

107年公務人員特種考試警察人員、一般警察人員考試及
107年特種考試交通事業鐵路人員考試試題

代號：20230

全一張
(正面)

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：(一)可以使用電子計算器。

(二)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。

(三)本科目除專門名詞或數理公式外，應使用本國文字作答。

- 一、一犯罪資料庫記錄前科犯所曾經犯過的不同刑案，資料如下：共 5 個欄位 A、B、C、D、E，有標 1 者分別代表該前科犯曾經犯過該刑案。試使用關聯規則探勘(association rule mining)方法，找出此五種犯罪刑案任二種間具最高支持度(support)與信心度(confidence)的前三條規則，並列出各規則的支持度和信心度的值。(25 分)
(注意：1.支持度高者優先列出，相同支持度時，請依信心度排列；2.必須寫出探勘過程。)

前科犯	A (偷竊)	B (恐嚇)	C (搶劫)	D (傷害)	E (販毒)
001	1			1	1
002		1		1	
003		1	1	1	
004	1		1	1	1
005	1				
006	1	1			1
007			1	1	
008	1		1		1
009			1		
010		1	1		

- 二、分類演算法(classification algorithm)是資料探勘中重要的分析方法。針對以下的訓練資料集：9 筆資料 $x_1, x_2, \dots, x_9$ ，每筆資料有 4 個屬性(A1, A2, A3, Class)，其中 A1、A2、A3 為特徵屬性(feature)，Class 為類別屬性。

ID	A1	A2	A3	Class
x_1	S	2	Y	C1
x_2	B	2	Y	C1
x_3	B	1	Y	C1
x_4	S	1	Y	C1
x_5	S	3	Y	C2
x_6	B	3	N	C2
x_7	B	2	N	C2
x_8	B	3	Y	C2
x_9	S	2	N	C2

現今希望用決策分類樹(decision tree)來學習該資料集的分類規則，其中分類樹內部節點(internal node)屬性以選擇最小的 $Gini(A)$ 值來決定分支， $Gini(A)$ 函數定義如下：

$$Gini(A) = \sum_{t \in A} p(t) \cdot [1 - \sum_{i=1}^C (p(i|t))^2],$$

t 為屬性 A 中的某特定值， $p(t)$ 是屬性 A 中值為 t 所出現的資料比率， $p(i|t)$ 是在屬性值為 t 時，資料集隸屬於類別 i 的比率， C 為所有類別的個數。分支的停止條件則是當節點資料完全隸屬同一類別時停止，該樹葉節點(leaf node)即是分類結果。

請以此訓練資料集和 $Gini(A)$ 屬性決策條件要求，產生一決策分類樹(decision tree)，答案中請清楚表達該決策樹的每一內部節點的屬性、分支條件值和該決策點的 $Gini(A)$ 值，以及葉節點的分類結果。(25 分)

(請接背面)

107年公務人員特種考試警察人員、一般警察人員考試及
107年特種考試交通事業鐵路人員考試試題

代號：20230

全一張
(背面)

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】)

三、某公路交通事故資料庫系統的部分相關關聯式表格名稱、欄位屬性名稱和資料如下：

ACC [事故紀錄]

事故編號	發生日期	發生時間	道路編號	道路位置	事故類別	受傷人數	死亡人數
ACC_ID	DATE	TIME	R_ID	R_KM	ACC_C	H_NUM	D_NUM
2100001	02/01/2016	05:13	T9	152	A1	4	1
2100002	02/12/2016	08:30	N1	178	A2	1	2
2100003	02/07/2016	18:01	N3	352	A3	0	0
2100004	05/16/2016	04:58	T9	152	A1	5	1
2100006	09/02/2016	05:32	N1	175	A3	0	0
2100007	09/12/2016	07:21	T3	202	A2	4	0
2100009	09/15/2016	07:55	N1	58	A2	2	0
2100010	10/01/2016	17:30	T1	312	A3	0	0
2100011	10/02/2016	01:32	C118	13	A2	6	1
2100013	12/12/2016	05:58	T3	202	A1	1	3

(一)請寫出 SQL 語法查詢(query)：列出曾發生單件事故中有受傷 3 人(含)以上或有死亡人數的道路編號。(10 分)

(二)請寫出 SQL 語法查詢(query)：列出各個道路編號的受傷人數與死亡人數統計數據。(10 分)

四、承續上題，交通管理單位高層決策希望能夠從 ACC 資料庫中，建立線上分析處理(On-line analytical processing; OLAP)系統，分析各道路容易出現重大事故的路段與時間，以利掌握並研擬改善措施。為了達成目標，首先必須構建資料倉儲(data warehouse)。假設所要分析的主要目標資料項包括：事故件數(ACC_NUM)、受傷人數(H_NUM)和死亡人數(D_NUM)；分析的維度包括有：事故日期、事故時段和道路種類、事故類別等 4 個維度(dimension)，分述如下：

(1)事故日期(ACC_DATE)：本維度中分析發生日期(DATE)，包含有日期的概念階層(concept hierarchy)；如下：

年(year) > 季(quarter) > 月(month) > 日(day)

(2)事故時段(ACC_TIME)：本維度分析發生時間(TIME)，包含上下午(Half_day)與尖、離峰時段(Rush_hour)

上午：0 時至 12 時，下午：12 時至 24 時；

尖峰：7 時至 9 時及 17 時至 19 時；其餘為離峰。

(3)道路種類(ROAD)：本維度分析道路編號(R_ID)，並將道路類別(Road_class)分為：國道(N)、省道(T)、縣道(C)

(4)事故類別(ACC_C)：本維度主要依事故類別(ACC_C)，A1、A2、A3 分類，無額外分析需求。

請應用資料倉儲模型(data warehouse model)為此 OLAP 系統設計資料倉儲綱要(schema)，並據以回答以下問題：

(一)請您繪出所設計的資料倉儲綱要圖，並明確指出所包含的事實表格(fact table)、維度表格(dimension table)，以及相對的所有參考屬性(referential attribute)。(20 分)

(二)請對 ACC 資料表寫出 SQL 語法查詢(query)：計算出在分析維度(尖峰時段,事故類別)的各類事故件數。(10 分)