108年公務人員特種考試警察人員、一般警察人員考試及108年特種考試交通事業鐵路人員、退除役軍人轉任公務人員考試試題

考 試 別：一般警察人員考試
等　　別：二等考試
類 科 別：刑事警察人員犯罪分析組
科　　目：資料探勘技術（包括資料庫管理與運用、線上交易處理【OLTP】、資料倉儲【Data Warehouse】、資料探勘【Data Mining】）
考試時間：2 小時　　座號：_____

※注意：㈠禁止使用電子計算器。
㈡不必抄題，作答時請將試題題號及答案依照順序寫在試卷上，於本試題上作答者，不予計分。
㈢本科目除專門名詞或數理公式外，應使用本國文字作答。

一、假設關聯資料庫的表格 R（A,B,C,D,E,F）有六個屬性 A,B,C,D,E,F，各屬性均無多值（Multi-Value）現象，其功能相依（Functional Dependency）有下列兩條：

FD1：{A,B} → {C,D,E,F}

FD2：C → B

㈠請以屬性封閉性（Closure）的概念，找出 R 的所有候選鍵（Candidate Key）（6 分）

㈡請證明 R 不滿足 Boyce-Codd 正規化（BCNF）。（3 分）

㈢請試圖將 R 分割，並先找出你分割出來每一表格的所有候選鍵，再證明分割出來的每一表格均滿足 BCNF，且同時證明你的分割滿足 lossless（無損）join 特性。（12 分）

㈣你是否會建議你上述的分割？為什麼？（4 分）

二、假設有個資料庫記錄了對全國某種受刑人數萬人做過的某次心理測驗，B、C、D、E、F 分別代表其具有某種行為傾向。以下 P{α}代表受刑人有 α 行為傾向的機率，P{α,β}代表受刑人同時有 α 與 β 行為傾向的機率。

P{B}= 0.08, P{C}= 0.06, P{D}= 0.04, P{E}= 0.07, P{F}= 0.02,
P{B,C}= 0.04, P{ B,D}= 0.04, P{B,E}= 0.06, P{B,F}= 0.02, P{C,D}= 0.04,
P{C,E}= 0.04, P{C,F}= 0, P{D,E}= 0.02, P{D,F}= 0, P{E,F}= 0.2

我們欲進行關聯規則（Association Rule）的資料探勘：

㈠請先解釋何謂支持度（Support）、信心度（Confidence）的概念。（6 分）

㈡假設支持度最低門檻是 0.05、信心度最低門檻是 0.7，請指出上述那些是 Large-1、Large-2 的項目集合（Item-set）；並找出所有只包含 2 個項目集合的強（Strong）關聯規則。（14 分）

㈢在尋找關聯規則時，有個重要的反單調（Anti-monotonicity）特性可減低運算成本，請先說明何謂此特性？再請以上述例子來說明應如何運用此特性。（5 分）

三、假設我們對某種犯罪資料要進行研究，資料庫收集了 4,000 筆個人的心理、行為、參與社群等詳細資料，其中 1,900 人實際有過該犯罪事實，2,100 人則無該犯罪事實。使用兩種方法來做集群（Cluster）分析。α 方法可分出 1,400 位犯罪人，但其中 100 位未有犯罪事實，但被錯誤歸為此犯罪群；此外有 600 位實際有過犯罪事實，卻未被歸為此群。β 方法可分出 1,600 位犯罪人，但其中 200 位未有犯罪事實，但被錯誤歸為此犯罪群；此外有 500 位實際有過犯罪事實，卻未被歸為此群。請問應如何評估此兩方法的優劣？你會建議選擇那個方法？為什麼？（25 分）

四、歡樂暢飲公司是一間行銷全世界的茶飲料公司，它的資料庫至少記錄了 2000-2018 年的 30 種產品每季在全世界各地區的銷售數量與金額。請以此為背景來說明資料倉儲的下列一些概念：

㈠何謂主題導向（Subject-Oriented）？此處的主題是什麼？（3 分）

㈡在為它建立模型時，有所謂的事實表格、維度表格，請舉例說明。（3 分）

㈢資料倉儲操作上有「向上擷取」（Roll-Up）及「向下探究」（Drill-Down）功能，這與維度設計有何關係？（3 分）

㈣請依此背景，設計出星型模式（Star Schema）（8 分）

㈤請依此背景，設計出雪花模式（Snowflake Schema）（8 分）