考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

C	1. 下列何項非資料前處理的步驟？ (A) 資料清理（Cleaning） (B) 資料操弄（Manipulation） (C) 資料建模（Modeling） (D) 資料變形（Reshaping）
A	2. 假設 Facebook 公司給您 1000 位用戶的基本資料，如：姓名、性別、年齡、學校、居住地，最可能是 R 語言中的何種資料結構？ (A) 資料框架（Data frame） (B) 串列（List） (C) 向量（Vector） (D) 矩陣（Matrix）
D	3. 使用下列何種方法，可以知道資料之中有偏差甚大的離群值存在？ (A) 將該欄位資料繪製成盒鬚圖（Box plot） (B) 將資料以直方圖（Histogram）表示 (C) 計算平均值與中位數的差異 (D) 以上皆是
D	4. 下列何者不是資料倉儲的特性？ (A) 主題導向的（Subject-oriented） (B) 經過整合的（Integrated） (C) 不會流失的（Non-volatile） (D) 屬於 OLTP 系統
D	5. 下列何者為資料遺缺的狀況？ (A) 完全隨機誤差（Missing Completely at Random，MCAR） (B) 隨機誤差（Missing at Random，MAR） (C) 非隨機誤差（Not Missing at Random，NMAR） (D) 以上皆是
C	6. 繪製下列何種圖表，資料集內至少需要包含兩個變量？ (A) 直方圖（Histogram） (B) 圓餅圖（Pie chart） (C) 散佈圖（Scatter plot） (D) 盒鬚圖（Box plot）
D	7. 下列何者不是用於資料的相關性分析（Correlation Analysis）？ (A) 卡方檢定 (B) 相關係數 (C) 共變異數 (D) 四分位數

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

C	8. 從 SQL Database 的角度，如果要輕易計算不同性別的平均體重，資料表欄位應該要怎樣設計比較恰當？ (A) 男性，女性，其他，男性體重，女性體重，其他體重 (B) 性別，男性體重，女性體重 (C) 性別，體重 (D) 以上皆非
C	9. 下列何種圖表適合用來展示時間序列（Time Series）類型的資料？ (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	10. 下列何者是利用時間序列來觀察不同維度之間隨時間變化的資訊？ (A) 勝率比（Odds ratio） (B) 平行座標圖（Parallel coordinates） (C) 目標投影追蹤（Targeted projection pursuit） (D) 運行圖（Run chart）
B	11. 有一群客戶的消費額最大為 3800 元、最小為 1800 元。假設將資料經過最小最大正規化（Min-Max Normalization）轉換成 0 到 1 的範圍區間，則若一客戶的消費額為 2300 元時，該消費額會被轉換為什麼數字？ (A) 0.2 (B) 0.25 (C) 0.4 (D) 0.5
A	12. 下列何者不是常用來儲存 log file 的資料格式？ (A) Doc (B) Csv (C) Textfile (D) Parquet
D	13. 下列何種方法可以用來進行特徵轉換？ (A) Diffusion maps (B) Locally-linear embedding (C) Relational perspective map (D) 以上皆是
D	14. 下列何者不是降維的好處？ (A) 減少運算時間與儲存空間 (B) 移除共線性資料能有效提高線性模型的效能 (C) 當資料維度降至 2～3 維時，能很容易的直接視覺化展示資料分佈 (D) 降維後的資料集訊息量增加，不會減少

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

D	15. 下列何項不是迴歸分析常用的維度縮減技術？ (A) 係數縮減法（Shrinkage） (B) 逐步迴歸法（Stepwise Regression） (C) 子集挑選法（Subset Selection） (D) 事後修剪法（Post-pruning）
A	16. 欲擷取網頁內容時，若發現網頁內容改變但網址不變時，較有可能為何請求方法？ (A) POST (B) PUT (C) GET (D) READ
D	17. 下列何者並非現今巨量資料系統架構的設計趨勢？ (A) 主從式分散架構（Master-Slave） (B) P2P 架構（P2P Architecture） (C) 分片機制（Sharding） (D) 高度集中化運算平台（Centalized Computing Platform）
B	18. 關於巨量資料平台 Hadoop，下列敘述何者正確？ (A) Name-Node 節點需要配置較多的記憶體，用來儲存文件資料 (B) 在 HDFS（Hadoop Distributed File System）上的文件，不支援隨機存取 (C) 支援一次寫入一次存取，確保資料完整存取 (D) 以上皆是
A	19. 下列何者不是 HDFS（Hadoop Distributed File System）的特色？ (A) 不需要 Master Node 來管理集群 (B) 可以將文件分散式儲存 (C) 適合儲存文字型資料 (D) 自動備份存入的檔案
A	20. 在撰寫 MapReduce 的程式時，下列何者操作不適合在 Reducer 中實現？ (A) x - y (B) x * y (C) x + y (D) count
D	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？ (A) 變異數 (B) 全距 (C) 平均數 (D) 變異係數

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

D	22. 盒鬚圖（Box plot）沒有顯示哪個統計量？ (A) 第一四分位數 (B) 中位數 (C) 第三四分位數 (D) 標準差
D	23. 下列何種情形適合使用單因子變異數分析（One-way Analysis of Variance）？ (A) 檢驗數據是否服從常態分配 (B) 比較某班級男生與女生數學成績的變異數 (C) 比較兩間輪胎工廠，輪胎平均使用年限是否不同 (D) 比較某工廠 4 部機器由不同人員操作下，其每小時平均產量是否不同
C	24. 二個獨立事件 A 與 B，機率分別是 60%與 40%，則 Pr{A∪B}=？ (A) 50% (B) 20% (C) 76% (D) 100%
B	25. 下列敘述何者正確？ (A) 若一組資料的最大值為 90，最小值為 0，其中位數為 60，則此資料為右偏 (B) 一組資料的所有數值與其算術平均數的差，其總和為 0 (C) 若二組資料有相同標準差，且平均數皆為正數，則平均數愈大者，變異係數愈大 (D) 兩組不同單位的資料可藉標準差來比較資料之離散程度
B	26. 若有四群學生的人數分別為 10、20、30、40 人，平均體重依序為 60、70、55、65 公斤，則全部學生的平均體重是？ (A) 60 公斤 (B) 62.5 公斤 (C) 65 公斤 (D) 67.5 公斤
C	27. 有一汽車業務員隨機拜訪 3 位客戶，依過去經驗客戶購買車的機率為 10%，試問這三位客戶中，至少有一位會購買車的機率？ (A) 23.1% (B) 25.1% (C) 27.1% (D) 29.1%

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

D	28. 統計資料分為離散型與連續型，請問下列何項與其他不同？ (A) 體重 (B) 身高 (C) 成績 (D) 國家數目
D	29. 關於連續型機率分配，下列敘述何者正確？ (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 (B) 已知均勻分配為 U(a, b)，則平均值為(a-b)/2 (C) 伽瑪分配是指數分配的特例 (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
C	30. 下列何者不是卡方檢定（Chi-square Test）的功能？ (A) 適合度檢定 (B) 獨立性檢定 (C) 變異數檢定 (D) 齊一性檢定
C	31. 下列何者為「非監督式學習」演算法？ (A) 決策樹（Decision tree） (B) 集成方法（Ensemble Methods） (C) K 平均法（K-Means） (D) 支援向量機（Support Vector Machine）
B	32. 關於非監督式學習，下列敘述何者正確？ (A) 意指不需要人看著就能學習 (B) 常見的集群分析屬於非監督式學習 (C) 常見的分類模型屬於非監督式學習 (D) 以上皆非
B	33. 關於 K 平均法（K-means）的分群，下列敘述何者不正確？ (A) 一開始群的中心點可以是隨機選擇的 (B) 每次分群的結果都一模一樣 (C) 每次分群結果必須讓組內平方和最小 (D) 一開始必須告知該演算法欲分群的群數
A	34. 下列何種分群演算法，是基於「密度」概念所設計的？ (A) OPTICS 演算法（Ordering Points To Identify the Clustering Structure） (B) K 平均法（K-means） (C) 聚合式階層分群法（Agglomerative Hierarchical Clustering） (D) 社群偵測（Community Detection）

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

A	35. 計算資料百分位數的 R 指令為何？ (A) quantile (B) percent (C) median (D) sum
D	36. 在 R 語言中使用 arules 套件，下列哪一個指令可將 dataset 轉換成關聯規則分析用資料？ (A) as(arules, "dataset") (B) as(dataset, "arules") (C) as(transactions, "dataset") (D) as(dataset, "transactions")
B	37. 欲呈現二維平面中檢視資料點之間的關係（例如：相似度或距離），一般會使用下列哪種方法？ (A) 詞頻-逆文檔頻率（Term Frequency Inverse Document Frequency，TF-IDF） (B) 多維尺度法（Multidimensional Scaling，MDS） (C) 最鄰近搜索（Approximate Nearest Neighbor，ANN） (D) 社會網路分析（Social Network Analysis，SNA）
C	38. 有一個數列[1,2,3,4,5,7,20]，若要找出此數列中的離群值，下列何者計算是不必要的？ (A) 計算此數列的平均數 (B) 計算此數列的標準差 (C) 計算此數列的峰度係數 (D) 將各數值標準化
B	39. 有一筆資料[1,2,5,6,10,22,...]，下列何種方式無法測量數列集中趨勢？ (A) 平均數 (B) 標準差 (C) 眾數 (D) 中位數
C	40. 下列何者不屬於非監督式學習？ (A) 局域離群因子（Local Outlier Factor） (B) 獨立成份分析（Independent Component Analysis） (C) 最近鄰法（Nearest Neighbor Methods） (D) 奇異值分解（Singular Value Decomposition）
D	41. 下列何者不是決策樹產生的基本演算法？ (A) ID3（Iterative Dichotomiser） (B) C4.5 (C) CART（Classification and Regression Trees） (D) 貝氏分類（Bayesian Classification）

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

C	42. 關於熱切式學習（Eager Learner）與偷懶式學習（Lazy Learner），下列敘述何者不正確？ (A) 熱切式學習是先利用訓練資料建立一個判別模型，以便進行測試 (B) 決策樹屬於熱切式學習 (C) 偷懶式學習會花很多時間在事先利用訓練資料建立判斷模型 (D) k-最近鄰分類法（K-Nearest-Neighbor Classifiers）屬於偷懶式學習
B	43. 若希望能透過學生基本資料與參與社團資料，來預測新生會選擇的社團，運用以下何種工具較為適當？ (A) 線性迴歸模型 (B) 分類模型 (C) 集群分析 (D) 探索式分析
D	44. 關於迴歸分析的基本統計假設，下列敘述何者正確？ (A) 依變數和自變數之間的關係必須是線性 (B) 資料呈現常態分配（Normal Distribution） (C) 自變數的誤差項，相互之間應該是獨立的 (D) 以上皆是
B	45. 如果判定係數為 0.8，則依變數能被自變數解釋的變異百分比為？ (A) 0.8% (B) 80% (C) 0.64% (D) 不一定
C	46. 假設在一混淆矩陣（Confusion Matrix）中，真陽性（True positive）為 100，假陽性（False Positive）為 50，真陰性（True Negative）為 50，假陰性（False Negative）為 800，請問該混淆矩陣的準確度（Accuracy）為？ (A) 0.6667 (B) 0.9412 (C) 0.15 (D) 0.84
D	47. 下列哪種方法可以避免機器學習模型過度配適（Overfitting）？ (A) 選擇特徵（Feature Selection） (B) 交叉驗證（Cross Validation） (C) 對目標函數施加懲罰（Penalty） (D) 以上皆是
C	48. 假設建立一個能夠辨識汽車的模型系統，在照片資料集共有 100 萬張照片，其中有 1000 張已標註汽車貼標的照片，接下來可用哪種學習方法找出剩下的照片當中是否有汽車？

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

	(A) 監督式學習（Supervised learning） (B) 非監督式學習（Unsupervised learning） (C) 半監督式學習（Semi-supervised learning） (D) 增強式學習（Reinforcement learning）
D	49. 下列何者情況不適合使用邏輯迴歸（Logistic Regression）模型？ (A) 明天是否下雨 (B) 鐵達尼號乘客是否存活 (C) 顧客是否會購買週年慶商品 (D) 行動通訊用戶國際電話服務用量預測
B	50. 當使用線性模型時，哪種方法對於學習預測線性不可分的資料集也許有幫助？ (A) 交叉驗證（Cross validation） (B) 核方法（Kernel method） (C) 過採樣（Over sampling） (D) 降採樣（Down sampling）
A	51. 將網頁資料擷取下來之後，應先進行下列何步驟？ (A) 資料清理（Cleaning） (B) 資料建模（Modeling） (C) 資料變形（Reshaping） (D) 趨勢預測（Prediction）
B	52. 假設 Facebook 公司給您 1000 位用戶的基本資料及文章資料，如：姓名、性別、年齡以及最近十篇發文的時間、點讚數、回應數與分享該文章所有人的基本資料，最適合 R 語言中的何種資料結構？ (A) 資料框架（Data frame） (B) 串列（List） (C) 向量（Vector） (D) 矩陣（Matrix）
A	53. 若資料表中只出現了一個遺缺值（NA）值，下列何項處理方式最不適當？ (A) 刪除整欄（變數） (B) 刪除整列（觀測值） (C) 以該欄其餘的資料平均值取代 NA 值 (D) 往回追溯資料源頭，尋找 NA 的來源
D	54. 下列何者不是資料倉儲的資料類型？ (A) 運算資料 (B) 預先加總資料 (C) 中繼資料（Metadata）

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

	(D) 即時更新資料
B	55. 一般來說，下列何者不是資料清理的目的？ (A) 將資料轉為可以分析的格式 (B) 發現資料之間的相關性 (C) 處理遺缺值 (D) 讓計算及分析上，更為方便及降低偏誤
B	56. 在一次考試中，由於班上同學考試成績最高分僅有 70 分，為了能夠讓學期成績比較好看，老師決定幫每個人的考試成績都加 10 分，請問這個數值樣本中的哪個統計量不會因為調分而有差別？ (A) 平均值 (B) 標準差 (C) 中位數 (D) 第一四分位數
A	57. 在統計學中，下列哪一個選項的分佈類型與其他不相同？ (A) 二項分佈（Binomial Distribution） (B) 指數分佈（Exponential Distribution） (C) t 分佈（t Distribution） (D) 常態分佈（Normal Distribution）
C	58. 關於資料探索，下列敘述何者不正確？ (A) 透過工具函數（例如：R 語言當中的 summary 函數）可了解關於資料集內容的整體結構、變數情況、分佈指標、遺缺值 (B) 視覺化工具可幫忙了解變數間的關係，以利後續資料探勘作業 (C) 定性變數可計算出最小值、分位數、中位數、平均值與最大值進行觀察 (D) 透過平均值和中位數的差異程度來判斷資料的偏倚程度，可用來判斷資料之左偏或右偏情況
A	59. 巨量資料中，以資料類別出現頻率排列下出現的長尾現象，一般可利用哪種統計工具來描述資料分佈？ (A) Zipf（齊夫分佈） (B) Gaussian（高斯分佈） (C) Dirichlet（狄利克雷分佈） (D) Uniform（均勻分佈）
D	60. 如果整理不同品項與業績的報表，最適合使用下列何種圖表？ (A) 盒鬚圖（Box plot） (B) 直方圖（Histogram） (C) 分位數圖（QQ plot） (D) 長條圖（Bar chart）

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

A	61. 若一個樣本的平均數為 100，標準差為 10，擇一個數值為 200，則該數值標準化（Standardarization）後，數值將會轉變為？ (A) 10 (B) 20 (C) 100 (D) 200
A	62. 使用下列何種方法，可以重新組合資料屬性，產生新的維度？ (A) 主成分分析法（PCA，Principle Component Analysis） (B) K 平均法（K-means） (C) C50 (D) 卡方檢定（Chi-square test）
B	63. 下列何種方法可以把學生的成績從連續型數值轉變為離散型的級距？ (A) 最大正規化（Min-Max Normalization） (B) 裝箱法（Binning Method） (C) 數值標準化（Standardarization） (D) Z-分數正規化（Z-score Normalization）
D	64. 下列何者不是特徵萃取所要達到的目的？ (A) 降低資料維度 (B) 提高學習模型時的效率與效能 (C) 過濾無用資訊 (D) 評估學習得到的模型效能
D	65. 下列何者不是常見的資料維度降維方法？ (A) 主成分分析（Principle Component Analysis） (B) 核主成分分析（Kernel PCA） (C) 多維尺度法（Multidimensional Scaling） (D) K 平均法（K-means）
C	66. 欲擷取網頁內容時，若發現網頁內容隨著網址而規律的改變，較有可能為何請求方法？ (A) POST (B) PUT (C) GET (D) READ
B	67. 下列何者並非串流計算（Streaming）的特性？ (A) 可擴展性（Scalable） (B) 批次運算（Batch Processing） (C) 低延遲（Low-Latency）

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

	(D) 高度容錯性（Fault-Tolerance）
B	68. 關於巨量資料運算平台上的資料處理，下列敘述何者不正確？ (A) 以分散式方式將資料分片儲存於各節點，以利一次寫入、多次讀取 (B) 資料本身可進行壓縮（gzip），分散到各資料節點，以增加運算效能 (C) 透過網路傳輸將執行程式送到資料端進行運算 (D) Map 程式的輸出的結果是中間檔（IFILE），Reducer 程式輸出的結果是在 HDFS（Hadoop Distributed File System）的檔案
B	69. 下列何者對於 HDFS（Hadoop Distributed File System）的使用是不恰當的？ (A) 存入過大（>100GB）的文字檔案 (B) 將檔案分隔成小單位（<4MB）存入 (C) 存入串流（streaming）資料 (D) 將文字檔壓縮存入
D	70. 在 MapReduce 架構中，資料由輸入到輸出的處理順序，下列何者正確？ (A) Map＞Reduce＞Sort＞Merge (B) Reduce＞Sort＞Map＞Merge (C) Sort＞Map＞Reduce＞Merge (D) Map＞Sort＞Merge＞Reduce
A	71. 下列何者為互斥事件（Mutually Exclusive Event）？ (A) 某公司 58.3%為男性，41.7%為女性 (B) 顧客在購買產品時，67.9%會考慮品質，34.1%會考慮價格 (C) 有 44.5%的顧客會選擇 X 產品，32.9%的顧客會選擇 Y 產品，29.5%的顧客會選擇 Z 產品 (D) 投擲一枚骰子骰到 1、2、3 的機率和骰到偶數的機率
C	72. 已知 4 組樣本資料：(2,5), (1,3), (5,6), (0,2)，試計算樣本相關係數 r？ (A) r=0.72 (B) r=0.83 (C) r=0.93 (D) r=1.0
C	73. 某工廠有 4 部機器生產同一產品，各機器生產之產品數量各佔總產量之比例為 0.4, 0.3, 0.2, 0.1。各機器產品的不良率分別為 0.02, 0.05, 0.01, 0.02，試問若隨機抽取一產品，其為不良品的機率為？ (A) 0.008 (B) 0.02

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

	(C) 0.027 (D) 0.05
A	74. 下列哪個統計圖表可以用來判定資料之離群值（Outlier）？ (A) 盒鬚圖（Box plot） (B) 圓餅圖（Pie chart） (C) 直方圖（Histogram） (D) 長條圖（Bar chart）
A	75. 一組數據資料中，若平均數減去中位數的值是很大的正數時，則下列敘述何者正確？ (A) 資料分佈呈右偏 (B) 中位數必須小於零，同時平均數必須大於零 (C) 平均數必須是大的正數 (D) 中位數必須小於零
B	76. 某疾病的發生率為 5%。某藥廠發展出一種檢測藥劑，若「有病」則檢測結果為「陽性（有病）」的機率為 99%，若「無病」則檢測結果為「陰性（沒病）」的機率亦為 99%。現隨機選一人，則檢測結果為陰性的機率，最接近下列何者？ (A) 95% (B) 94% (C) 93% (D) 92%
A	77. 一串聯系統有兩個獨立運作之零件 A1 與 A2，其故障機率分別為 40%與 50%，下列敘述何者正確？ (A) 整體系統能運作機率為 30% (B) 整體系統能運作機率為 20% (C) 若改成並聯，整體系統運作機率為 90% (D) 若改成並聯，整體系統運作機率為 70%
A	78. 某鄉鎮人口男性佔 60%，女性佔 40%，男性中有 30%有買基金，女性中有 10%有買基金。今從此鄉鎮隨機選出一人，若已確定此人有買基金，則此人為男性的機率為何？ (A) 81.82% (B) 83.82% (C) 18.18% (D) 20.18%
A	79. 何種測量離散程度的測度量最易受到極端值的影響？ (A) 全距 (B) 變異數

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

	(C) 四分位距 (D) 標準差
C	80. 下列敘述何者正確？ (A) 樣本變異數的值一般介於 1 與-1 之間 (B) 母體變異數計算公式與樣本變異數相同 (C) 變異係數(Coefficient of Variance)=(標準差/平均數)*100% (D) 相關係數介於 0 與 1 之間
D	81. 關於 K 平均法（K-means），下列敘述何者不正確？ (A) 希望找出 k 個互不交集的群集 (B) 不同的起始群集中心，可能會造成不同的分群結果 (C) 容易受雜訊與離群值影響其群集中心 (D) 可以處理類別型資料
A	82. 探索式資料分析的主要目的為何？ (A) 熟悉資料 (B) 視覺化資料 (C) 測試模型 (D) 資料分群
C	83. 下列哪一項技術屬於非監督式學習？ (A) 決策樹（Decision Tree） (B) 類神經網路（Neural Network） (C) 集群分析（Clustering Analysis） (D) 支援向量機（Support Vector Machine）
A	84. 關於探索式資料繪圖，下列敘述何者不正確？ (A) 直方圖之 X 軸資料是間斷不連續的 (B) 長條圖適合用於類別型資料分析 (C) QQ plot 可用於常態分佈視覺化檢驗 (D) ROC 曲線（Receiver Operating Characteristic Curve）用於分類模型評估
D	85. 關於集群分析（Clustering Analysis），下列敘述何者不正確？ (A) 依照相似度將資料分群 (B) 同一群內的相似度大 (C) 各群之間的相似度小 (D) K-means 每次分群結果一定會相同
C	86. 關於階層式集群分析（Hierarchical Clustering），下列敘述何者不正確？ (A) 一般採用樹狀圖（Dendrogram）表示 (B) 樹狀圖根節點（Root）為單一群集

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

	(C) 聚合法（Agglomerative）是由上方根節點往下進行計算 (D) 分裂法（Divisive）是一開始將所有資料視為一個大群集
C	87. 關於階層式集群分析（Hierarchical Clustering）的方法，下列敘述何者不正確？ (A) 單一連結法（Single Linkage Method）採用兩群間最小距離 (B) 完全連結法（Complete Linkage Method）採用兩群間最大距離 (C) 平均連結法（Average Linkage Method）採用兩群間中心點距離 (D) 華德法（Ward's Method）是計算組內變異作為評估群集相似性
B	88. 推薦系統（Recommender System）通常採用下列哪一個方法作為核心技術，來分析產品與使用者間的關係？ (A) 支援向量機（Support Vector Machine） (B) 矩陣分解（Matrix Factorization） (C) 線性判別分析（Linear Discriminative Analysis） (D) 詞性標記（Part-of-Speech（POS）Tagging）
A	89. 下列何種統計學習的演算法是用來進行資料的分群（Clustering），但不能用來進行資料分類（Classification）？ (A) 基於密度的集群分析算法（Density-Based Spatial Clustering of Applications with Noise，DBSCAN） (B) 貝氏網路（Bayisian Network） (C) 隨機森林（Random Forest） (D) 支援向量機（Support Vector Machine）
D	90. 在非監督式學習方法中，下列何者最常被做為資料降維的方法使用？ (A) K 平均法（K-means） (B) 最大期望算法（Expectation-maximization） (C) 模糊 C 平均法（Fuzzy C-means） (D) 主成分分析（Principle Component Analysis）
B	91. 試問下列哪一項不包含在一個「多層前向式（Multilayer Feed-Forward）」類神經網路架構？ (A) 輸入層（Input Layer） (B) 實體層（Physical Layer） (C) 隱藏層（Hidden Layer） (D) 輸出層（Output Layer）
B	92. 不同的決策樹方法，我們可以透過屬性選擇指標（Attribute Selection Measure），將資料分割成個別類別，使其所包含的資料群組具有相同的類別，試問下列何者不是屬性選擇指標？ (A) 資訊獲利（Information Gain） (B) 拉普拉斯估計式（Laplace Estimator）

(C) 獲利比率（Gain Ratio）
(D) 吉尼係數（Gini Index）

B 93. 根據下表，我們要預測該公司的顧客是否會買電腦，若規則 R:(年齡=青年)Λ (學生=是)=>(購買電腦=是)，則規則 R 的覆蓋率與正確率分別為何？

某公司顧客資料庫的訓練資料

No.	年齡	收入	學生	購買電腦
1	青年	高	否	否
2	青年	高	否	否
3	中年	高	否	是
4	老年	中	否	是
5	老年	低	是	是
6	老年	低	是	否
7	中年	低	是	是
8	青年	中	否	否
9	青年	低	是	是
10	老年	中	是	是
11	青年	中	是	是
12	中年	中	否	是
13	中年	高	是	是
14	老年	中	否	否

(A) 覆蓋率為 14.28% 與正確率為 92.86%
(B) 覆蓋率為 14.28% 與正確率為 100%
(C) 覆蓋率為 28.57% 與正確率為 92.86%
(D) 覆蓋率為 28.57% 與正確率為 100%

A 94. 若希望能透過歷史氣溫與菜價資料來預測未來菜價，運用以下何種工具較為適當？
(A) 線性迴歸模型
(B) 分類模型
(C) 集群分析
(D) 探索式分析

C 95. 關於過度配適（Overfitting），下列敘述何者不正確？
(A) 知識發掘的方法在建立模型的過程中容易出現過度配適的情形，模型可能陷入只能解釋在訓練集樣本的關聯，而沒辦法一體適用
(B) 機器學習所學到的假設（Hypothesis）過度貼近訓練資料（Training Data），而導致測試資料（Testing Data）錯誤率變得更大
(C) 過度配適表示測試資料的正確率極高
(D) 為了避免過度配適現象，必須使用額外的技巧，如交叉驗證、Early Stopping、貝斯信息量準則（BIC）、赤池信息量準則（AIC）等

考科 2：資料處理與分析概論-參考樣題

提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

B	96. 關於 Apriori 演算法，下列敘述何者不正確？ (A) 使用於關聯規則分析 (B) 屬於協同過濾推薦的一環 (C) 需要產生大量候選項集和需要重複掃描資料庫 (D) 支持度大於最小支持度的項集稱為頻繁項目集
C	97. 假設在一混淆矩陣（Confusion Matrix）中，真陽性（True Positive）為 100，假陽性（False Positive）為 50，真陰性（True Negative）為 50，假陰性（False Negative）為 800，請問該混淆矩陣的真陽性率（True Positive Rate）為？ (A) 0.15 (B) 0.9 (C) 0.6667 (D) 0.1
C	98. 下列何種機器學習方法不屬於監督式學習演算法？ (A) 邏輯迴歸（Logistic Regression） (B) 類神經網路（Neural Network） (C) 多維尺度法（Multidimensional Scaling） (D) 最近鄰居法（Nearest Neighbor）
D	99. 假設行動廣告推薦系統，透過每次廣告推薦的決策中，得到用戶點擊與否的資訊，並不斷進行改進，修正其策略以得到最佳的廣告推薦效果，適合下列何種學習方法？ (A) 監督式學習（Supervised Learning） (B) 非監督式學習（Unsupervised Learning） (C) 半監督式學習（Semi-supervised Learning） (D) 增強式學習（Reinforcement Learning）
D	100. 關於隨機森林（Random Forest），下列敘述何者不正確？ (A) 包含多個決策樹 (B) 可以幫助篩選重要的自變項 (C) 輸出可以採用投票或是平均 (D) 依變項只能使用類別型變項