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

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) 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）

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) 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
A	101. 下列何項工作通常較耗費時間？ (A) 資料前處理（清理、操弄及變形等） (B) 統計分析 (C) 資料視覺化 (D) 資料建模
D	102. 下列何者並非資料清理的範疇？ (A) 檢查資料的不一致性

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！

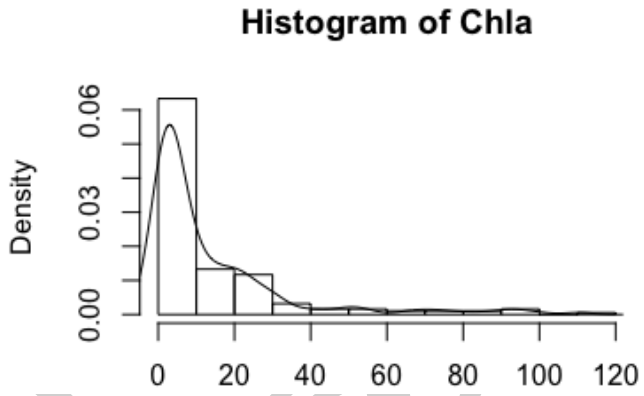
	(B) 移除遺缺值（Missing Value） (C) 去除重複資料 (D) 資料合併
D	103. 下列何者不是進行資料操弄（Manipulation）與清理時，較常會用到的方法？ (A) 資料萃取（Extract） (B) 資料轉換（Transform） (C) 資料載入（Load） (D) 資料視覺化（Visualization）
C	104. 下列何者並非資料清理時，所需處理之項目？ (A) 處理資料遺缺值（Missing Value） (B) 處理資料離群值（Outlier） (C) 處理資料變異性（Variability） (D) 處理資料之錯字
D	105. 在進行資料處理時，會使用正規表達式（Regular Expression）來處理何種類型的資料？ (A) 視頻 (B) 音頻 (C) 圖像 (D) 文字
B	106. 在一個網站使用者活動的日報表，某天造訪該網站的使用者中，來自台灣的使用者人數為 3,200 人，來自美國的使用者人數為 1,000 人，來自日本的使用者人數為 500 人，來自其他國家的使用者人數為 300 人。請問台灣使用者的人數占總比多少百分比？ (A) 50 (B) 64 (C) 80 (D) 32
C	107. 我們常用盒鬚圖（Box plot）觀察資料分散情況資料，下列敘述何者不正確？ (A) 四分位距（Interquartile Range）可用來計算隔離離群值（Outlier）的邊界 (B) 資料分配有對稱分配、左偏分配、右偏分配與均勻分配 (C) 需要 5 個統計資料，包含最大值、最小值、眾數、下四分位數及上四分位數 (D) 將某些集中量數與分散量數，以長盒形圖表現出來的一種圖示法
D	108. 下列何者不是探索式資料分析（Exploratory Data Analysis）常用的圖形

	化方法？ (A) 長條圖（Bar Chart） (B) 圓餅圖（Pie Chart） (C) 次數分配圖（Frequency Chart） (D) 流程圖（Flow Chart）
D	109. 當使用直方圖（Histogram）進行資料探索時，下列何者無法從圖形中看出？ (A) 資料中的異常值 (B) 資料的變異程度 (C) 資料的分布狀況 (D) 資料的排列順序
B	110. 關於探索式資料分析（Exploratory Data Analysis），下列敘述何者不正確？ (A) 不拘泥於統計方法 (B) 著重於對資料的整理 (C) 可用於驗證算法的假設 (D) 更直觀的發現隱含的資訊
A	111. 如果要使用邏輯式迴歸（Logistic Regression）建立客戶流失分類模型，則我們必須將標籤（例如名為：流失，不流失）做什麼處理？ (A) 轉換為二元（Binary）數值 (B) 必須將標籤從中文名稱轉換成英文 Yes 與 No (C) 保留原始標籤名稱即可 (D) 必須將標籤轉換為數字型態
B	112. 下列何者不是讀取文字檔時，通常需要注意的事項？ (A) 檔案編碼格式 (B) 檔案擁有者 (C) 檔案大小 (D) 欄位分隔符號
D	113. 下列何者並非屬性選取（Feature Selection）的目的？ (A) 避免過度配適（Overfitting） (B) 提高機器學習效能 (C) 篩選最佳屬性 (D) 提高資料維度
A	114. 關於資料標準化（Normalization），下列敘述何者不正確？ (A) 通常會在建模後進行 (B) 消除不同觀測值間的差異性 (C) 減少不必要的變異

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) 將不同變量縮放到一定範圍
B	115. 若您想對 30000 筆的房屋資料進行機器學習，欄位包含房價、樓層、距市區距離、車位數量和面積等，應該先對資料進行下列何項步驟，使得變量尺度統一？ (A) 集中化（Centralization） (B) 標準化（Normalization） (C) 聚合（Aggregation） (D) 離散化（Discretization）
D	116. 下列何者不屬於 HTTP/1.1 協定中定義的方法？ (A) POST (B) GET (C) DELETE (D) TCP
C	117. 在大數據系統中，為了能夠有效的分散式運算資料，資料通常會被擺置在下列哪個地方？ (A) Storage Area Network（SAN）儲存設備 (B) 集中化的儲存設備 (C) 分散擺放在各個機器節點上 (D) 一台主要節點之中
A	118. 關於巨量資料運算平台架構之設計策略，不包含下列何者？ (A) 垂直式擴充（Scale-up） (B) 水平式擴充（Scale-out） (C) 大量平行化運算（Massively Parallel Processing） (D) 無共享架構（Shared-Nothing Architecture）
A	119. 下列何者並非提升巨量資料處理效能的方法？ (A) 採用關聯式資料庫 (B) 使用分散式架構處理數據 (C) 提升為較好的硬體配備 (D) 進行平行運算
C	120. 若欲透過巨量資料分析來剖析核心的商業價值，下列何者需最先被執行？ (A) 數據建模 (B) 資料解析 (C) 定義問題 (D) 資料理解
B	121. 關於卡方分配、F 分配、t 分配之比較，下列敘述何者不正確？ (A) 皆為小樣本

	(B) 皆為不連續分配 (C) 三者皆有自由度 (D) 皆來自於常態母體
C	122. 關於卜瓦松分配之特性，下列敘述何者不正確？ (A) 某一時段內發生的次數與其他時段發生的次數相互獨立 (B) λ 與所選擇之時間長度成比例 (C) 在極短時間內成功兩次以上之機率不可忽略不計 (D) 在相同長度的時段內發生事件的機率皆相同
D	123. 當所有觀察值都落在迴歸直線上，則 x 與 y 之間的相關係數為何？ (A) -1 < r < 1 (B) 僅 r = 1 (C) 僅 r = -1 (D) r = 1 或 r = -1
B	124. 下圖為海藻資料集的變數 Chla 分佈狀況，請問最接近何種機率分佈？  (A) 常態分佈 (B) 偏態分佈 (C) 幾何分佈 (D) 均勻分佈
A	125. 考慮簡單線性迴歸（Linear Regression）模型，其變異數分析表中，迴歸模型的自由度為何？ (A) 1 (B) 2 (C) 12 (D) 20
D	126. 機器學習（Machine Learning）是從所搜集的資料中建構出（學習出 learning 或配適出 fitting）X 與 Y 之間的函數關係 Y = f(X)，下列敘述何者不正確？ (A) X 稱為預測變數（Predictors） (B) X 稱為獨立變數（Independent Variables）

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) Y 稱為反應變數（Responses） (D) Y 稱為屬性（Features）
C	127. 關於模型績效評估，下列敘述何者不正確？ (A) 殘差（或稱預測誤差）是真實的反應變數值減去預測的反應變數值 (B) 所有的模型績效衡量準則都基於殘差 (C) 均方根預測誤差（Root Mean Squared Error, RMSE）其單位為反應變數原始單位的平方 (D) 瞭解殘差的分佈情形有助於評估模型的優劣與適用情境
B	128. 抽樣（Sampling）是統計學重要的概念之一，下列敘述何者正確？ (A) 巨量資料趨近於母體，因此分析時一定要採用抽樣方法 (B) 資料分析師如欲估計線性迴歸模型的變異性，可以用重抽樣方法（Resampling Methods）對每一組重抽樣訓練樣本配適模型後，檢視各模型績效的差異程度，這種作法使我們可以獲得只以原訓練集配適一次因而無法獲得的有用資訊 (C) 隨機抽樣（Random Sampling）是反覆地從訓練集或資料集中抽出或有不同的各組樣本，並重新配適各組樣本的模型，以獲得模型相關的額外資訊 (D) 拔靴取樣法（Bootrapping）是隨機 k 等分（通常是十等份）訓練集樣本後，每次留下一份作為測試集樣本，而以其餘 k-1 份樣本進行模型訓練
C	129. 關於模型訓練與測試機制，下列敘述何者正確？ (A) 用測試集對最佳模型未來之績效估計工作，又可稱為（候選）模型挑選（Model Selection）的階段 (B) 模型建立與優化的步驟，又可稱為模型評定/績效估計（Model Assessment/Performance Estimate）階段 (C) 模型建立後，常透過交叉驗證（Cross-Validation）進行優化，避免過度配適（Overfitting）的情況發生 (D) 最佳模型的績效估計工作來說，績效準則的計算速度是首要考量
B	130. 關於機率與統計，下列敘述何者不正確？ (A) 統計分析的對象是樣本而非母體時，可引入能合宜刻畫抽樣變異的機率模型，以了解源自於抽樣變異，所產生的統計估計值不確定程度 (B) 大資料時代下各式資料充斥，資料分析師必須使用全部的可用資料進行分析 (C) 抽樣理論說明如何有效率地從母體中萃取所需要的訊息，許多統計推論均假設樣本已經隨機抽出，可以直接進行分析

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) 資料探勘與機器學習領域中必須運用基本的統計分析，以及進階的重抽樣（Resampling）與模擬（Simulation）等方法。因此，統計學仍然是大資料時代下必備且基本的武器之一
D	131. 關於「非監督式學習」（Unsupervised Learning），下列敘述何者正確？ (A) 輸出屬性值是可以被預測的 (B) 需用測試集測試所建立模型的正確性 (C) 輸入資料的形式只能是類別資料型態 (D) 建立模式用的資料，並不是事前定義好的
B	132. 關於離群值（Outlier），下列敘述何者不正確？ (A) 與一般資料極度不同的資料個體 (B) 異常值偵測時可以忽略之 (C) 經由量測錯誤所引起 (D) 資料與生俱有的變異性造成
C	133. 關於 K-means 演算法，下列敘述何者不正確？ (A) 將 n 筆待分群的資料選出 k 個資料點為集群的中心點 (B) 將所有資料與此 k 個中心點做距離運算 (C) 穩定性高，對異常值或極值不敏感 (D) 每次計算 K-means，分群結果不一定相同
D	134. 關於階層式分群，下列敘述何者不正確？ (A) 階層式分群法可以用樹狀結構呈現計算過程 (B) 使用階層式分群法時，必須定義資料群之間的距離計算方式 (C) 階層式分群可應用於小資料集 (D) 階層式分群法需一開始給定群的數目
A	135. 關於探索式資料分析（Exploratory Data Analysis），下列何者常用來呈現非類別資料的情況？ (A) 莖葉圖（Stem-and-Leaf Plot） (B) 長條圖（Bar Chart） (C) 桑基圖（Sankey Diagram） (D) 圓餅圖（Pie Chart）
B	136. 關於多變量探索式分析（Multivariate Exploratory Analysis），下列敘述何者不正確？ (A) 多個量化變數（Quantitative Variables）之間的關係，通常可以散佈圖矩陣（Scatterplot Matrix）來表達 (B) 共變異數（Covariance）是相對指標（Relative Index），而相關係數（Correlation Coefficient）是絕對指標（Absolute Index） (C) 計算數值資料矩陣中所有成對的相關係數，即可獲得相關係數矩陣（Correlation Matrix）

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) 透過數據排名值（Rank）的方式，將共變異數與相關係數的概念延伸到類別變數上
C	137. 屬性挑選（Feature Selection）是指挑選原始資料中的合宜屬性，或可視為移除缺乏訊息內涵之變數的維度縮減策略。下列常用的降維方法中，何者不屬於屬性挑選的方式？ (A) 遺缺值比率（Missing Values Ratio） (B) 前向式屬性構模（Forward Feature Construction） (C) 主成分分析（Principal Component Analysis） (D) 卡方檢定與信息增益（Chi-square and Information Gain）
B	138. 關於機器學習的方式，下列敘述何者正確？ (A) 非監督式學習的目標通常清晰 (B) 一般而言我們很難評估非監督式學習結果的好壞 (C) 監督式學習通常是探索式資料分析的一部分 (D) 非監督式學習的目標就是預測反應變數
A	139. 下列何者不是非監督式學習的任務？ (A) 上傳至社群網路的相片是否有人臉 (B) 在癌症樣本中，或是基因變數中尋找可能的子群，以對此疾病有更好的瞭解 (C) 線上購物網站嘗試依相似的瀏覽習慣與採購記錄將消費者分群，並將同群消費者購買的品項集合在一起 (D) 搜尋引擎也會依據具有相同搜尋型態的使用者點擊歷程，決定呈現哪些搜尋結果
D	140. 關於關聯規則分析（Association Rule），下列敘述何者不正確？ (A) 典型的關聯規則分析，是分析超市中顧客購買的品項集合資料（通常被稱為交易資料，或是購物籃資料），其個別品項間或品項群間的關聯 (B) 關聯規則分析最常見的模型，是以品項集合出現的次數，來量化彼此間的關聯程度，以此探勘出來的品項集合稱為頻繁品項集（Large Itemsets or Frequent Itemsets） (C) 關聯規則分析的目的，是基於某些品項出現的前提下，挖掘出可預測其它品項發生之可能性的規則，這些規則就被稱為關聯規則 (D) 關聯規則分析在資料探勘領域中，不容易與其它方法論結合運用，例如：集群、分類與離群值分析等
A	141. 下列何種方法，最適合解決非線性分類問題及高維空間數據識別問題？ (A) 支援向量機（Support Vector Machine） (B) K 平均法（K-means）

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！

<table>
<tr><td></td><td>(C) 貝式判別（Bayesian Classifier）
(D) 模糊理論（Fuzzy Set）</td></tr>
<tr><td>B</td><td>142. 下列何者為「常用來分析客戶資料，並建構模型以預測顧客流失（Churn）與否」的方法？
(A) 關聯規則分析
(B) 分類模型
(C) 迴歸分析
(D) 集群分析</td></tr>
<tr><td>A</td><td>143. 關於訓練資料集與測試資料集，下列敘述何者正確？
(A) 訓練資料是從要分析的資料庫中隨機取樣
(B) 訓練資料可以不知道其類別
(C) 測試資料可以包含訓練資料集中的資料
(D) 測試資料可以不知道其類別</td></tr>
<tr><td>C</td><td>144. 關於線性迴歸模型（Linear Regression）的假設，下列敘述何者不正確？
(A) 依變項（要預測的變項）呈現常態分佈
(B) 自變項（預測變項）之間是獨立事件
(C) 自變項的組合來自隨機抽樣（例如五個自變項隨機抽三個出來做迴歸）
(D) 依變項之誤差為常態分佈</td></tr>
<tr><td>A</td><td>145. 下列何者為「學習模型時，為了得到好的準確度」而容易造成的問題？
(A) 過度擬合（Overfitting）
(B) 資料標籤不均（Imbalanced Data）
(C) 資料維度過高
(D) 資料集過大</td></tr>
<tr><td>B</td><td>146. 在線性迴歸中，以最小平方法計算迴歸參數時，殘差需符合四大條件。下列何者非屬四大條件之一？
(A) 殘差期望值為零
(B) 殘差必須符合均勻分配
(C) 殘差之間沒有自相關（殘差獨立性）
(D) 殘差需符合變異數同質性</td></tr>
<tr><td>D</td><td>147. 選用線性模型或統計分析方法時，會根據依變數與自變數的屬性而有所不同，請問下列表格與配對的選項，何者不適當？
<table>
<tr><td></td><td>依變數 - 連續型</td><td>依變數 - 離散型</td></tr>
<tr><td>自變數 - 連續型</td><td>(A)線性迴歸</td><td>(B)線性判別分析（Linear Discriminant Analysis）</td></tr>
<tr><td>自變數 - 離散型</td><td>(C)變異數分析</td><td>(D)皮爾森相關係數（Pearson</td></tr>
</table>
</td></tr>
</table>

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！

		（Analysis of Variance）	Correlation Coefficient）

(A) 線性迴歸
(B) 線性判別分析（Linear Discriminant Analysis）
(C) 變異數分析（Analysis of Variance）
(D) 皮爾森相關係數（Pearson Correlation Coefficient）

D

148. 根據資料分佈選擇模型假設時，會遇到不適合直接建立線性模型的情況，此時可以透過轉換資料的方式進行修正。請問透過何種模型，適合將圖 1 的資料分佈轉換為圖 2？

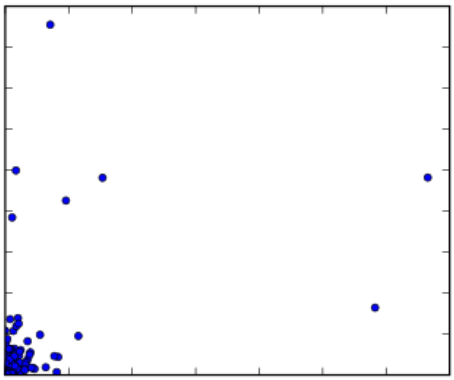


圖 1

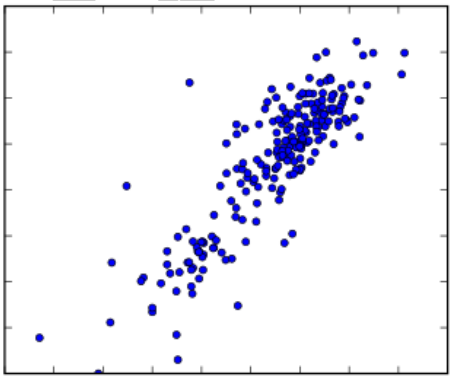


圖 2

(A) $y=\beta 0+\beta 1\ x$
(B) $y=\beta 0+\beta 1\ x+\beta 2\ x2$
(C) $y2=\beta 0+\beta 1\ x$
(D) $\log(y)= \beta 0+\beta 1\ \log(x)$

B

149. 關於線性迴歸模型績效評估，下列敘述何者不正確？
(A) 評估模型績效的方式不只一種
(B) 通常可透過混淆矩陣評估模型績效
(C) 殘差繪圖（Residual Plots）是以視覺化的方式檢視模型的配適狀況
(D) 許多績效評量的計算是基於殘差（Residual），它是各觀測值減去其模型的預測值，而常用的 SSE 是殘差平方的總和

A

150. 關於資料解析思維，下列敘述何者不正確？
(A) 資料建模時，規範性（Prescriptive）模型、敘述性（Descriptive）模型與預測性（Predictive）模型必須謹慎擇一使用
(B) 許多資料建模方法成功的關鍵是決定合宜的接近性衡量，以同中求異、異中求同的思維解決問題

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) 資料前處理就是避免垃圾進垃圾出（Garbage In Garbage Out, GIGO） (D) 不確定型態建模時須以合宜的理論（Theoretical）機率分佈，捕捉實際資料之經驗（Empirical）機率分佈變異的型態
D	151. 資料分析時常產生一些特殊值，請問下列何者並非 R 語言的特殊值？ (E) NULL (F) NA (G) Inf (H) Error
C	152. 在一份調查資料中，下列何者遺缺值（Missing Value）較適合使用平均值（Mean）填充？（假設遺缺值的比例小於萬分之一，且為隨機遺缺） (E) 遺缺的家庭收入 (F) 遺缺的居住地區 (G) 遺缺的身高 (H) 遺缺的家庭支出
B	153. 在資料建模前經常會對髒資料（Dirty Data）進行預處理，下列何者不屬於髒資料？ (E) 資料傳輸時遺漏的數據 (F) 資料接收時間延遲 (G) 資料來源不一致 (H) 資料包含雜訊
A	154. 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) 串列（List） (D) 資料框架（Data frame）
B	155. 網路上提供了許多不同型態的資料，在處理時通常需要將其轉化為何種類型，以便進行資料探勘（Data Mining）？ (E) 中繼資料（Metadata） (F) 結構化資料（Structured Data） (G) 非結構化資料（Unstructured Data） (H) 半結構化資料（Semi-structured Data）
C	156. 下列何種圖表較適合呈現國內各手機品牌的市占率？ (E) 散佈圖（Scatter plot）

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！

	(F) 直方圖（Histogram） (G) 圓餅圖（Pie Chart） (H) 肩形圖（Ogive）
A	157. 下列何者在繪製時需要使用到資料的四分位數？ (A) 盒鬚圖（Box plot） (B) 目標投影追蹤（Targeted projection pursuit） (C) 散點圖（Scatter plot） (D) 平行座標圖（Parallel coordinates）
A	158. 若要較直觀的觀察各字詞的出現次數，最適合使用下列何種圖表？ (A) 克里夫蘭點圖（Cleveland Dot Plot） (B) 散點圖（Scatter Plot） (C) 文字雲（Word Cloud） (D) 氣泡圖（Bubble Chart）
A	159. 使用視覺化進行資料探索時，下列用法何者不正確？ (A) 使用點圖（Dot Plot）觀察房價跟建物坪數之間的關係 (B) 透過折線圖（Line Chart）觀察房價的走勢 (C) 以長條圖（Bar Chart）檢視不同縣市的房價 (D) 利用圓餅圖（Pie Chart）來辨別不同建物類型的比例
D	160. 若對類別變量進行資料探索，下列敘述何者正確？ (E) 得知資料的變異程度 (F) 得知資料的分佈情形 (G) 得知資料間的相關性 (H) 得知資料的出現頻次
A	161. 請問經過下列轉換方法做資料轉換後，哪個方法會產生出較原資料多的變數？ (E) 虛擬編碼（Dummy Encoding） (F) 降低維度（Dimension Reduction） (G) 特徵縮放比例（Feature Scaling） (H) 特徵選擇（Feature Selection）
A	162. 有一個類別型變項名稱為「鞋子顏色」，值域（亦即所有可能出現的值）為{紅, 藍, 綠}，下列何者為正確的 One-Hot Encoding 方式（變項名稱：值域）？ (E) 鞋子_紅:{1,0}，鞋子_藍:{1,0}，鞋子_綠:{1,0} (F) 鞋子顏色:{紅, 藍, 綠} (G) 鞋子顏色:{1, 2, 3} (H) 鞋子_紅:{1, 2}，鞋子_藍:{1, 2}，鞋子_綠:{1, 2}
A	163. 下列哪個方法需要類別標籤（Label）資訊？

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) 線性判別分析（Linear Discriminant Analysis） (B) 主成分分析（Principle Component Analysis） (C) 潛在語意分析（Latent Semantic Analysis） (D) 獨立成分分析（Independent Component Analysis）
A	164. 關於屬性挑選，下列敘述何者正確？ (E) 過濾式屬性挑選法（Filter）單就預測變數空間中進行挑選工作 (F) 封裝式屬性挑選法（Wrapper）不考慮後續建模方式的方法 (G) 大數據（Big Data）時代下，用越多屬性詮釋反應變數越好 (H) 屬性挑選時通常僅需考慮個別屬性的分佈狀況，無須考慮屬性間的互動關係
C	165. 若您在分析資料後發現，由於資料的某些欄位具有高度相關性而影響了分析結果，請問可能是忘了進行下列哪一步驟？ (E) 清理資料 (F) 轉換資料結構 (G) 屬性挑選 (H) 蒐集的資料不正確
C	166. 擷取網頁資料時，通常會透過 HTML 的規則進行網頁解析（Parse），下列敘述何者不正確？ (E) CSS 選擇器（Selector）是常用解析網頁元素的一種語法 (F) XPath（XML Path Language）是常用解析網頁元素的一種語法 (G) cURL 是常用解析網頁元素的一種語法 (H) 絕大部分的網頁都是以 HTML 格式來呈現的
C	167. 可以透過哪些方式取得 HTML 表單？ (1) GET (2) HOLD (3) POST (4) PUSH (E) (1) 和 (2) (F) (2) 和 (3) (G) (1) 和 (3) (H) (1) 和 (4)
D	168. 關於巨量資料進行機器學習建模的觀念，下列敘述何者正確？ (A) 以原始資料進行建模模型 (B) 資料重量不重質 (C) 無須對資料背景理解 (D) 資料轉為結構化資料
B	169. 在擁有大量且需要即時處理的資料時，下列何者並非最必要之技術？

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) 訊息佇列（Message Queue） (B) 虛擬化（Virtualization） (C) 非關聯式資料庫（NoSQL） (D) 串流技術（Streaming）
C	170. 關於巨量資料處理，下列敘述何者正確？ (A) 任何資料皆可作為特定分析目的訓練樣本 (B) 巨量資料缺少某些變量不會影響判斷結果 (C) 可透過網路爬蟲或 API 來搜集大量外部資料 (D) 透過巨量資料處理可由機器完全取代人工判斷
C	171. 關於卡方分配（Chi-squared Distribution），下列敘述何者不正確？ (A) 卡方分配的曲線為非對稱的 (B) 卡方分配的期望值為其自由度 (C) 卡方分配的期望值與變異數相等 (D) 卡方分配的自由度越大會使其變異數越大
B	172. 參考以下報表之結果，何者敘述為正確？ `> summary(faithful)` `eruptions waiting` `Min. :1.600 Min. :43.0` `1st Qu.:2.163 1st Qu.:58.0` `Median :4.000 Median :76.0` `Mean :3.488 Mean :70.9` `3rd Qu.:4.454 3rd Qu.:82.0` `Max. :5.100 Max. :96.0` (A) eruptions 變數的最小值為 43.0 (B) eruptions 變數的 75 百分位數為 4.454 (C) waiting 變數的最大值為 82.0 (D) waiting 變數的中位數為 70.9
A	173. 針對 3 種不同飲料與 4 種不同商店進行二因子變異數分析，則交互作用的自由度為多少？ (A) 6 (B) 7 (C) 12 (D) 20
D	174. 請參考以下迴歸分析結果：

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！

	`` Call: lm(formula = weight ~ height, data = women) Residuals: Min 1Q Median 3Q Max -1.7333 -1.1333 -0.3833 0.7417 3.1167 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -87.51667 5.93694 -14.74 1.71e-09 *** height 3.45000 0.09114 37.85 1.09e-14 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 1.525 on 13 degrees of freedom Multiple R-squared: 0.991, Adjusted R-squared: 0.9903 F-statistic: 1433 on 1 and 13 DF, p-value: 1.091e-14 ``` 下列敘述何者正確？ (A) 模型不具有線性模型解釋能力 (B) 模型的截距為 3.45000 (C) 模型的樣本數為 13 (D) 模型的判定係數（Coefficient of Determination）為 0.991
A	175. 預測建模前經常會先以各式統計量數，移除無用的預測變數，下列敘述何者正確？ (A) 數值型預測變數可以其間的相關係數，剔除贅餘的預測變數 (B) 線性迴歸模型中納入退化分佈（Degenerate Distribution）的預測變數，並不會損傷其績效 (C) percentUnique 是以最常見的類別值頻次，除以次常見類別值頻次的比值，來辨識有退化分佈現象的類別型變數 (D) freqRatio 是以獨一無二的類別值數量與樣本大小的比值，來辨識有退化分佈現象的類別型變數
A	176. 重抽樣方法（Resampling Methods）是當代統計學不可或缺的工具之一，下列敘述何者錯誤？ (A) 拔靴取樣法（Bootrapping）採行多次不置回抽樣的方式，取出與原樣本大小相同的子集 (B) k 摺交叉驗證（K-fold Cross Validation）是隨機 k 等分（通常是十等份）訓練集樣本後，每次留下一份作為測試集樣本，而以其餘 k-1 份樣本進行模型訓練 (C) 重抽樣方法是反覆地從訓練集或資料集中抽出或有不同的各組樣本，並重新配適各組樣本的模型，以獲得模型相關的額外資訊 (D) 拔靴取樣法與 k 摺交叉驗證兩種重抽樣方法的差別只在於樣本子集如何被挑出，而彙整與摘要統計的方式則是相同的
D	177. 機器學習經常以樣本子集進行模型訓練與測試，下列敘述何者正確？ (A) 測試集（Test Set）用以最佳化模型參數

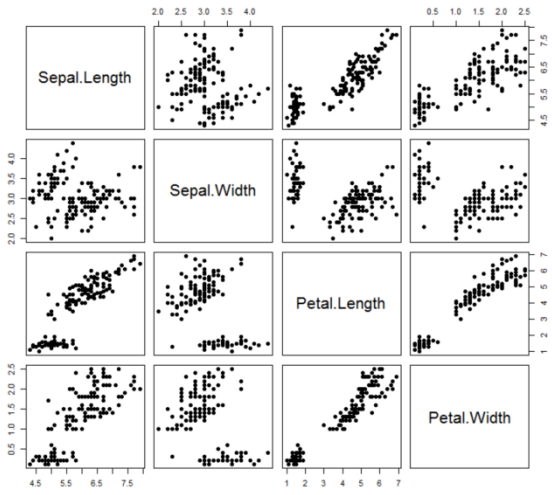
考科 2：資料處理與分析概論-參考樣題

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！

	(B) 核驗集（Validation Set）用以建立模型 (C) 訓練樣本集（Training Set）用以合理估計模型績效 (D) 訓練與測試機制經常運用重抽樣（Resampling）策略
C	178. 關於模型訓練與測試機制，下列敘述何者不正確？ (A) 模型建立與優化的步驟又可稱為（候選）模型挑選（Model Selection）的階段 (B) 用測試集對最佳模型未來之績效估計工作又可稱為模型評定/績效估計（Model Assessment/Performance Estimate）階段 (C) 模型挑選與未來績效評定估計工作應儘量使用相同的樣本進行訓練、調校與測試 (D) 對於模型挑選工作來說，當欲評估的參數組合眾多，績效準則的計算速度應為首要考量
C	179. 關於模型訓練與測試機制中的資料切分，下列敘述何者不正確？ (E) 在樣本充足的情況下，通常將之切割為兩或三個子集，分別肩負模型建立、最佳化與估計最佳化模型，對新案例的預測績效 (F) 決定最佳的模型複雜度或參數組合後，最後再以整個校驗集建立最佳複雜度或最佳參數組合下的最終模型 (G) 保留法（holdout）包含內外兩圈的重抽樣機制，分別負責模型最佳化與績效估計的工作，如此內外圈反覆執行所需計算量應是負擔最重的訓練與測試機制 (H) 雙重抽樣法充分運用了資料集中的各個樣本，基本上所有樣本都會作為訓練樣本、驗證樣本與測試樣本，但不會有樣本同時參與模型建立與未來績效的估計工作上，也就是說不會發生球員兼裁判的狀況
D	180. 關於獨立（independence）與相依（dependency），下列敘述何者正確？ (E) 統計獨立的觀念對於類別與數值變數兩種變數有不同的定義 (F) 相關（correlation）係數為 0，代表兩變數統計獨立 (G) 類別與數值變數的相依定義是相同的 (H) 類別變數的關聯（association）衡量較數值變數的相關衡量更為複雜，因為關聯性的衡量方式遠比相關性的為多
C	181. 關於「基於密度的集群分析算法（Density-Based Spatial Clustering of Applications with Noise，DBSCAN）」下列敘述何者不正確？ (A) 是以中心點為基礎的方法 (B) 能抵抗雜訊，且能處理任何形狀和大小的群集 (C) 需要人工指定集群的數目 (D) 找出遠離低密度區域之高密度的區域
C	182. 下列關於階層式集群，何者不正確？

考科 2：資料處理與分析概論-參考樣題

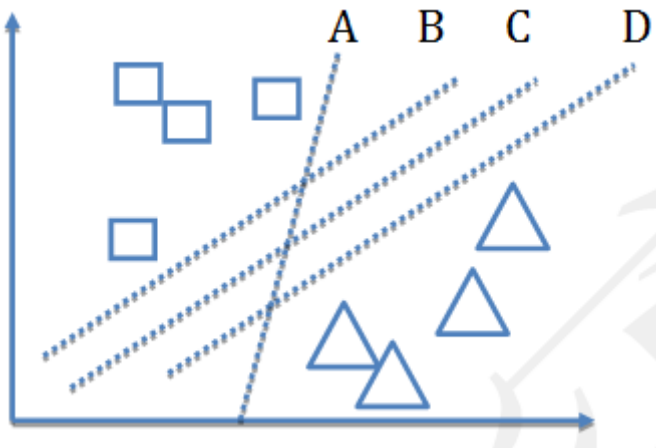
提醒！參考樣題僅協助考生瞭解考試題型及考試準備方向，並非正式的考題！

	(E) 概念簡單，可用樹狀結構來表現整個計算過程 (F) 只需要資料點兩兩之間的距離，就可以建構集群結果 (G) 即使處理大量資料也能保持高效率 (H) 不需要事先定義群數
C	183. 請問我們可以使用哪一種方法進行屬性萃取？ (E) 交替最小次方法（Alternating Least Square, ALS） (F) 二元搜索樹（Binary Search Tree） (G) 主成分分析（Principal Component Analysis, PCA） (H) K 平均法（K-Means）
C	184. 已知 iris 資料集前 3 筆資料如下圖所示： `> head(iris, n=3)` `  Sepal.Length Sepal.Width Petal.Length Petal.Width Species` `1 5.1 3.5 1.4 0.2 setosa` `2 4.9 3.0 1.4 0.2 setosa` `3 4.7 3.2 1.3 0.2 setosa` 完成以下散佈圖矩陣的 R 語言指令為何？  (A) pairs(iris) (B) pairs[iris] (C) pairs(iris[-5]) (D) pairs[iris(-5)]
B	185. 關於 K 平均法（K-Means），下列敘述何者不正確？ (E) 將資料分割成不相交的 K 個群集 (F) 如果資料與某群集中心的相似度大於其他群集，則該資料歸類於其他群集 (G) 目標是達到群集內的距離平方達到最小 (H) 目標是達到群集間的距離平方達到最大
C	186. 下列何種問題適合使用 K 平均法（K-Means）？ (A) 家庭背景對薪資的影響

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) 男性的長相對其女朋友數量的影響 (C) 根據花蕊長和花瓣寬將花朵集群 (D) 根據上下文填入克漏字
C	187. 關於 K 平均法（K-Means）集群分析，下列敘述何者不正確？ (E) 事前需要估算資料中有多少集群存在，方能執行演算法 (F) 不適合非球形或數據密度變化大的集群問題 (G) 算法只要收斂，保證可以獲得最佳的集群結果 (H) 算法可彈性變化，經過簡單的調整後，可以解決大部份的缺點
A	188. 屬性轉換（Feature Transformation）與資料縮減（Data Reduction）屬於資料前處理（Data Preprocessing）的重要工作，下列敘述何者不正確？ (A) 線性迴歸、偏最小平方法（PLS）、類神經網絡（NN）等算法內嵌有變數選擇機制的方法，對於預測變數中的雜訊，或是無訊息力的變數等較不敏感 (B) 文字資料探勘（Text Mining）中的詞頻-逆文件頻率可視為維度縮減 (C) 詞組提取（Chunk Extraction）與 N 元（N-gram）字組，算是文字資料探勘的降維方法 (D) 選用的模型種類決定資料前處理的需求
D	189. 關於非監督式學習，下列敘述何者正確？ (A) 以預測變數（predictors）來準確預測反應變數（response variable） (B) 瞭解反應變數與預測變數兩者間的關係 (C) 線性迴歸與羅吉斯迴歸都屬於非監督式學習 (D) 能發現變數間或觀測值間的子群體
A	190. 關於關聯型態探勘的特點，下列敘述何者不正確？ (A) 關聯型態探勘所得到的結果，因為可以直接進行應用，所以廣受歡迎 (B) 關聯型態分析容易從隨機的型態中妄下虛假的結論 (C) 關聯型態探勘符合資料探勘挖掘資料庫中無預期知識的理念 (D) 關聯型態探勘的分析方法對於小資料集的用處不大
A	191. 關於多元迴歸（Multiple Regression）與廣義線性模型（Generalized Linear Model，GLM），下列敘述何者不正確？ (E) 多元迴歸的自變項必須是類別資料 (F) GLM 的自變項可以是類別資料 (G) 多元迴歸的應變項必須為單一個 (H) GLM 的自變項可以是兩個以上
D	192. 關於 K 近鄰法（K-Nearest Neighbor），下列敘述何者不正確？

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) 根據現有已分類好的資料集合，找出 K 個與待分類資料最鄰近的現存資料 (B) 根據 K 個最鄰近資料之類別，以多數決的方式決定待分類資料的所屬類別 (C) 建議 K 為奇數值 (D) 根據 K 值將資料分為 K 群
A	193. 下列何者屬於線性的機器學習模型？ (A) 偏最小平方法（Partial Least Squares） (B) 貝氏分類法（Naïve Bayes） (C) 支援向量機（Support Vector Machine） (D) 類神經網路（Neural Network）
A	194. 關於常用於資料建模的演算法，下列敘述何者不正確？ (A) K 平均法（K-Means）演算法是分類演算法的一種 (B) ID3（Iterative Dichotomiser 3）演算法是決策樹演算法的一種 (C) C4.5 演算法是分類演算法的一種 (D) CART（Classification and Regression Trees）演算法是分類演算法的一種
D	195. 深度學習（Deep Learning）和下列哪一個演算法沒有直接相關？ (A) 卷積神經網路（Convolutional Neural Networks） (B) 遞歸神經網路（Recurrent Neural Network） (C) 感知學習演算法（Perceptron Learning Algorithm） (D) 關聯規則演算法（Apriori Algorithm）
C	196. 支援向量機（Support Vector Machine）源自最大邊界分類法（Maximun Margin Classifier），請問哪一條線是最大邊界分類法建立的決策邊界？  (E) A (F) B (G) C

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！

	(H) D
C	197. 梯度遞減法（Gradient Descent）是機器學習中常使用的參數收斂方式，我們可以透過參數 alpha 來調整整體收斂的速度（Step Size），請問如果 alpha 過大時，會導致什麼狀況發生？ (A) 太快收斂 (B) 收斂速度過慢 (C) 無法收斂 (D) 以上皆有可能發生
A	198. 下列何者不適合用來「預測明天會不會下雨」？ (A) 支援向量迴歸（Support Vector Regression） (B) 支援向量機（Support Vector Machine） (C) 決策樹（Decision Tree） (D) 類神經網路（Neural Network）
D	199. 關於迴歸模型績效評估，下列敘述何者正確？ (E) 評估模型績效的方式不只一種，通常只用一種模型績效評估的方式來決定模型的優劣 (F) 許多績效評量的計算是基於殘差（residual），它是模型的預測值減去觀測值 (G) 常用的績效評量 SSE 是殘差平方的平均值，而另一個常用的績效評量 MSE 是殘差平方的總和 (H) R2 也是常用的績效評量，其值表示資料中的訊息被模型所解釋的比例，R2 的解釋與結果變數的變異有關
D	200. 關於線性迴歸，下列敘述何者不正確？ (A) 並非任何資料集均可建立多元線性迴歸模型（multiple linear regression），有時會有建模失敗的狀況發生 (B) 線性迴歸屬於無母數（nonparametric）的統計建模方法 (C) 迴歸方程式係數估計最佳化問題是最小化誤差平方和（Sum of Squared Error, SSE） (D) 相較於類神經網路與支援向量機等監督式學習模型，迴歸建模所獲得的模型可解釋性較低