

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 1 頁，共 11 頁

單選題 50 題 (佔 100%)

D	1. 關於遺缺值 (NA) 的處理方式，下列敘述何者較「不」正確？ (A) 以貝氏定理公式計算最可能的值填入 (B) 以決策樹歸納法計算最可能的值填入 (C) 以迴歸分析計算最可能的值填入 (D) 無須考慮遺缺值比例，全部刪除
A	2. 將非結構化的資料轉變為結構化的資料，這樣的過程屬於下列何種工作？ (A) 資料前處理 (B) 資料標準化 (C) 資料視覺化 (D) 資料載入
A	3. 假設您每分鐘都會收到某張股票的開盤價、收盤價、最低價、最高價、成交量，若您只想儲存收盤價，最適合 R 語言中的何種結構？ (A) 向量 (Vector) (B) 矩陣 (Matrix) (C) 字串 (Character) (D) 資料框架 (Data frame)
D	4. 為找出某一篇英文文章中較為正確且重要的詞頻 (如 Cat 與 Cats 均併做 Cat 計算)，下列何者「不」是必要的步驟？ (A) 移除停用字 (Stop Words) (B) 詞幹提取 (Stemming) (C) 詞形還原 (Lemmatization) (D) 詞性標記 (Part of Speech Tagging)
C	5. 考慮某資料欄位為銷售地區，資料包括北部、中部、南部與離島。如果須使用單熱編碼 (One-Hot Encoding)，則離島值最合適的編碼為何？ (A) [1 1 1 1] (B) [1 0 0 1] (C) [0 0 0 1] (D) [1 1 1 0]
A	6. 下列何種圖表最適合用來展示資料中各類型數據所佔比例？ (A) 圓餅圖 (Pie chart) (B) 散點圖 (Scatter plot) (C) 折線圖 (Line chart) (D) 長條圖 (Bar chart)
C	7. 下列何種圖表最適合用來展示時間序列 (Time Series) 類型的資料？

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 2 頁，共 11 頁

	(A) 圓餅圖 (Pie chart) (B) 散佈圖 (Scatter plot) (C) 折線圖 (Line chart) (D) 長條圖 (Bar chart)
D	8. 關於將資料去識別化，下列敘述何者「不」正確？ (A) 將姓名轉換成 MD5 雜湊值是一種去識別化的方式 (B) 將年齡資料，例如：「39 歲」轉換成「>35 歲」是一種去識別化的方式 (C) 將資料進行分群後，取平均或中位數取代原本的資料也是一種去識別化的方式 (D) 所有去識別化的方法百分之百無法找回原本的訊息
D	9. R 語言中，下列函數何者可以回傳資料向量中各百分位數？ (A) var() (B) sd() (C) mean() (D) quantile()
C	10. 關於資料彙總 (Data Aggregation)，下列敘述何者最為正確？ (A) 可降低資料尺度、資料偏斜性對於模型的不良影響 (B) 是運用推論統計學，以展現資料的基本特質 (C) 是以摘要的形式收集或呈現資訊的任何過程 (D) 可以統整不同尺度之連續屬性間的數值分佈
C	11. 資料縮減 (Data Reduction) 包括屬性挑選 (Feature Selection) 與屬性萃取 (Feature Extraction)，下列何者「不」是屬性萃取的方法？ (A) 主成份分析 (Principal Component Analysis) (B) 偏最小平方法 (Partial Least Squares) (C) 相關係數矩陣視覺化 (Correlation Matrix Visualization) (D) 因素分析 (Factor Analysis)
D	12. 關於屬性萃取 (Feature Extraction)，下列敘述何者「不」正確？ (A) 可以消除屬性間的相互影響，增加模型的效果 (B) 屬性萃取方法有監督式與非監督式兩種 (C) 能夠降低屬性的個數，提升運算效率 (D) 屬性挑選 (Feature Selection) 與屬性萃取不同的是，前者最後決定出的屬性，與原來的屬性有函數關係
A	13. 假設您要對一含有數百個生物特徵屬性的資料進行分析，可使用下列何種方法來萃取重要的訊息？ (A) 主成分分析 (Principle Component Analysis)

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 3 頁，共 11 頁

	(B) K 近鄰法 (K-nearest neighbors) (C) K 均值法 (K-means) (D) 關聯規則 (Association Rule)
A	14. 在資料處理中，有些資料可能不是連續型的數值，而是一些分類值，例如職業、性別等，對於這樣的特徵值，我們將該屬性的欄位的各種狀態，設置獨立的欄位，並在發生該狀態的欄位中，填入 1，請問此種資料轉換的方式稱為？ (A) 單熱編碼 (One-Hot Encoding) (B) 動態規劃 (Dynamic Programing) (C) 最大似然估計 (Maximum Likelihood Estimation) (D) 正規化 (Normalization)
B	15. 關於正規化 (Normalization)，下列敘述何者正確？ (A) 一定落在[-1, 1]區間內 (B) 為了消除數據特徵之間的量綱影響 (C) 針對類變變量進行處理 (D) 會影響資料原來的分佈情況
A	16. 透過 Web 瀏覽器上傳信用卡資料時，下列何者為最合適的請求方法？ (A) POST (B) PUT (C) GET (D) UPLOAD
A	17. 下列敘述何者「不」正確？ (A) GET 方法傳輸速度較 POST 慢 (B) POST 方法適合傳送較為隱私的資料 (C) POST 方法允許傳送 GET 方法更多的資料 (D) GET 和 POST 皆可以將資料送到 Web Server 端
A	18. 下列何者為進行資料分析時的首要步驟？ (A) 資料收集 (B) 資料清理 (C) 資料建模 (D) 資料分析
C	19. 關於集中式資料庫系統 (Centralized Database) 與分散式資料庫系統 (Distributed Database)，下列敘述何者正確？ (A) 集中式資料庫系統中，資料一般是存放在多台伺服器 (B) 分散式資料庫系統，容量擴充不易且資料很容易受到某一台伺服器毀損而遺失

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 4 頁，共 11 頁

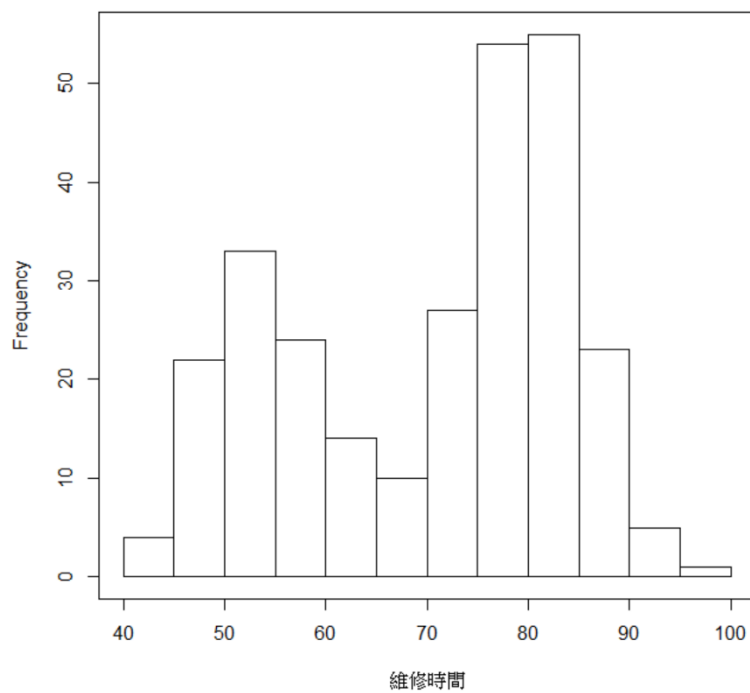
	<p>(C) 分散式資料庫系統中，某伺服器出現問題時（如斷線、當機等），仍有機會可維持資料庫運作</p> <p>(D) 集中式資料庫系統，運行較穩定不會出現問題，因此不需要定時備份</p>
C	<p>20. 建立數據分析工作流程時，除了要得到分析結果外，如何建立「有效率」、「易維護」、「可重複使用」的良好品質程式碼亦十分重要。下列敘述何者較「不」恰當？</p> <p>(A) 以 try-except 建立良好的錯誤處理機制，避免程式碼因部分錯誤而整體停止</p> <p>(B) 對輸出之靜態檔案（ex: .json, .csv, .txt, .ft）等進行結構式命名，以利後續資料存取</p> <p>(C) 為避免有太多的程式碼檔案（ex: .py, .r, .js），將 50 多個函式與主資料處理流程共 4000 行程式碼置放於同一個檔案之中進行管理</p> <p>(D) 將程式碼透過版控軟體（ex: git, VSS）進行紀錄，建立良好的開發分支與版本紀錄</p>
D	<p>21. 若欲比較兩公司員工薪資之離散程度，可採用下列何者統計量？</p> <p>(A) 變異數</p> <p>(B) 全距</p> <p>(C) 平均數</p> <p>(D) 變異係數</p>
A	<p>22. 「林書豪的球衣號碼」屬於下列那一種量度尺度分類？</p> <p>(A) 名目尺度（Nominal Scale）</p> <p>(B) 順序尺度（Ordinal Scale）</p> <p>(C) 比率尺度（Ratio Scale）</p> <p>(D) 區間尺度（Interval Scale）</p>
D	<p>23. 關於連續型機率分配，下列敘述何者正確？</p> <p>(A) 常態分配中，平均值為 0、變異數為 0 之分配，稱為標準常態分配</p> <p>(B) 已知均勻分配為 $U(a, b)$，則平均值為 $(a-b)/2$</p> <p>(C) 伽瑪分配是指數分配的特例</p> <p>(D) 已知隨機變數為標準常態分配，則取其平方為卡方分配且自由度為 1</p>
B	<p>24. 附圖為某機器維修時間之次數分配所繪製之結果，下列敘述何者正確？</p>

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 5 頁，共 11 頁



- (A) 圖形名稱為長條圖 (Bar Chart)
- (B) 維修時間出現最多的範圍為 80~85
- (C) 資料呈現常態分配
- (D) 資料平均值為 55

A 25. 變異數分析是檢定二個以上母體的何者統計量是否相等？

- (A) 平均數 (Mean)
- (B) 變異數 (Variance)
- (C) 標準差 (Standard Deviation)
- (D) 中位數 (Median)

A 26. 附圖為模型複雜度 (Model Complexity) 與預測誤差 (Prediction Error) 之間的變化關係，下列敘述何者正確？

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 6 頁，共 11 頁

	<p>(A) 戊段表過度配適 (Overfitting)，它代表模型越複雜時與訓練集配適的過好，但卻逐漸喪失對測試集的預測能力</p> <p>(B) 實曲線甲為測試集 (Test set) 樣本下的模型複雜度與預測誤差之間的變化關係</p> <p>(C) 虛曲線乙為訓練集 (Training set) 樣本下的模型複雜度與預測誤差之間的變化關係</p> <p>(D) 丙段表配適不足 (Underfitting)，此時訓練集預測誤差表現不佳，而測試集預測誤差表現良好</p>
C	<p>27. 請問下列何者運算後是類別型變數資料？</p> <p>(A) 銷售金額除以年齡</p> <p>(B) 銷售金額開根號</p> <p>(C) 性別數值化後取負號</p> <p>(D) 氣溫的平方</p>
C	<p>28. 當 A、B 的共變異數 (Covariance) $Cov[A, B] = 5$ 時，請問 $Cov[A+2, B+1] = ?$</p> <p>(A) 8</p> <p>(B) 7</p> <p>(C) 5</p> <p>(D) 3</p>
C	<p>29. Python 語言中，關於附圖繪製直方圖，下列敘述何者正確？</p>

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 7 頁，共 11 頁

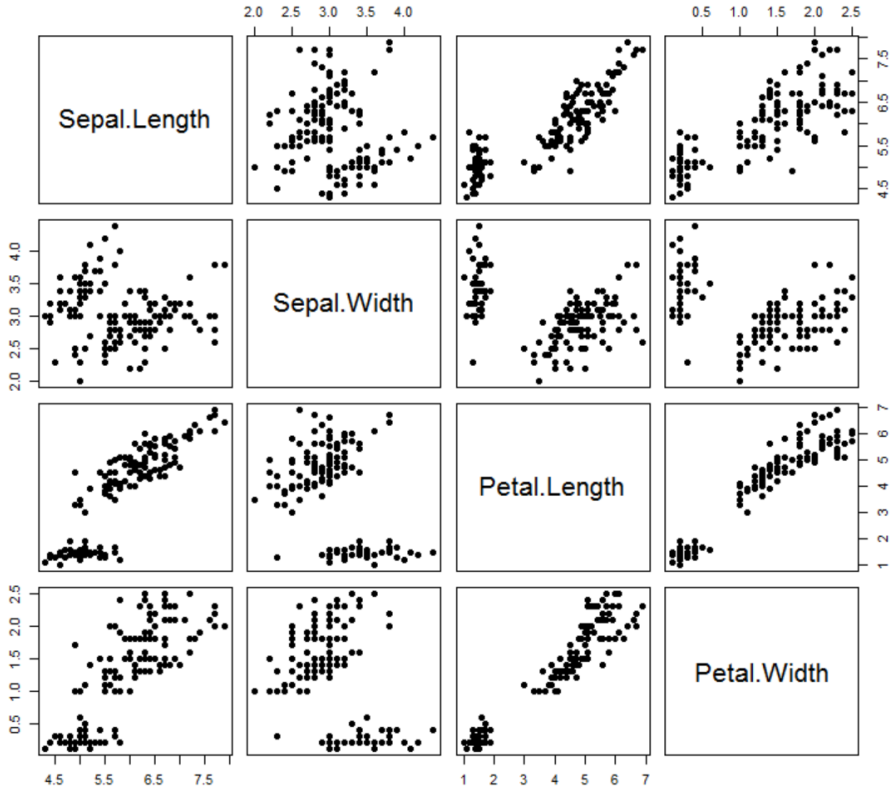
	<div><p>iPAS Data Analysis</p><table border="1"><caption>Histogram Data (Approximate)</caption><thead><tr><th>Bin Range</th><th>Frequency</th></tr></thead><tbody><tr><td>-3 to -2</td><td>10</td></tr><tr><td>-2 to -1</td><td>140</td></tr><tr><td>-1 to 0</td><td>350</td></tr><tr><td>0 to 1</td><td>360</td></tr><tr><td>1 to 2</td><td>170</td></tr><tr><td>2 to 3</td><td>110</td></tr><tr><td>3 to 4</td><td>140</td></tr><tr><td>4 to 5</td><td>190</td></tr><tr><td>5 to 6</td><td>200</td></tr><tr><td>6 to 7</td><td>150</td></tr><tr><td>7 to 8</td><td>90</td></tr><tr><td>8 to 9</td><td>40</td></tr><tr><td>9 to 10</td><td>20</td></tr></tbody></table></div> <div><p>(A) 資料呈現常態分配 (Normal Distribution)</p><p>(B) 資料有一個高峰值 (Peak)</p><p>(C) x 軸稱為組界 (Bin)，範圍為-4 至 10</p><p>(D) 數值 5 至 6 的範圍約有 350 筆資料</p></div>	Bin Range	Frequency	-3 to -2	10	-2 to -1	140	-1 to 0	350	0 to 1	360	1 to 2	170	2 to 3	110	3 to 4	140	4 to 5	190	5 to 6	200	6 to 7	150	7 to 8	90	8 to 9	40	9 to 10	20
Bin Range	Frequency																												
-3 to -2	10																												
-2 to -1	140																												
-1 to 0	350																												
0 to 1	360																												
1 to 2	170																												
2 to 3	110																												
3 to 4	140																												
4 to 5	190																												
5 to 6	200																												
6 to 7	150																												
7 to 8	90																												
8 to 9	40																												
9 to 10	20																												
A	<div><p>30. 參考附圖，Python 語言中，關於使用 <code>anova_lm</code> 函數進行不同機器 (machine)、不同操作員 (operator) 對於產量是否會有影響的二因子變異數分析，下列敘述何者「不」正確？</p><pre>In [2]: sm.stats.anova_lm(mod, typ = 2)</pre><pre>Out[2]:</pre><table><tr><th></th><th>sum_sq</th><th>df</th><th>F</th><th>PR(>F)</th></tr><tr><td>machine</td><td>1804.816667</td><td>2.0</td><td>41.666282</td><td>1.562273e-08</td></tr><tr><td>operator</td><td>60.208333</td><td>1.0</td><td>2.779958</td><td>1.084488e-01</td></tr><tr><td>machine:operator</td><td>17.504667</td><td>2.0</td><td>0.404115</td><td>6.720249e-01</td></tr><tr><td>Residual</td><td>519.792000</td><td>24.0</td><td>NaN</td><td>NaN</td></tr></table></div> <div><p>(A) 機器類型有二種</p><p>(B) 機器與操作員交互作用自由度為 2</p><p>(C) 不同機器之平均產量有顯著之差異</p><p>(D) 不同操作員之平均產量有顯著之差異</p></div>		sum_sq	df	F	PR(>F)	machine	1804.816667	2.0	41.666282	1.562273e-08	operator	60.208333	1.0	2.779958	1.084488e-01	machine:operator	17.504667	2.0	0.404115	6.720249e-01	Residual	519.792000	24.0	NaN	NaN			
	sum_sq	df	F	PR(>F)																									
machine	1804.816667	2.0	41.666282	1.562273e-08																									
operator	60.208333	1.0	2.779958	1.084488e-01																									
machine:operator	17.504667	2.0	0.404115	6.720249e-01																									
Residual	519.792000	24.0	NaN	NaN																									
B	<div><p>31. 關於資料探勘 (Data Mining)，下列敘述何者「不」正確？</p><p>(A) 利用一種或多種電腦技術來自動分析語或去擷取知識的過程</p><p>(B) 可找出遺失的歷史資料</p><p>(C) 通常被用來和知識發現相互交換使用的術語</p><p>(D) 獲得的知識通常是資料的模型或是歸納</p></div>																												
C	<div><p>32. 下列何者屬於「非監督式學習」(Unsupervised Learning) 演算法？</p><p>(A) 決策樹 (Decision Tree)</p><p>(B) 集成方法 (Ensemble Methods)</p><p>(C) K 平均法 (K-means)</p><p>(D) 支援向量機 (Support Vector Machine)</p></div>																												
C	<div><p>33. 關於 K 平均法 (K-means) 的分群，下列敘述何者「不」正確？</p></div>																												

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 8 頁，共 11 頁

	<p>(A) 一開始群的中心點是隨機選擇的</p> <p>(B) 每次分群結果必須讓組內平方和最小</p> <p>(C) 每次分群的結果都一模一樣</p> <p>(D) 一開始必須告知該演算法欲分群的群數</p>
D	<p>34. 參考附圖，下列敘述何者「不」正確？</p>  <p>(A) 此圖稱為散佈圖矩陣 (Scatter Plot Matrix)</p> <p>(B) Sepal.Length 與 Petal.Width 呈現正相關</p> <p>(C) Petal.Length 與 Petal.Width 呈現正相關</p> <p>(D) Sepal.Length 資料範圍為 2.0~4.0 之間</p>
B	<p>35. 考慮 R 語言之資料物件 x，如須找出大於三倍標準差的資料語法為下列何者？</p> <p>(A) $x[x > 3 * \text{var}(x)]$</p> <p>(B) $x[x > 3 * \text{sd}(x)]$</p> <p>(C) $x > 3 * \text{var}(x)$</p> <p>(D) $x > 3 * \text{sd}(x)$</p>
B	<p>36. 關於基於密度的聚類分析算法 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN)，下列敘述何者「不」正確？</p> <p>(A) DBSCAN 是一種基於密度的分群方法 (Density-Based Clustering)</p> <p>(B) 如果資料達到最小資料數目，則無法將該資料聚集成一群集</p> <p>(C) 如果資料點在所定義的半徑範圍內超過資料點密度，則稱為核心</p>

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 9 頁，共 11 頁

	<p>點 (Core)</p> <p>(D) 如果資料點位於核心點的半徑範圍內稱為境內點 (Border)</p>
C	<p>37. 關於群集分析 (Clustering Analysis)，下列敘述何者正確？</p> <p>(A) K 平均法 (K-Means) 不用事先決定群集數目</p> <p>(B) K 平均法 (K-means) 不用事先標準化資料即可建立較佳模型</p> <p>(C) 期望最大化法 (Expectation Maximization, EM) 是以模式為基礎的方法</p> <p>(D) 期望最大化法 (Expectation Maximization, EM) 不用事先決定群集數目</p>
B	<p>38. 下列何者常用來呈現資料的群聚情況？</p> <p>(A) 直方圖 (Histogram)</p> <p>(B) 熱圖 (Heat Map)</p> <p>(C) 折線圖 (Line Chart)</p> <p>(D) 趨勢圖 (Run Chart)</p>
D	<p>39. 若要描述非常態分佈的年收入，下列何者是最適當的指標？</p> <p>(A) 標準差 (Standard Deviation)</p> <p>(B) 平均值 (Mean)</p> <p>(C) 眾數 (Mode)</p> <p>(D) 中位數 (Median)</p>
B	<p>40. 關於機器學習，下列敘述何者正確？</p> <p>(A) 在沒有反應變數的監督學習情況下，我們無法知道監督式學習結果的真正答案</p> <p>(B) 非監督式學習通常更具挑戰性，其過程沒有單一的分析目標</p> <p>(C) 監督式學習通常是探索式資料分析的一部分</p> <p>(D) 非監督式學習的目標就是預測反應變數</p>
D	<p>41. 關於監督式學習 (Supervised Learning)，下列敘述何者「不」正確？</p> <p>(A) 可以由訓練資料中學到或建立一個模式 (Learning Model)</p> <p>(B) 訓練資料是由輸入和預期輸出所組成</p> <p>(C) 函數的輸出可以是一個連續的值 (也就是迴歸分析, Regression)，或是預測一個分類標籤 (也就是分類, Classification)</p> <p>(D) 多維尺度法 (Multidimensional Scaling) 屬於監督式學習</p>
D	<p>42. 關於決策樹 (Decision Tree)，下列敘述何者「不」正確？</p> <p>(A) 每個內部節點表示一個評估欄位</p> <p>(B) 每個分枝代表一個可能的欄位輸出結果</p> <p>(C) 每個樹葉節點代表不同分類的類別標記</p> <p>(D) 屬於非監督式演算法的一種</p>

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 10 頁，共 11 頁

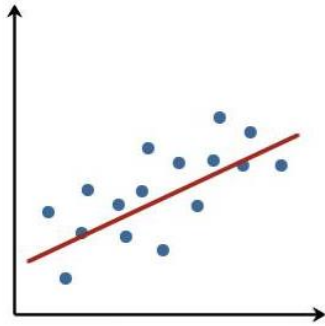
C	43. 關於線性迴歸，下列敘述何者正確？ (A) 迴歸方程式係數估計最佳化問題是最小化均方誤差 (Mean Squared Error, MSE) (B) 線性迴歸屬於無母數 (Non-parametric) 的統計建模方法 (C) 迴歸建模的好處是所獲得模型可解釋性高 (D) 任何資料集均可建立多元線性迴歸模型 (Multiple Linear Regression)，不會有建模失敗的狀況發生
C	44. 在相同的資料集下，請依模型複雜度從高到低排序線性建模的方法：壹、主成份迴歸 (principal component regression)；貳、多元線性迴歸；參、偏最小平方法 (partial least squares) (A) 貳 -> 參 -> 壹 (B) 參 -> 貳 -> 壹 (C) 貳 -> 壹 -> 參 (D) 壹 -> 貳 -> 參
B	45. 關於機器學習中的交叉驗證 (Cross validation)，下列何者「不」是其主要用途？ (A) 使用不同的資料組合來驗證訓練模型 (B) 讓兩人以上相互檢驗所搜集得到資料的正確性 (C) 避免過擬合 (over fitting) (D) 尋找模型適合的參數
C	46. 對於二元分類問題，依真實資料的真假值與模型預測輸出的真假值，可以組合出真陽性 (True Positive, TP)、真陰性 (True Negative, TN)、偽陽性 (False Positive, FP)、偽陰性 (False Negative, FN) 四種情況，組成混淆矩陣 (Confusion matrix)。其中對應於統計上的 Type I error 的是？ (A) TP (B) TN (C) FP (D) FN
A	47. 參考附圖，關於迴歸係數何者正確？

110 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：110 年 09 月 26 日

第 11 頁，共 11 頁

	 <p>(A) 迴歸係數>0 (B) 迴歸係數<0 (C) 迴歸係數$=0$ (D) 迴歸係數無法判斷</p>
D	<p>48. 關於簡單直線迴歸，下列敘述何者「不」正確？</p> <p>(A) 相關係數 (Coefficient of Correlation) 和斜率正負號必定相同 (B) 相關係數可用來表示線性關係強度 (C) 相關係數的平方等於判定係數 (Coefficient of Determination) (D) 兩變數間相關係數$=1$ 代表它們之間有因果關係</p>
C	<p>49. 下列關於線性迴歸模型，下列敘述何者正確？</p> <p>(A) 線性迴歸最多只能以 2 個自變數進行模型建立 (B) 線性迴歸方程式之繪圖呈現一定是直線 (C) 可透過最小平方法，優化線性迴歸模型 (D) 資料內的離群值 (Outlier)，不會對線性迴歸模型造成影響</p>
A	<p>50. 關於監督式學習 (Supervised Learning)，下列敘述何者正確？</p> <p>(A) 需要給定特徵輸入資料與預測輸出答案，以建立模型 (B) 僅能建立連續數值資料之預測模型，而無法建立類別之預測模型 (C) 監督式學習的特點是，能夠在沒有正確解答的情況下，對資料進行分群 (D) 監督式學習的特點是，不需要出現正確的輸入/輸出對，即可訓練出模型</p>