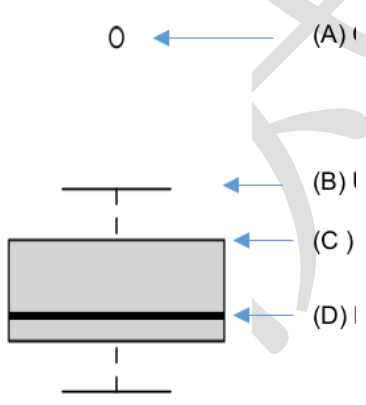


初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

第 1 頁，共 9 頁

單選題 50 題（佔 100%）

D	<p>1. 下列何者不是極端值或雜訊產生的主要原因？</p> <p>(A) 數據輸入錯誤</p> <p>(B) 測量儀器出錯</p> <p>(C) 人為故意謊報資料導致錯誤</p> <p>(D) 資料交給分析人員時，是透過電子郵件寄送而非隨身碟傳送</p>
C	<p>2. 在資料分析之前，需要花費很多力氣去整理資料，其中處理遺失值（Missing Value）便是一種，下列何者不是處理遺失值的手段？</p> <p>(A) 移除有遺失值的資料</p> <p>(B) 使用平均數或第一四分位數來填補</p> <p>(C) 將前一筆資料的值填入</p> <p>(D) 使用 K-近鄰法（K-Nearest Neighbours）搭配中位數進行填補</p>
B	<p>3. 經過網路爬蟲收集的網頁資料（如新聞網頁 HTML 格式資料）為半結構化的內容，經過解析器取得各式重要資訊，並透過詮釋資料（Metadata）結構化這些內容，這樣的過程與下列何者較為相符？</p> <p>(A) 資料擴增</p> <p>(B) 資料組織</p> <p>(C) 資訊分類</p> <p>(D) 模型預測</p>
D	<p>4. 關於盒鬚圖，下列敘述何者不正確？</p>  <p>(A) 為離群值（Outlier）</p> <p>(B) 為上圍籬（Upper Extreme）</p> <p>(C) 為 75 百分位數（Upper Quartile）</p> <p>(D) 為平均值（Mean）</p>
A	<p>5. 關於資料之遺缺值處理，下列何者不正確？</p> <p>(A) 無須考慮遺缺值比例，全部刪除</p> <p>(B) 類別資料補上眾數之值</p>

初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

第 2 頁，共 9 頁

	<p>(C) 利用模型補上估計產生之值</p> <p>(D) 透過差值法 (interpolation method) 補上該值</p>
A	<p>6. 根據下面提供的資料，老闆希望你一句話報告今年業績和去年業績的狀況，請問下列哪句話比較合適？</p> <p>2017 年業績：100 萬</p> <p>2018 年業績：120 萬</p> <p>(A) 今年業績成長了 20%</p> <p>(B) 今年業績增加了 20 萬</p> <p>(C) 今年的業績是 120 萬</p> <p>(D) 去年業績比今年少 20 萬</p>
A	<p>7. 下列何種圖形，較適合用來顯示資料隨著時間的變化趨勢？</p> <p>(A) 折線圖</p> <p>(B) 圓餅圖</p> <p>(C) 直方圖</p> <p>(D) 盒鬚圖</p>
C	<p>8. 統計圖常用來將統計資料繪製成幾何圖形，從其顯示出資料的規模、水平、結構、趨勢、比例關係，下列何者不是常用的統計圖？</p> <p>(A) 長條圖</p> <p>(B) 折線圖</p> <p>(C) 流程圖</p> <p>(D) 圓餅圖</p>
D	<p>9. 關於資料敘述與摘要統計之內容，下列敘述何者不正確？</p> <p>(A) 資料抽樣常見的有簡單隨機抽樣、系統抽樣、分層隨機抽樣</p> <p>(B) 將資料處理與製作圖表，例如：次數分配表、直方圖</p> <p>(C) 衡量資料集中趨勢的統計量，例如：平均數、中位數、眾數</p> <p>(D) 比較兩筆資料的分散程度，例如：相關係數</p>
A	<p>10. 下列哪個方法是將時間序列資料轉換到頻域空間？</p> <p>(A) 傅立葉轉換</p> <p>(B) 特徵值加權</p> <p>(C) 資料降維</p> <p>(D) 隨機抽樣</p>
B	<p>11. 對於某些資料屬性內出現異常大的值，有可能會導致誤導模型訓練的結果，此時會將該屬性值進行何種處理，使所有屬性值被轉換到 0 至 1 之間？</p> <p>(A) 資料組織</p> <p>(B) 資料特徵縮放</p> <p>(C) 資料清理</p>

初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

第 3 頁，共 9 頁

	(D) 資料分析
B	<p>12. 胖虎目前在分析一間公司的健康檢查資料，其中有一個欄位是 BMI 值，胖虎想要將其根據不同區段分類為過輕、正常、過胖、肥胖，請問胖虎正在做的是何種屬性轉換？</p> <p>(A) 二值化 (Binarization)</p> <p>(B) 分級 (Bining)</p> <p>(C) 捨入 (Rounding)</p> <p>(D) Log 轉換 (Log Transformation)</p>
D	<p>13. 下列何者不是屬性轉換的主要目的？</p> <p>(A) 轉換後可能更容易發現資料之間的關係，使沒有關係變成有關係</p> <p>(B) 資料可能呈現嚴重的偏態分布，經過轉換後差異可以拉開</p> <p>(C) 讓資料能夠符合模型所需要的假設，以利進行分析，例如經過轉換後的資料呈現正態分布</p> <p>(D) 能夠讓資料的可讀性更高</p>
D	<p>14. 下列哪種方法不屬於特徵選擇 (Feature-Selection) 的標準方法？</p> <p>(A) 嵌入方法 (Embedded)</p> <p>(B) 過濾方法 (Filter)</p> <p>(C) 包裝方法 (Wrapper)</p> <p>(D) 抽樣方法 (Sampling)</p>
A	<p>15. 關於資料特徵，下列敘述何者不正確？</p> <p>(A) 資料特徵個數越多，該模型所需的運算時間也就越短</p> <p>(B) 資料特徵個數越多，容易引起維度災難，而模型也會越複雜</p> <p>(C) 剔除不相關或多餘的資料特徵，以減少資料特徵個數，提高模型效果</p> <p>(D) 可透過模型計算資料特徵重要程度，例如：Random Forest</p>
B	<p>16. 關於巨量資料，下列敘述何者不正確？</p> <p>(A) 巨量資料分析始於找出大量資料之間的關聯性</p> <p>(B) 隨著巨量資料分析技術俱進，分析人員可以忽略數據的真實性，依然仍夠得到理想的結果</p> <p>(C) 好的巨量資料運算服務，是可以根據運算需求與時效性，平行擴增所需要的運算資源</p> <p>(D) 能妥善處理和保存大量的數據資料，即為巨量資料所談的量級 (Volume) 之特性</p>
D	<p>17. 關於 MapReduce 框架，下列敘述何者不正確？</p> <p>(A) Mapper 的輸出需要是鍵值組 (key-value pair) 的結構</p> <p>(B) 實現 Reducer，通常是定義如何處理個別鍵值下的值集合</p> <p>(C) Reducer 的輸出值通常也是鍵值組 (key-value pair) 的結構</p>

初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

第 4 頁，共 9 頁

	(D) 資料在進入 Map 階段之前會經過整理階段 (shuffle)
B	<p>18. 下列敘述何者在描述巨量資料中多樣性 (Variety) 的特性？</p> <p>(A) 能夠處理相當大的資料，例如 100TB 的歷史資料</p> <p>(B) 善於處理非結構化資料，例如各式網站資料等</p> <p>(C) 能夠大幅縮短分析的時間，能更快速反應商業需求</p> <p>(D) 能夠處理每天龐大的交易數據</p>
C	<p>19. 關於巨量資料技術架構，下列何者不是應具備的需求？</p> <p>(A) 可以被平行擴充</p> <p>(B) 儘可能能夠被分散式處理</p> <p>(C) 儘可能的使用單一節點資料庫</p> <p>(D) 具有高容錯性</p>
C	<p>20. 關於 HDFS 的文件寫入，下列敘述何者正確？</p> <p>(A) 支持多用戶對同一份文件的寫入操作</p> <p>(B) 用戶可以在文件的任意位置進行修改</p> <p>(C) 預設將文件複製三份存放</p> <p>(D) 複製的文件預設都存在同一個主機上</p>
C	<p>21. 下列何種統計量無法由盒鬚圖 (box-and-whisker plot, boxplot) 得知？</p> <p>(A) 最小值</p> <p>(B) 中位數</p> <p>(C) 變異數</p> <p>(D) 全距</p>
A	<p>22. 若兩事件 X、Y 為某試驗可能發生之二獨立事件，$P(X)>0$，$P(Y)>0$，下列何者不正確？</p> <p>(A) $P(X \cup Y) = P(X) + P(Y)$</p> <p>(B) $P(X Y) = P(X)$</p> <p>(C) $P(X Y)P(Y) = P(Y X)P(X)$</p> <p>(D) $P(X \cap Y) = P(X)P(Y)$</p>
D	<p>23. 對自變數 X 與依變數 Y 作簡單線性迴歸得到的相關係數 r，下列敘述何者正確？</p> <p>(A) $r = -1$ 代表 X 與 Y 完全無關</p> <p>(B) $r = 0$ 代表數據點恰好落在同一條水平直線上</p> <p>(C) $r > 0$ 代表 X、Y 間有因果關係</p> <p>(D) $r = 1$ 代表 $Y = aX + b$ (a、b 是常數，$a > 0$)</p>
B	<p>24. 關於單一變量的 (univariate) 統計量數，下列敘述何者不正確？</p> <p>(A) 變異係數 (coefficient of variation) 適用於量化變數</p> <p>(B) 四分位距 (inter-quartile range) 可由類別變數的次數分佈進行計算</p> <p>(C) 熵係數 (entropy coefficient) 可用於檢視類別變數次數分佈的異質</p>

初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

第 5 頁，共 9 頁

	<p>性</p> <p>(D) 異質性 (heterogeneity) 最低時集中度 (concentration) 達到最大；而異質性最高時集中度則最小</p>
A	<p>25. 關於邏輯斯迴歸中的迴歸係數，可以使用下列何種方法求解？</p> <p>(A) 最小平方法</p> <p>(B) 牛頓迭代法</p> <p>(C) 馬可夫鏈演算法</p> <p>(D) 最大概似估計法</p>
A	<p>26. 行銷部選擇部分客戶進行簡訊產品推薦，同時獲取了客戶是否願意購買產品的資訊；而通過這些已知資訊，用來判斷其他用戶的購買意願，請問屬於下列何種方法？</p> <p>(A) 推薦系統</p> <p>(B) 預測模型</p> <p>(C) 探索性分析</p> <p>(D) 關聯法則</p>
D	<p>27. 下列何者不屬於非監督式學習？</p> <p>(A) 關聯法則</p> <p>(B) K-Means</p> <p>(C) Word2Vec</p> <p>(D) K Nearest Neighbor</p>
B	<p>28. 請問下列敘述何者不正確？</p> <p>(A) 機器學習 (machine learning) 某種程度來說亦可稱為統計學習 (statistical learning)</p> <p>(B) 從所搜集的資料中建構出 X 與 Y 之間模型的過程，電腦科學的人群偏好敘述為從資料中「估計」模型參數這樣的說法，勝於從資料中「學習」的說法</p> <p>(C) Y 稱為結果變數 (outcome)</p> <p>(D) X 稱為屬性 (attributes)</p>
A	<p>29. 關於模型績效評估，下列敘述何者不正確？</p> <p>(A) 殘差 (或稱預測誤差) 是預測的反應變數值減去真實的反應變數值</p> <p>(B) 迴歸模型績效衡量大多基於殘差</p> <p>(C) 赤池弘次訊息準則 (Akaike's Information Criterion, AIC) 與舒瓦茲貝氏訊息準則 (Schwarz's Bayesian Information Criterion, BIC) 的不同在於懲罰過多變數入模的方式不同</p> <p>(D) Mallows's Cp 準則有考慮建模所用的變數數量，因此適合用來比較不同變數數量下的模型績效</p>
B	<p>30. 關於獨立 (independence) 與相依 (dependency)，下列敘述何者不正確？</p>

初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

第 6 頁，共 9 頁

	<p>(A) 獨立與相依是描述兩變數之間關係的概念</p> <p>(B) 相關（correlation）係數為 0，代表兩變數統計獨立</p> <p>(C) 關聯（association）衡量是基於頻次進行計算，用以表達兩類別變數之間的相依性</p> <p>(D) 數值變數以相關係數代表兩變數之間的相依性</p>																		
D	<p>31. 下列何種方法通常應用在集群（Clustering）問題？</p> <p>(A) Support Vector Machine</p> <p>(B) Random Forest</p> <p>(C) K Nearest Neighbors</p> <p>(D) K-Means</p>																		
C	<p>32. 下列何者不是資料降維的方法？</p> <p>(A) Principal Component Analysis</p> <p>(B) Linear Discriminant Analysis</p> <p>(C) K Nearest Neighbors</p> <p>(D) Isomap</p>																		
A	<p>33. 下列哪種圖表最能展現所有類別的總和為 100%？</p> <p>(A) 圓餅圖</p> <p>(B) 折線圖</p> <p>(C) 散布圖</p> <p>(D) 雷達圖</p>																		
C	<p>34. 某公司員工 8 人，月薪如下：</p> <table><tr><td>編號</td><td>#1</td><td>#2</td><td>#3</td><td>#4</td><td>#5</td><td>#6</td><td>#7</td><td>#8</td></tr><tr><td>月薪(千元)</td><td>22</td><td>25</td><td>25</td><td>28</td><td>30</td><td>30</td><td>60</td><td>100</td></tr></table> <p>下列敘述何者不正確？</p> <p>(A) 薪資中位數為 29 千元</p> <p>(B) 有 50%的員工，薪資≥第二四分位數</p> <p>(C) 有 50%的員工，薪資≥平均值</p> <p>(D) 繪製成箱形圖（Box plot，盒鬚圖），呈現右偏</p>	編號	#1	#2	#3	#4	#5	#6	#7	#8	月薪(千元)	22	25	25	28	30	30	60	100
編號	#1	#2	#3	#4	#5	#6	#7	#8											
月薪(千元)	22	25	25	28	30	30	60	100											
D	<p>35. 關於 K-means 集群演算法，下列敘述何者正確？</p> <p>(A) 當集群中心不再變動，就達到全局最佳解（global optimum）</p> <p>(B) 必須事先給定群組數目 K 值</p> <p>(C) 集群結果只與資料群聚分佈方式有關</p> <p>(D) 對異常值、極值的資料敏感</p>																		
D	<p>36. 以下何者不是探索性資料分析經常關心的議題？</p> <p>(A) 資料的四分位數</p> <p>(B) 資料是否有離群值</p> <p>(C) 與應變數相關的自變數</p>																		

初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

第 7 頁，共 9 頁

	(D) 資料模型的準確度
B	<p>37. 史皮爾曼相關係數 (Spearman correlation coefficient) 是一種兩兩變數相關係數計算的方式，下列敘述何者不正確？</p> <p>(A) 順序值類別變數 (ordinal qualitative variables) 適合此計算方式</p> <p>(B) 名目值類別變數 (nominal qualitative variables) 適合此計算方式</p> <p>(C) 實數值量化變數 (real-valued quantitative variables) 適合此計算方式</p> <p>(D) 史皮爾曼相關係數又稱為等級相關係</p>
D	<p>38. 關於階層式分群法 (Hierarchical Clustering)，下列敘述何者不正確？</p> <div data-bbox="635 689 1088 1032" data-label="Figure"> </div> <p>(A) 若採用聚合的方式，則由樹狀結構的底部開始，將資料或群集逐次合併</p> <p>(B) 若採用分裂的方式，則由樹狀結構的頂端開始，將群集逐次分裂</p> <p>(C) 群與群的距離定義為不同群聚中最近的兩個點的距離，該法稱為單一聯結法 (Single Linkage，又稱「最近法」)</p> <p>(D) 事先必須告知分群數量，以利分群法之進行</p>
C	<p>39. 關於 K-Means 與 DBSCAN，下列敘述何者不正確？</p> <p>(A) 兩者都是集群分析</p> <p>(B) K-Means 基於距離的概念，而 DBSCAN 基於密度的概念</p> <p>(C) 兩者都需要事先告知分群的數量</p> <p>(D) K-Means 集群結果易受離群值的影響</p>
D	<p>40. 下列何者不屬於非監督式學習的演算法？</p> <p>(A) PCA</p> <p>(B) Hierarchical-Clustering</p> <p>(C) Auto-Encoder</p> <p>(D) XGBoost</p>
B	<p>41. 下列何種方法常應用在分類問題？</p> <p>(A) Linear regression</p> <p>(B) Logistic regression</p> <p>(C) Polynomial Regression</p>

初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

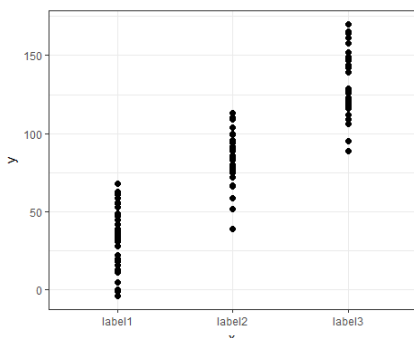
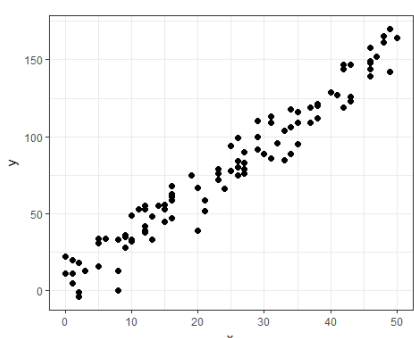
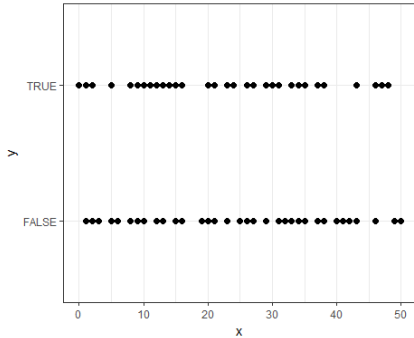
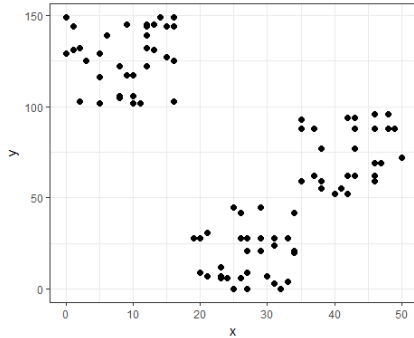
第 8 頁，共 9 頁

	(D) Support vector regression
C	<p>42. 下列何者不適合用來預測「句子的下一個詞」？</p> <p>(A) Hidden Markov model</p> <p>(B) Conditional random field</p> <p>(C) Linear Regression</p> <p>(D) N-gram</p>
A	<p>43. 下列學習方法，何者難以獲得人類容易理解的知識或特徵？</p> <p>(A) Multilayer perceptron</p> <p>(B) Decision tree</p> <p>(C) Logistic regression</p> <p>(D) Association rule mining</p>
C	<p>44. 關於配適不足（under-fitting），下列何者正確？</p> <p>(A) 訓練誤差較大，測試誤差較小</p> <p>(B) 訓練誤差較小，測試誤差較大</p> <p>(C) 訓練誤差較大，測試誤差較大</p> <p>(D) 訓練誤差較小，測試誤差較小</p>
A	<p>45. 利用多個分類器的預測來提高分類的準確率之技術為下列何者？</p> <p>(A) Ensemble</p> <p>(B) Dimensionality reduction</p> <p>(C) Pruning</p> <p>(D) Feature selection</p>
A	<p>46. 下列哪種的資料可以無需經過前處理，直接使用線性模型（Linear Model）進行學習？</p> <p>(A) 身高（公分）、體重（公斤）</p> <p>(B) 性別（男、女）、腰圍（公分）</p> <p>(C) 最高時速（公里/小時）、車款（車種型號）</p> <p>(D) 氣候（晴、陰、雨）、溫度（攝氏溫度）</p>
D	<p>47. 下列何者不為監督式學習（Supervised Learning）方法？</p> <p>(A) K 近鄰法（K-Nearest Neighbor）</p> <p>(B) 支援向量機（Support Vector Machine）</p> <p>(C) 邏輯迴歸（Logistic Regression）</p> <p>(D) 自我組織映像圖（Self-Organizing Map）</p>
B	<p>48. 建立簡單線性迴歸模型之前常會根據資料的散佈圖進行模型假設，則下列四張資料的散佈圖，何者最適合使用簡單線性迴歸模型？</p> <p>(A) (B)</p>

初級巨量資料分析師能力鑑定樣題

科目 2：資料處理與分析概論

第 9 頁，共 9 頁

	<div style="display: flex; justify-content: space-around;">   </div> <div style="display: flex; justify-content: space-around; margin-top: 20px;"> <div style="text-align: center;"> <p>(C)</p>  </div> <div style="text-align: center;"> <p>(D)</p>  </div> </div>
C	<p>49. 關於監督式學習，下列敘述何者不正確？</p> <ul style="list-style-type: none"> (A) 模型的訓練資料必須有應變項 (B) 訓練資料不一定為連續型資料 (C) 主成分分析是一種監督式學習的方法 (D) 訓練資料過少時，可利用 Bootstrap（拔靴法）進行修正
B	<p>50. 關於資料解析思維，下列敘述何者不正確？</p> <ul style="list-style-type: none"> (A) 巨量資料中雜訊多，穩健統計方法可降低雜訊對模型的影響 (B) 機器學習模型不需要考慮資料是否與背景假設吻合 (C) 利用重抽樣樣本中的不確定性，可以強化參數估計過程與避免過度配適 (D) 集成（或稱系集）模型（ensemble models）可以發揮團結力量大的效果，解決困難的問題