.25|, |2-3.25|, |3-3.25|, |3-3.25|, |3-3.25|, |4-3.25|, |4-3.25|, |5-3.25|}

   ={1.25, 1.25, 0.25, 0.25, 0.25, 0.75, 0.75, 1.75}

   排序後得 { 0.25, 0.25, 0.25, 0.75, 0.75, 1.25, 1.25, 1.75}

   中位數爲 (0.75 + 0.75) /2 = 0.75

g. $x_{100\%}$ =max{4, 3, 2, 3, 3, 4, 5, 2} = 5

   $x_{0\%}$ =min{4, 3, 2, 3, 3, 4, 5, 2} = 2

   $x_{50\%}$ =(5 + 2) /2 = 3.5

   $x_{75\%}$ =(5 + 3.5) /2 = 4.25

   $x_{25\%}$ =(3.5 + 2) / 2 = 2.75

   IQR = $x_{75\%}$ - $x_{25\%}$ =2.25-2.75 = 1.5