資料探勘期中考(2021/11/23)

1. （20％）(關聯法則）給予一個交易資料表如下，請回答下列各子題：

表一、交易資料表

|  |  |
| --- | --- |
| 交易編號 | 交易項目 |
| 1 | {A, B, C, D} |
| 2 | {A, C, D} |
| 3 | {A, B, C, D, E} |
| 4 | {B, C, D} |
| 5 | {A, B, D} |
| 6 | {B, C, D} |
| 7 | {B, D, E} |
| 8 | {B, C,E} |

1. 假設最小支持度為0.4，請列出所有的Frequent Itemsets
2. 計算{B, C} -> {D} 的lift與信賴度(confidence)
3. (20%) (迴歸) 表二為一個包含X, Y兩個項目的資料表，建立一個以X為自變數，Y為因變數的迴歸方程式(15%)。請計算此一迴歸方程式的**R2**值 （5%）

表二、XY資料集

|  |  |
| --- | --- |
| X | Y |
| 2 | 5 |
| 2 | 6 |
| 3 | 5 |
| 3 | 8 |
| 4 | 7 |
| 4 | 10 |
| 5 | 10 |
| 5 | 12 |

Mean(Y) =7.875, Mean(X)=3.5

b1=Sxy/Sxx；b0=mean(y)-b1\*mean(x)；

R2=S2xy/(Sxx\*Syy)=SSR/SST=(SST-SSE)/SST; Sxy=cov(X,Y); Sxx=Varance(X) (先算 b1, 再算 b0，建立迴歸 方程式)









3. (20%)有一個資料集包含10個數值 5, 7, 11,25, 23, 26, 18, 19, 15,16；請回答下列兩個子題

a. 以階層式分群 (hierarchical clustering) 法將這些資料分成三群，使用Complete Link方式計算群與群的距離，劃出樹狀圖

b. 以階層式分群法將這些資料分成三群，使用Single Link 方式計算群與群的距離，劃出樹狀圖

4. (20%) 為下列資料集建立兩層的決策樹，並計算該決策樹對資料集的正確率；class欄位為類別欄位，A, B, C為屬性欄位

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 編號 | A | B | C | Class |
| 1 | 0 | 0 | 1 | P |
| 2 | 1 | 0 | 1 | P |
| 3 | 0 | 1 | 0 | P |
| 4 | 1 | 0 | 0 | P |
| 5 | 1 | 0 | 1 | P |
| 6 | 0 | 1 | 1 | N |
| 7 | 1 | 1 | 0 | N |
| 8 | 0 | 0 | 0 | N |
| 9 | 0 | 1 | 0 | N |
| 10 | 1 | 1 | 1 | N |

5. (20%) 問答題

a. 甚麼是資料探勘 (data mining)? (10%)

b. 請舉例說明，那些工作是監督式學習 (supervised learning)? 那些工作是非監督式學習 (unsupervised learning)? (請從”關聯法則”、”回歸”、”分類”、”分群”中舉例) (10%)