

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 1 頁，共 15 頁

單選題 50 題（佔 100%）

| B | 1. 關於支援向量機（Support Vector Machine, SVM）的模型超參數（hyperparameters），下列敘述何者「不」正確？ (A) 懲罰係數 C 越高，越容易過度最佳化 (B) 支援向量的數目要事先決定 (C) 核函數（kernel function）要事先決定 (D) 網格搜尋（Grid Search）常用來尋找超參數（hyperparameters） | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|----------|--|----------|---|---|----|---|---|----|---|---|----|---|---|----|---|---|----|---|----|----|---|----|----|---|---|----|---|----|----|----|---|-----|---|---|
| C | 2. 關於求解線性迴歸 $Y=\beta^T X=\beta_0*1+\beta_1x_1+\beta_2*x_2+...+\beta_m*x_m$ ，其中 $\beta^T=[\beta_0, \beta_1, ..., \beta_m]$ ，下列敘述何者「不」正確？ (A) 簡單線性迴歸，只包括一個自變量和一個因變量，可使用最小平方方法求解 (B) 簡單線性迴歸解得模型是否有效，需要計算相關係數 r 以確認 X 對 Y 有顯著的影響，並呈密切的線性相關 (C) 多元線性迴歸自變數之間的相關程度可以高於自變數與因變數之間的相關程度 (D) 多元迴歸公式的結構等於一個只有輸入和輸出層的神經網路，可用隨機梯度下降法（Stochastic gradient descent, SGD）去找 β^T 的最佳解 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| A | 3. 關於下表資料，使用 k 平均數集群（k-means clustering）分析法與歐氏距離（Euclidean distance），將資料分成三個集群（假設 k=3）。 <table><tr><th>Point ID</th><th>X</th><th>Y</th></tr><tr><td>P1</td><td>0</td><td>0</td></tr><tr><td>P2</td><td>2</td><td>3</td></tr><tr><td>P3</td><td>4</td><td>2</td></tr><tr><td>P4</td><td>0</td><td>6</td></tr><tr><td>P5</td><td>3</td><td>10</td></tr><tr><td>P6</td><td>4</td><td>11</td></tr><tr><td>P7</td><td>6</td><td>9</td></tr><tr><td>P8</td><td>8</td><td>10</td></tr><tr><td>P9</td><td>12</td><td>6</td></tr><tr><td>P10</td><td>7</td><td>9</td></tr></table> (A) C1: (P1, P2, P3, P4), C2: (P5, P6, P7, P8, P10), C3: (P9) (B) C1: (P1, P2, P4, P5), C2: (P3, P6, P7, P8, P10), C3: (P9) (C) C1: (P1, P2, P3, P4, P5, P6), C2: (P7, P8, P10), C3: (P9) (D) C1: (P1, P2, P3), C2: (P4, P5, P6, P7, P8, P10), C3: (P9) | Point ID | X | Y | P1 | 0 | 0 | P2 | 2 | 3 | P3 | 4 | 2 | P4 | 0 | 6 | P5 | 3 | 10 | P6 | 4 | 11 | P7 | 6 | 9 | P8 | 8 | 10 | P9 | 12 | 6 | P10 | 7 | 9 |
| Point ID | X | Y | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P1 | 0 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P2 | 2 | 3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P3 | 4 | 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P4 | 0 | 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P5 | 3 | 10 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P6 | 4 | 11 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P7 | 6 | 9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P8 | 8 | 10 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P9 | 12 | 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P10 | 7 | 9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 2 頁，共 15 頁

| C | 4. 用主成分分析 (principle component analysis) 來處理多維度資料時，會利用相關矩陣 (correlation matrix) 來計算特徵值 (eigenvalue) 與特徵向量 (eigenvector)，如果特徵向量 $\lambda = [4.32, 1.07, 0.49, 0.10, 0.01, 0.01]$ ，下列敘述何者正確？ (A) 主特徵值的貢獻率達到 80% (B) 主特徵值的貢獻率達到 90% (C) 前兩個特徵值的貢獻率達到 90% (D) 前兩個特徵值的代表性不足 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---------------|---|---------------|--------------|----------|------------|---|-----|---|----|---|-----|-----|----|---|-----|---|----|---|-----|-----|----|---|------|-----|----|---|-----|---|----|---|-----|-----|----|---|-----|-----|----|---|---|---|----|
| C | 5. 下列何者是用來衡量「類別變數次數分佈」異質性的方法？ (A) 變異數 (B) 四分位距 (C) 熵 (entropy) 係數 (D) 中位數絕對離差 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| D | 6. 當訓練資料有遺缺資料 (missing data)，某些樣本缺少某些維度的資料時，下列何者「不是」處理方式？ (A) 將遺缺資料的樣本移除 (B) 以原始資料的平均值補上遺缺資料 (C) 將遺缺資料用「合理」的數字補上 (D) 將資料分成 n 組交叉檢驗 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B | 7. 當身體肌肉收縮時，肌電訊號 (Electromyography, EMG) 肌電信號可用於控制計算機，作為一種用戶界面。EMG 訊號一般可由：頻率 (Frequency, F)、強度 (Strength, S) 與時間 (Time, T) 來表示，下表為 EMG 的實驗資料 (F, S, T) 和相應的動作分類 (Action, A)，若用 Gini 係數來建立決策樹模型，第一個分類屬性為下列何者？ <table><tr><th>Frequency (F)</th><th>Strength (S)</th><th>Time (T)</th><th>Action (A)</th></tr><tr><td>1</td><td>810</td><td>1</td><td>A1</td></tr><tr><td>1</td><td>864</td><td>0.5</td><td>A2</td></tr><tr><td>1</td><td>485</td><td>1</td><td>A3</td></tr><tr><td>1</td><td>950</td><td>0.5</td><td>A2</td></tr><tr><td>1</td><td>1003</td><td>0.5</td><td>A2</td></tr><tr><td>1</td><td>524</td><td>1</td><td>A3</td></tr><tr><td>1</td><td>736</td><td>0.5</td><td>A4</td></tr><tr><td>1</td><td>661</td><td>0.5</td><td>A4</td></tr><tr><td>2</td><td>*</td><td>*</td><td>A5</td></tr></table> | Frequency (F) | Strength (S) | Time (T) | Action (A) | 1 | 810 | 1 | A1 | 1 | 864 | 0.5 | A2 | 1 | 485 | 1 | A3 | 1 | 950 | 0.5 | A2 | 1 | 1003 | 0.5 | A2 | 1 | 524 | 1 | A3 | 1 | 736 | 0.5 | A4 | 1 | 661 | 0.5 | A4 | 2 | * | * | A5 |
| Frequency (F) | Strength (S) | Time (T) | Action (A) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 810 | 1 | A1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 864 | 0.5 | A2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 485 | 1 | A3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 950 | 0.5 | A2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 1003 | 0.5 | A2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 524 | 1 | A3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 736 | 0.5 | A4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 661 | 0.5 | A4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | * | * | A5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 3 頁，共 15 頁

| | |
|---|--|
| | (A) 頻率 (Frequency, F) (B) 強度 (Strength, S) (C) 時間 (Time, T) (D) 動作分類 (Action, A) |
| A | 8. 分類問題當不同類的樣本數不平衡時，下列何者「不是」處理方式？ (A) 使用丟棄 (dropout) 方法從大類中剔除一些樣本 (B) 使用降抽樣 (undersampling) 方法從大類中選取部分樣本 (C) 使用權重 (weighting) 方法調整樣本權重 (D) 使用數據合成 (synthetic) 方法生成新的樣本 |
| D | 9. 下列何者「不是」用來評估模型的驗證指標 (validation index)？ (A) 均方誤差 (Mean Squared Error, MSE) (B) 混淆矩陣 (confusion matrix) 與敏感度 (sensitivity) (C) ROC 曲線與曲線下方面積 (Area Under Curve, AUC) (D) 特徵值 (eigenvalue) |
| B | 10. 關於接收者操作特性曲線 (Receiver Operating characteristic Curve, ROC)，下列敘述何者「不」正確？ (A) ROC 曲線往左上角移動，代表模型的敏感度越高 (B) AUC=0.8，代表無鑑別力 (C) ROC 常用於分析二元分類模型 (D) ROC 曲線是以假陽性率 (False Positive Rate, FPR) 為 X 軸，以真陽性率 (True Positive Rate, TPR) 為 Y 軸 |
| A | 11. 與隨機誤差建模相關的參數有模型參數和超參數 (hyperparameters)，下列何者「不是」超參數？ (A) 迴歸方程式的截距 (B) 支援向量機中徑向基底核函數的 σ (C) 人工神經網路的隱藏層節點數 (D) 建構迴歸方程式的預測變數集 |
| B | 12. 真實的反應變數值與預測的反應變數值之間的差，稱為殘差或 (預測) 誤差，下列何者是應用殘差平方值的算術平均來評估迴歸模型的績效？ (A) 誤差平方和 (Sum of the Squared Errors, SSE) (B) 均方預測誤差 (Mean Squared Error, MSE) (C) 均方根預測誤差 (Root Mean Squared Error, RMSE) (D) 誤差絕對值和 (Sum of Absolute Error, SAE) |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 4 頁，共 15 頁

| | |
|---|---|
| C | <p>13. 交叉驗證 (cross-validation) 主要用於模型訓練或建模應用中，目的是為了得到可靠穩定的模型。請問下列敘述何者「不」正確？</p> <p>(A) k 摺交叉驗證 (k-fold cross validation)，若 $k=10$，代表將數據集分成 10 份，將其中 9 份做訓練、1 份做驗證</p> <p>(B) 交叉驗證經常用於分類預測、偏最小平方迴歸 (Partial Least Squares regression, PLS 迴歸) 建模等</p> <p>(C) 採用 k 摺交叉驗證 (k-fold cross validation) 通常會重複 k 次以上，以 k-1 次的結果均值作為對算法精度的估計</p> <p>(D) 交叉驗證的類型，常見的有保留法 (holdout) 驗證、k 摺驗證</p> |
| B | <p>14. 當資料科學家建模時，如果有測試誤差高，訓練誤差低的狀況時，稱為下列何者？</p> <p>(A) 配適不足</p> <p>(B) 過度配適</p> <p>(C) 配適良好</p> <p>(D) 配適狀況不明</p> |
| D | <p>15. 模型選擇與評定時，經常運用重抽樣方法進行模型訓練與測試，下列敘述何者正確？</p> <p>(A) 模型評定 (model assessment) 的工作包括同一模型不同參數的調校 (within model)，以及跨越不同模型的比較 (between models)</p> <p>(B) 模型優化 (model optimization) 的工作則是在確定最優模型後，合理地估計其未來實際應用上可能的績效表現</p> <p>(C) 與隨機誤差建模相關的參數 (parameters) 有兩種：一種可以直接利用資料估計其值的超參數 (hyperparameters)，另一種則是不易從資料中估計的模型參數</p> <p>(D) 一般而言 k 摺交叉驗證 (k-fold cross validation) 相較於他法有較高的變異，但當訓練集大時則此問題較不嚴重</p> |
| C | <p>16. 有一個混淆矩陣 (confusion matrix)，橫列表示預測類別，縱行表示真實類別，假設有一個預測類別矩陣為 [0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1]，真實類別矩陣為 [1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1]，則假陽數 (false positive) 的值為何？</p> <p>(A) 3</p> <p>(B) 5</p> <p>(C) 2</p> <p>(D) 6</p> |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 5 頁，共 15 頁

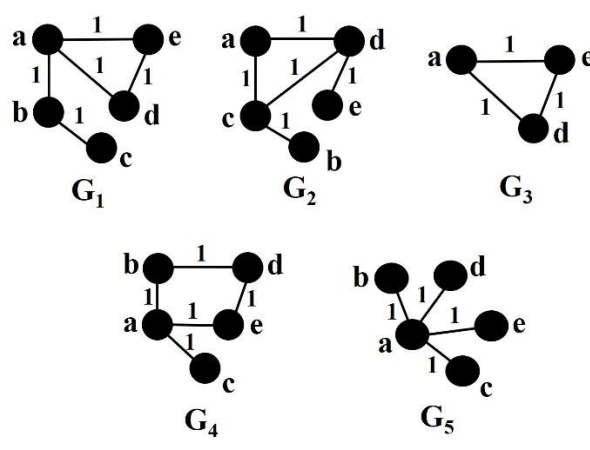
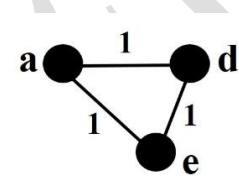
| D | <p>17. 此例交易資料用 Apriori 演算法尋找頻繁項目集 (frequent itemsets)，最小支持度 (minimum support) 要設多少，才會有長度為 3 的頻繁項目集？</p> <table><tr><th>TID</th><th>List of item IDs</th></tr><tr><td>T100</td><td>I1, I2, I5</td></tr><tr><td>T200</td><td>I2, I4</td></tr><tr><td>T300</td><td>I2, I3</td></tr><tr><td>T400</td><td>I1, I2, I4</td></tr><tr><td>T500</td><td>I1, I3</td></tr><tr><td>T600</td><td>I2, I3</td></tr><tr><td>T700</td><td>I1, I3</td></tr><tr><td>T800</td><td>I1, I2, I3, I5</td></tr><tr><td>T900</td><td>I1, I2, I3</td></tr></table> <p>(A) 0.3 (B) 0.27 (C) 0.25 (D) 0.22</p> | TID | List of item IDs | T100 | I1, I2, I5 | T200 | I2, I4 | T300 | I2, I3 | T400 | I1, I2, I4 | T500 | I1, I3 | T600 | I2, I3 | T700 | I1, I3 | T800 | I1, I2, I3, I5 | T900 | I1, I2, I3 |
|------|--|-----|------------------|------|------------|------|--------|------|--------|------|------------|------|--------|------|--------|------|--------|------|----------------|------|------------|
| TID | List of item IDs | | | | | | | | | | | | | | | | | | | | |
| T100 | I1, I2, I5 | | | | | | | | | | | | | | | | | | | | |
| T200 | I2, I4 | | | | | | | | | | | | | | | | | | | | |
| T300 | I2, I3 | | | | | | | | | | | | | | | | | | | | |
| T400 | I1, I2, I4 | | | | | | | | | | | | | | | | | | | | |
| T500 | I1, I3 | | | | | | | | | | | | | | | | | | | | |
| T600 | I2, I3 | | | | | | | | | | | | | | | | | | | | |
| T700 | I1, I3 | | | | | | | | | | | | | | | | | | | | |
| T800 | I1, I2, I3, I5 | | | | | | | | | | | | | | | | | | | | |
| T900 | I1, I2, I3 | | | | | | | | | | | | | | | | | | | | |
| D | <p>18. 集群 (clustering) 是以非監督 (unsupervised) 方式定義其欲解決的問題，所以只能透過一些常用的內部核驗準則來評估結果，下列何者「不是」內部核驗準則？</p> <p>(A) 各群樣本點到中心距離的平方和 (B) 群內距離相對於群間距離的比值 (C) 側影係數 (silhouette coefficient) (D) 類別標籤</p> | | | | | | | | | | | | | | | | | | | | |
| A | <p>19. 下列何種集群法 (clustering) 是利用兩兩樣本 (或群) 間的距離與樹狀結構，將資料進行分群，一開始會將所有的資料視為一個完整的群體，在迭代過程中不斷的分裂為較小的群體，直到所有的資料都成為單獨的個體？</p> <p>(A) 階層式集群 (B) 密度集群 (C) k-means 集群 (D) k-medoids 集群</p> | | | | | | | | | | | | | | | | | | | | |
| B | <p>20. 下列何者是 k 平均數集群 (k-means clustering) 的優點？</p> <p>(A) 算法涉及隨機抽樣，每次運行的結果不盡相同 (B) 原理簡單，容易以非統計的詞彙解釋說明之</p> | | | | | | | | | | | | | | | | | | | | |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 6 頁，共 15 頁

| | |
|---|---|
| | <p>(C) 形成的群多為類圓球狀且大小相近</p> <p>(D) 不易受到離群值的影響</p> |
| B | <p>21. 圖形資料集如圖 (1)：</p>  <p>(1)</p> <p>則圖 (2)</p>  <p>(2)</p> <p>的支持度為？</p> <p>(圖取自：資料探勘/Pang-Ning Tan, Michael Steinbach, 作者：施雅月，出版社：臺灣培生教育/歐亞)</p> <p>(A) 20%</p> <p>(B) 40%</p> <p>(C) 60%</p> <p>(D) 80%</p> |
| B | <p>22. 關於關聯分析：FP-growth 演算法 (Frequent Pattern-growth)，下列敘述何者「不」正確？</p> <p>(A) 這是一種沒有生成候選項目集的頻繁項目集探勘方法</p> <p>(B) 採用類似 Apriori 方法的生成和測試 (generate-and-test) 策略</p> <p>(C) 它構造了一個高度緊湊的資料結構 (FP-tree) 來壓縮原始交易資料庫</p> <p>(D) 它著重於頻繁項目的增長，避免昂貴的候選生成</p> |
| C | <p>23. 下表是 7 個資料點的 (x, y) 值，假設現在分成三個集群 $A=\{P1, P3, P6\}$, $B=\{P2, P4\}$, $C=\{P5, P7\}$，若以歐氏距離 (Euclidean distance) 平</p> |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 7 頁，共 15 頁

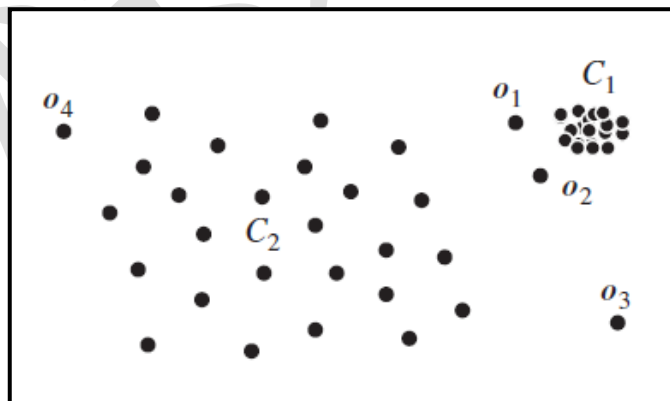
方作為衡量相似度的依據，則集群 A 與 B 間共有 6 個兩兩資料點的距離： $D(P1, P2)=233$, $D(P1, P4)=261$, $D(P3, P2)=149$, $D(P3, P4)=169$, $D(P6, P2)=80$, $D(P6, P4)=104$ 。

| Point ID | x | y |
|----------|----|----|
| P1 | 14 | 15 |
| P2 | 22 | 28 |
| P3 | 15 | 18 |
| P4 | 20 | 30 |
| P5 | 30 | 35 |
| P6 | 18 | 20 |
| P7 | 32 | 30 |

則下列敘述何者「不」正確？

- (A) 若以最小距離 (minimum distance) 作為集群相似度的衡量，則集群 A 與 B 間的距離是 80
- (B) 若以最大距離 (maximum distance) 作為集群相似度的衡量，則集群 A 與 B 間的距離是 261
- (C) 若以平均距離 (average distance) 作為集群相似度的衡量，則集群 A 與 B 間的距離是 160
- (D) 若以中心值距離 (centroid distance) 作為集群相似度的衡量，則集群 A 與 B 間的距離是 156.89

- D 24. 以密度為基礎的離群值偵測方法，下圖中有哪些點最可能是離群值 (Outlier)？



- (A) O4
- (B) O1, O2
- (C) O3
- (D) O1, O2, O3

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 8 頁，共 15 頁

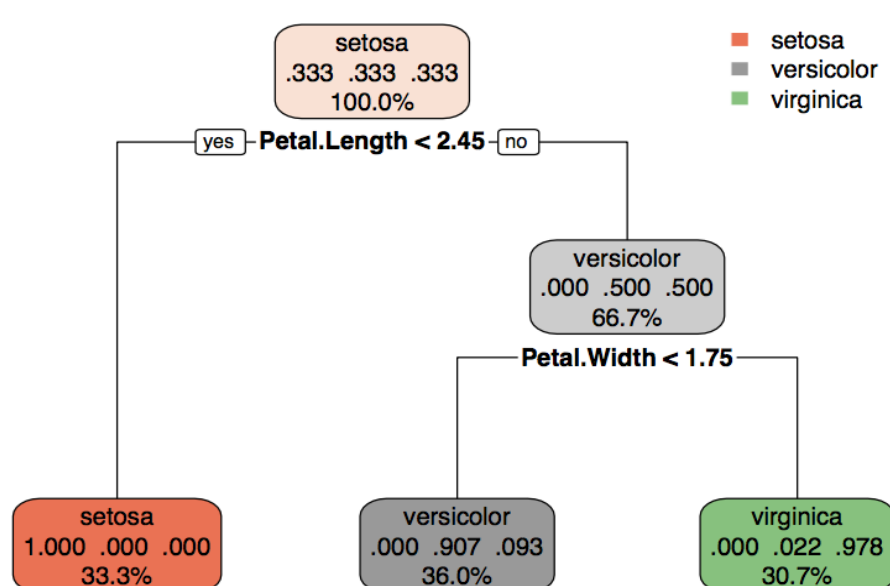
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|-----------|---|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------|-----|-----|------|------|------|------|------|-----|------|------|-----|-----|------|-----|------|-----|------|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| B | <p>25. 生產智慧自動化系統，達成即時多結構化的資料彙整，資料來源包括現場生產管制（Shop Floor Control, SFC）所產生的生產機台測試資料、生產機台參數與生產執行系統（Manufacturing Execution System, MES）中的工單相關資料，依需求定義「資料處理與運算邏輯」，產生資料分析所需要生產總表與對應統計表，透過監控製程數據早期發現異常，產生警告訊號。</p> <p>下圖如同時對「Sample #1~4」以每小時取樣，同時了解資料的偏斜性（skewness）及離群值（outliers），建議使用何種圖表分析？</p> <table><tr><td>TIME</td><td>23:33</td><td>0:33</td><td>1:33</td><td>2:33</td><td>3:33</td><td>4:33</td><td>5:33</td><td>6:33</td><td>7:33</td><td>8:33</td><td>9:33</td><td>10:33</td><td>11:33</td><td>12:33</td><td>13:33</td><td>14:33</td><td>15:33</td></tr><tr><td>Sample #0</td><td>24.6</td><td>35.8</td><td>32.5</td><td>34.5</td><td>34.8</td><td>31.5</td><td>27.9</td><td>30.7</td><td>35.6</td><td>31.9</td><td>24.1</td><td>27.5</td><td>33.8</td><td>22.7</td><td>35.0</td><td>32.0</td><td>35.0</td></tr><tr><td>Sample #1</td><td>29.4</td><td>29.0</td><td>29.8</td><td>26.9</td><td>34.6</td><td>24.5</td><td>32.2</td><td>37.0</td><td>25.3</td><td>33.9</td><td>25.8</td><td>27.2</td><td>26.0</td><td>26.6</td><td>30.5</td><td>36.1</td><td>28.4</td></tr><tr><td>Sample #2</td><td>32.4</td><td>31.0</td><td>27.6</td><td>22.9</td><td>24.6</td><td>25.5</td><td>25.6</td><td>34.0</td><td>36.0</td><td>33.6</td><td>31.0</td><td>26.2</td><td>33.0</td><td>28.1</td><td>23.1</td><td>27.7</td><td>23.6</td></tr><tr><td>Sample #3</td><td>23.6</td><td>36.5</td><td>37.3</td><td>33.7</td><td>28.7</td><td>35.5</td><td>37.0</td><td>37.3</td><td>34.5</td><td>24.3</td><td>28.7</td><td>32.7</td><td>36.9</td><td>28.3</td><td>24.5</td><td>26.8</td><td>35.7</td></tr><tr><td>Sample #4</td><td>26.7</td><td>32.6</td><td>22.8</td><td>25.4</td><td>28.4</td><td>37.1</td><td>31.5</td><td>31.6</td><td>24.4</td><td>37.4</td><td>23.9</td><td>32.7</td><td>27.2</td><td>29.6</td><td>27.3</td><td>34.1</td><td>35.5</td></tr><tr><td>MEAN</td><td>27.3</td><td>33.0</td><td>30.0</td><td>28.7</td><td>30.2</td><td>30.8</td><td>30.9</td><td>34.1</td><td>31.2</td><td>32.2</td><td>26.7</td><td>29.3</td><td>31.4</td><td>27.1</td><td>28.1</td><td>31.3</td><td>31.7</td></tr><tr><td>RANGE</td><td>8.7</td><td>7.5</td><td>14.5</td><td>11.6</td><td>10.2</td><td>12.6</td><td>11.4</td><td>6.6</td><td>11.6</td><td>13.0</td><td>7.1</td><td>6.5</td><td>10.9</td><td>6.9</td><td>11.9</td><td>9.3</td><td>12.0</td></tr><tr><td>SUM</td><td>136.6</td><td>164.9</td><td>150.0</td><td>143.4</td><td>151.1</td><td>154.1</td><td>154.3</td><td>170.7</td><td>155.8</td><td>161.1</td><td>133.5</td><td>146.3</td><td>156.9</td><td>135.4</td><td>140.4</td><td>156.6</td><td>158.3</td></tr></table> <p>(A) 管制圖（control chart） (B) 盒鬚圖（boxplot） (C) 圓餅圖（pie chart） (D) 直方圖（histogram）</p> | TIME | 23:33 | 0:33 | 1:33 | 2:33 | 3:33 | 4:33 | 5:33 | 6:33 | 7:33 | 8:33 | 9:33 | 10:33 | 11:33 | 12:33 | 13:33 | 14:33 | 15:33 | Sample #0 | 24.6 | 35.8 | 32.5 | 34.5 | 34.8 | 31.5 | 27.9 | 30.7 | 35.6 | 31.9 | 24.1 | 27.5 | 33.8 | 22.7 | 35.0 | 32.0 | 35.0 | Sample #1 | 29.4 | 29.0 | 29.8 | 26.9 | 34.6 | 24.5 | 32.2 | 37.0 | 25.3 | 33.9 | 25.8 | 27.2 | 26.0 | 26.6 | 30.5 | 36.1 | 28.4 | Sample #2 | 32.4 | 31.0 | 27.6 | 22.9 | 24.6 | 25.5 | 25.6 | 34.0 | 36.0 | 33.6 | 31.0 | 26.2 | 33.0 | 28.1 | 23.1 | 27.7 | 23.6 | Sample #3 | 23.6 | 36.5 | 37.3 | 33.7 | 28.7 | 35.5 | 37.0 | 37.3 | 34.5 | 24.3 | 28.7 | 32.7 | 36.9 | 28.3 | 24.5 | 26.8 | 35.7 | Sample #4 | 26.7 | 32.6 | 22.8 | 25.4 | 28.4 | 37.1 | 31.5 | 31.6 | 24.4 | 37.4 | 23.9 | 32.7 | 27.2 | 29.6 | 27.3 | 34.1 | 35.5 | MEAN | 27.3 | 33.0 | 30.0 | 28.7 | 30.2 | 30.8 | 30.9 | 34.1 | 31.2 | 32.2 | 26.7 | 29.3 | 31.4 | 27.1 | 28.1 | 31.3 | 31.7 | RANGE | 8.7 | 7.5 | 14.5 | 11.6 | 10.2 | 12.6 | 11.4 | 6.6 | 11.6 | 13.0 | 7.1 | 6.5 | 10.9 | 6.9 | 11.9 | 9.3 | 12.0 | SUM | 136.6 | 164.9 | 150.0 | 143.4 | 151.1 | 154.1 | 154.3 | 170.7 | 155.8 | 161.1 | 133.5 | 146.3 | 156.9 | 135.4 | 140.4 | 156.6 | 158.3 |
| TIME | 23:33 | 0:33 | 1:33 | 2:33 | 3:33 | 4:33 | 5:33 | 6:33 | 7:33 | 8:33 | 9:33 | 10:33 | 11:33 | 12:33 | 13:33 | 14:33 | 15:33 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sample #0 | 24.6 | 35.8 | 32.5 | 34.5 | 34.8 | 31.5 | 27.9 | 30.7 | 35.6 | 31.9 | 24.1 | 27.5 | 33.8 | 22.7 | 35.0 | 32.0 | 35.0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sample #1 | 29.4 | 29.0 | 29.8 | 26.9 | 34.6 | 24.5 | 32.2 | 37.0 | 25.3 | 33.9 | 25.8 | 27.2 | 26.0 | 26.6 | 30.5 | 36.1 | 28.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sample #2 | 32.4 | 31.0 | 27.6 | 22.9 | 24.6 | 25.5 | 25.6 | 34.0 | 36.0 | 33.6 | 31.0 | 26.2 | 33.0 | 28.1 | 23.1 | 27.7 | 23.6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sample #3 | 23.6 | 36.5 | 37.3 | 33.7 | 28.7 | 35.5 | 37.0 | 37.3 | 34.5 | 24.3 | 28.7 | 32.7 | 36.9 | 28.3 | 24.5 | 26.8 | 35.7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sample #4 | 26.7 | 32.6 | 22.8 | 25.4 | 28.4 | 37.1 | 31.5 | 31.6 | 24.4 | 37.4 | 23.9 | 32.7 | 27.2 | 29.6 | 27.3 | 34.1 | 35.5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MEAN | 27.3 | 33.0 | 30.0 | 28.7 | 30.2 | 30.8 | 30.9 | 34.1 | 31.2 | 32.2 | 26.7 | 29.3 | 31.4 | 27.1 | 28.1 | 31.3 | 31.7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| RANGE | 8.7 | 7.5 | 14.5 | 11.6 | 10.2 | 12.6 | 11.4 | 6.6 | 11.6 | 13.0 | 7.1 | 6.5 | 10.9 | 6.9 | 11.9 | 9.3 | 12.0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| SUM | 136.6 | 164.9 | 150.0 | 143.4 | 151.1 | 154.1 | 154.3 | 170.7 | 155.8 | 161.1 | 133.5 | 146.3 | 156.9 | 135.4 | 140.4 | 156.6 | 158.3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| B | <p>26. 關於離群值檢測(outlier detection)可能面臨的挑戰，下列敘述何者「不」正確？</p> <p>(A) 異常值檢測高度依賴於正常（非異常值）和異常值的有效建模 (B) 一般檢測方法與應用無關（application-independent） (C) 異常值與雜訊不同 (D) 檢測到異常值具備易理解性（understandability）</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| C | <p>27. 關於決策樹（decision tree）的使用，下列敘述何者「不」正確？</p> <p>(A) 每次分割要讓子節點不純度降低 (B) 樹葉節點中的樣本應屬於同一類別 (C) 為了進行準確的預測，要讓決策樹盡量長大長深 (D) 結合多個決策樹形成隨機森林（random forest），模型較穩健，不易過度配適</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| A | <p>28. 關於 k 近鄰（k-nearest neighbors）學習，下列敘述何者「不」正確？</p> <p>(A) 資料需服從常態分配 (B) 因為沒有模型，所以限制了我們瞭解預測變數與目標變數之間關係的能力 (C) 度量綱不一的屬性、名目屬性與遺缺數據需要額外處理 (D) k 近鄰法計算耗時，計算時須將數據載入記憶體中，當資料量大時通常用節省記憶體的資料結構，以加快計算，如：k 維樹（k-</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 9 頁，共 15 頁

| | |
|---|---|
| | dimensional tree, k-d tree) |
| B | <p>29. 鳶尾花 iris 資料集共有 150 個樣本，花瓣長寬、花萼長寬與花種 (setosa、versicolor、virginica) 等五個變數，關於下圖中樹狀模型的說明何者「不」正確？</p>  <p>(A) 是預測類別變數的分類樹</p> <p>(B) 樹深為 3</p> <p>(C) 共有五個節點，弧角方框內由下往上的資訊分別是：落入此節點的樣本比例、三類樣本的比例、以及比例最高的類別標籤（若有平手狀況，則優先取排在前面的類別標籤）</p> <p>(D) 第一次切分條件為 $Petal.Length < 2.45$，滿足的樣本子集有左邊的 50 株（33.3%），此子集中全為 setosa，三類樣本比例分別 (1.000, .000, .000)</p> |
| D | <p>30. 下列何者通常「不」用來處理連續值的預測問題？</p> <p>(A) 簡單線性迴歸 (simple linear regression)</p> <p>(B) 多元迴歸分析 (multiple regression analysis)</p> <p>(C) 支援向量迴歸 (support vector regression)</p> <p>(D) 羅吉斯迴歸 (logistic regression)</p> |
| A | <p>31. 下列何種方法，訓練的速度通常最快？</p> <p>(A) k 近鄰法 (k-nearest neighbors)</p> <p>(B) 支援向量機 (support sector machine)</p> <p>(C) 決策樹 (decision tree)</p> <p>(D) 多層感知器 (multilayer perceptron)</p> |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 10 頁，共 15 頁

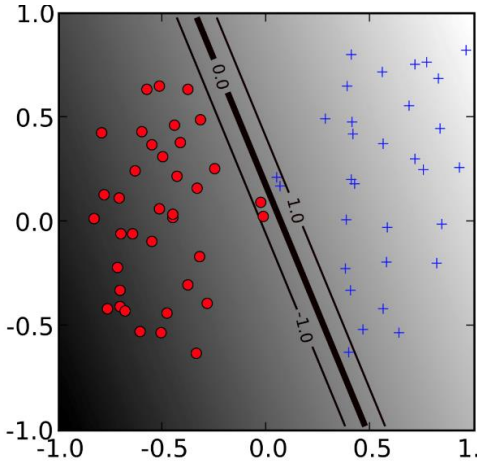
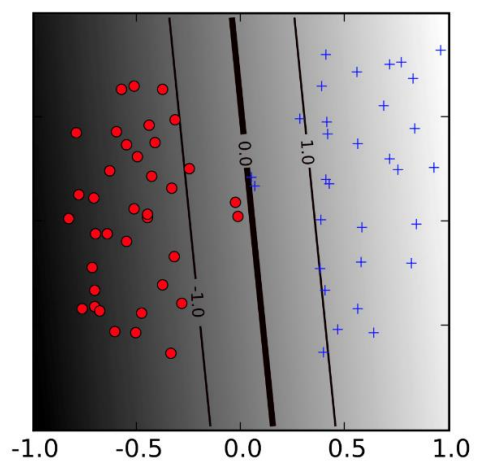
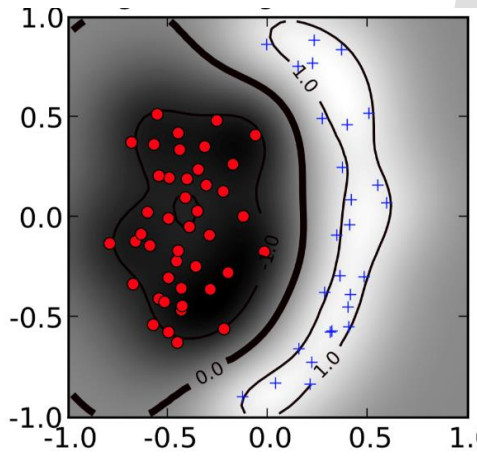
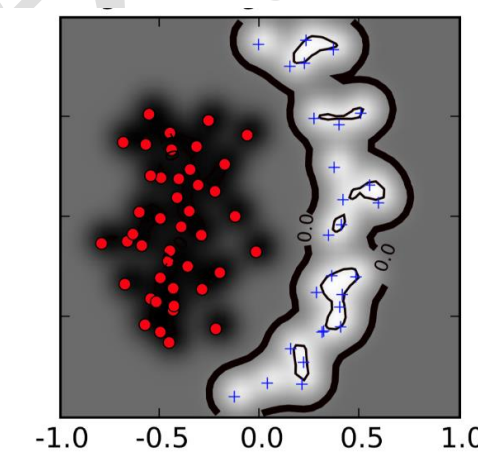
| | |
|---|--|
| C | <p>32. 下列技術之應用，何者最「不」適當？</p> <p>(A) 用卷積網路 (Convolutional Neural Networks, CNN) 辨識影像內容</p> <p>(B) 用遞歸網路 (Recurrent Neural Networks, RNN) 進行文字翻譯</p> <p>(C) 用 k 平均數演算法 (k-means) 學習多分類問題</p> <p>(D) 用自動編碼器 (autoencoder) 將資料降維</p> |
| A | <p>33. 關於類神經網路 (artificial neural networks)，下列敘述何者「不」正確？</p> <p>(A) 類神經網路至少包括投入層與隱藏層，投入層各個節點接收資料表中的自變數，隱藏層則試圖預測因變數</p> <p>(B) 除了投入層之外，其餘各層的神經元節點均需設置活化函數 (activation function)</p> <p>(C) 最簡單的類神經網路是感知機 (perceptron)，它是一種線性分類模型</p> <p>(D) 近年來一些分段線性活化函數漸受歡迎，整流線性單元 (Rectified Linear Unit, ReLU) 和硬式雙曲正切函數 (hard hyperbolic tangent function) 大範圍地取代了 S 型函數 (Sigmoid function) 與雙曲正切函數 (hyperbolic tangent function)，因為兩者更適合訓練多層神經網路</p> |
| B | <p>34. 關於 k 近鄰法 (k-nearest neighbors)，下列敘述何者正確？</p> <p>(A) 若 k=1，設得太低會導致配適不足 (underfitting)</p> <p>(B) k 取較大的值，由較多的訓練樣本共同決定待測樣本的類別，比較穩定抗雜訊</p> <p>(C) 若 k=樣本數，每個待測樣本必須跟所有訓練樣本計算距離，計算量太大</p> <p>(D) k 取奇數，可以避免鄰近樣本在不同類別的數目相等，無法判定待測樣本類別</p> |
| B | <p>35. 關於軟性邊界支援向量機 (soft-margin support vector machine)，下列敘述何者正確？</p> $\underset{\mathbf{w}, b}{\text{minimize}} \quad \frac{1}{2} \ \mathbf{w}\ ^2 + C \sum_{i=1}^n \xi_i$ <p>subject to: $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$</p> <p>公式(1)</p> <p>(公式與圖取自 Asa Ben-Hur, Jason Weston : A User's Guide to Support Vector Machines. http://pymml.sourceforge.net/svm_howto.html)</p> |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 11 頁，共 15 頁

| | |
|---|--|
| | <div style="display: flex; justify-content: space-around;">   </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <p>(a)</p> <p>(b)</p> </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;">   </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <p>(c)</p> <p>(d)</p> </div> <p>(A) 若只調整參數 C，則(a) (b)圖中，(b)圖的參數 C 比較大</p> <p>(B) 若使用 Gaussian kernel: $k(x, x') = \exp(-\gamma \ x - x'\ ^2)$ 且只調整 γ，則(c) (d)圖中，(d)圖的 γ 比較大</p> <p>(C) 在(a)圖中+1.0 與-1.0 直線間的樣本，$\xi_i=0$</p> <p>(D) 因為公式(1)中有$\ w\ ^2$，所以又稱 L2-SVM</p> |
| C | <p>36. 以拔靴集成法 (Bootsrap AGGregatING, BAGGING) 產生的 5 株裝袋樹 (bagged tree)，對某個觀測值的預測結果分別 1,0,0,0,1，請問最終預測結果為下列何者？</p> <p>(A) 1</p> <p>(B) 3</p> <p>(C) 0</p> <p>(D) 2</p> |
| A | <p>37. 假設有一訓練資料集為 $S = \{1(-), 1(-), 2(+), 2(+), 2(+), 3(-), 3(-), 5(-), 7(-), 9(-)\}$，其中數字代表特徵值，括號內的正負號代表類別。若使用拔靴</p> |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 12 頁，共 15 頁

| | |
|---|--|
| | <p>集成法 (Bootstrap AGGREGatING, BAGGING) 建立分類器時，使用的抽樣方法為下列何者？(假設抽樣的資料筆數亦為 10)</p> <p>(A) 隨機從 S 中抽取 10 筆資料，S 中的每筆資料可被抽取不只一次，例如：{5(-), 1(-), 9(-), 7(-), 9(-), 3(-), 1(-), 2(+), 5(-), 2(+)}</p> <p>(B) 隨機從 S 中抽取 10 筆資料，S 中的每筆資料至多只能抽取一次，例如：{5(-), 1(-), 9(-), 2(+), 3(-), 2(+), 7(-), 2(+), 3(-), 1(-)}</p> <p>(C) 隨機從 S 中抽取 10 筆資料，S 中的每筆資料可被抽取不只一次，但必須維持正負兩個類別的比例各半(1:1)，例如：{2(+), 1(-), 9(-), 2(+), 2(+), 3(-), 2(+), 7(-), 2(+), 3(-)}</p> <p>(D) 隨機從 S 中抽取 10 筆資料，S 中的每筆資料可被抽取不只一次，但必須維持正負兩個類別的比例和 S 中的比例相同(3:7)，例如：{2(+), 1(-), 9(-), 7(-), 9(-), 3(-), 1(-), 2(+), 5(-), 2(+)}</p> |
| D | <p>38. 俗話說：「三個臭皮匠，勝過一個諸葛亮」，恰好呼應了拔靴集成法 (Bootstrap AGGREGatING, BAGGING) 中，會產生多個弱分類器，並將這些弱分類器組合在一起，以獲得一個強分類器，有機會做出更準確的判斷。請問以下敘述何者正確？</p> <p>(A) 假如臭皮匠們沒有贏過諸葛亮，只要持續增加臭皮匠的人數，必定可以勝過諸葛亮。(亦即：假如強分類器表現不好，只要增加弱分類器的數量，必定能提升強分類器的分類準確度。)</p> <p>(B) 每個臭皮匠的意見越接近越好，因為表示他們越團結一心。(亦即：各個弱分類器的判斷結果越接近，強分類器的分類結果越準確。)</p> <p>(C) 增加臭皮匠的數量會導致意見分歧，反而無法贏過諸葛亮。(亦即：增加弱分類器數量，反而會導致強分類器的分類準確度變差。)</p> <p>(D) 即使沒有一個臭皮匠能答對所有的問題，但這些臭皮匠一起討論後，仍有可能答對所有的問題。(亦即：雖然各個弱分類器的分類準確度都不到 100%，但強分類器的分類準確度仍有可能達到 100%。)</p> |
| A | <p>39. 關於機器學習，下列敘述何者「不」正確？</p> <p>(A) 薈萃式學習 (ensemble learning) 著眼於不同分類模型的特質，及其對訓練資料中隨機噪訊的不同敏感程度，它只能集合多個不同類模型的預測結果，總結出來的預測值能準確命中標的</p> <p>(B) 薈萃式學習以重抽樣方法，產製多個基本模型 (base learner)，成為共同決策的一系列模型</p> <p>(C) 非監督式屬於問題定義尚未十分清楚時探索與知識發現的方法，其目的仍是為了建立更好的監督式學習模型</p> <p>(D) 大數據時代下我們面臨的問題益形複雜，專家學者們在非監督式</p> |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 13 頁，共 15 頁

| | |
|---|--|
| | 與監督式學習的基礎上，延伸了更多解決複雜問題的統計機器學習方式，薈萃式學習因應而生 |
| A | <p>40. 使用拔靴集成法 (Bootsrap AGGREGatING, BAGGING) 需要注意的地方，「不」包含下列何者？</p> <p>(A) 預先指定特徵進行訓練</p> <p>(B) 拔靴集成法用抽樣資料建構的模型，可能有好有壞，所以通常以取平均（或投票法）的方式，來獲得較穩定的表現</p> <p>(C) 拔靴集成法能提升機器學習演算法的穩定性和準確性，它可以減少模型的變異數 (variance) 從而避免出現過度配適 (overfitting) 的現象</p> <p>(D) 拔靴集成法可以改良比較不穩定演算法的性能，例如：類神經網路、CART</p> |
| D | <p>41. 一般說來，拔靴集成法 (Bootsrap AGGREGatING, BAGGING) 其集成模型中各株決策樹 (decision tree) 是_____；效能提升法 (boosting) 其集成模型中各株樹則是_____。</p> <p>(A) 經過修剪的 (pruned)；經過修剪的 (pruned)</p> <p>(B) 經過修剪的 (pruned)；未經修剪的 (unpruned)</p> <p>(C) 未經修剪的 (unpruned)；未經修剪的 (unpruned)</p> <p>(D) 未經修剪的 (unpruned)；經過修剪的 (pruned)</p> |
| D | <p>42. 效能提升法 (boosting) 是將弱分類器 (weak classifiers) 集合起來，轉換為強分類器 (strong classifier)。請問下列敘述何者「不」正確？</p> <p>(A) 弱分類器是指，僅比隨機亂猜好一點點的模型，例如：擲銅板</p> <p>(B) 對於訓練好的弱分類器，可按照權重進行投票</p> <p>(C) 自適應效能提升法 (Adaptive Boosting, AdaBoost) 方法，是一個具有即時調節觀測值抽樣權重的演算法</p> <p>(D) 自適應效能提升法 (Adaptive Boosting, AdaBoost) 方法，會出現有過度配適 (overfitting) 的情況</p> |
| D | <p>43. 下列何者「不是」薈萃式學習 (ensemble learning) 常用的集成技術？</p> <p>(A) 投票法 (voting)</p> <p>(B) 平均法 (averaging)</p> <p>(C) 提升法 (boosting)</p> <p>(D) 平衡法 (balance)</p> |
| A | <p>44. 關於效能提升法 (boosting)，下列敘述何者「不」正確？</p> <p>(A) 又稱為層積法 (stacking)</p> <p>(B) 前面模型預測不準的樣本，後續被抽出的機率增大，使得後面的</p> |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 14 頁，共 15 頁

| | |
|---|---|
| | <p>模型加強對這些樣本作出更準確的預測，互補合作提升整體效能</p> <p>(C) 效能提升之意是建立多個互補的弱模型 (weak learners)，將之集成後發揮團結力量大的綜效</p> <p>(D) 極端梯度多模激發法 (eXtreme Gradient BOOSTing, XGBOOST) 結合排序算法與直方圖計算最佳的屬性分割值</p> |
| D | <p>45. 關於自適應效能提升法 (Adaptive Boosting, AdaBoost) 的訓練過程，下列敘述何者「不」正確？</p> <p>(A) 如果有 M 個樣本，則每一個訓練樣本最開始時都被賦予相同的權值：$1/M$</p> <p>(B) 某個樣本點若已經被準確地分類，在建構下一個訓練集中，它的權重就被降低；若沒有被準確地分類，那麼它的權重就得到提高</p> <p>(C) 將每個訓練得到的弱分類器 (weak classifiers) 集成強分類器 (strong classifier)</p> <p>(D) 分類錯誤率低的弱分類器 (weak classifiers) 在最終分類器中占的權重較小，否則較大</p> |
| B | <p>46. 關於機器學習的建模方式，下列何者比較不會因為增加模型的複雜度而產生過度配適 (overfitting) 的現象？</p> <p>(A) 多層感知機 (multi-layer perceptron)</p> <p>(B) 隨機森林 (random forest)</p> <p>(C) 決策樹 (decision tree)</p> <p>(D) 羅吉斯迴歸 (logistic regression)</p> |
| C | <p>47. 關於隨機森林 (random forest)，下列敘述何者「不」正確？</p> <p>(A) 可用於具有遺缺值的資料</p> <p>(B) 可用來進行預測、分類</p> <p>(C) 在薈萃式學習 (ensemble learning) 裡面，隨機森林採用效能提升的方式 (boosting) 進行系集模型的建構</p> <p>(D) 提供變數重要度分數</p> |
| D | <p>48. 關於決策樹 (decision tree) 與隨機森林 (random forest) 的比較，下列敘述何者正確？</p> <p>(A) 決策樹屬於監督式學習 (supervised learning)，隨機森林屬於非監督式學習 (unsupervised learning)</p> <p>(B) 兩者皆屬於薈萃式學習 (ensemble learning)</p> <p>(C) 隨機森林的每一棵決策樹之間是有關聯的</p> <p>(D) 隨機森林能處理離散型資料，也能處理連續型資料</p> |
| C | <p>49. 自適應效能提升法 (Adaptive Boosting, AdaBoost) 與隨機森林 (random</p> |

108 年度中級巨量資料分析師能力鑑定試題

科目 1：進階機器學習

考試日期：108 年 9 月 21 日

第 15 頁，共 15 頁

| | |
|---|--|
| | <p>forest) 的關鍵差異在於？</p> <p>(A) 前者的基本模型是強模型 (strong learners)；後者則是集成數個弱模型 (weak learners)</p> <p>(B) 前者的拔靴抽樣各樣本權重相同；後者則依前面模型預測結果的良窳，對各樣本進行加權抽樣</p> <p>(C) 前者運用資料集中的全部屬性；後者只使用資料集中的部份屬性</p> <p>(D) 前者的基本模型是樹狀模型 (tree-like models)；後者則是非樹狀模型</p> |
| C | <p>50. R 語言的 rpart 套件，實現了_____算法的諸多概念？</p> <p>(A) 迭代二分樹第三代 (Iterative Dichotomiser 3, ID3) 算法</p> <p>(B) C4.5</p> <p>(C) 分類與迴歸樹 (Classification and Regression Trees, CART)</p> <p>(D) C5.0</p> |