台灣電力公司112學年度大學及研究所獎學金甄選試題

類科:電力、資訊與用戶資料運用　　節次:第一節

科目:巨量資料概論與智慧電網

注意事項	1.本試題共5頁，採雙面印刷，請注意正、背面試題。 2.僅限使用簡易型計算器（不限廠牌、型號，功能以不超出＋、－、×、÷、%、√、MR、MC、MU、M+、M-、GT、TAX+、TAX-之運算為限；其他具有文數字編輯、發聲、振動、記憶儲存、內建程式、外接插卡、通訊或類似功能之計算工具一律禁止使用）。 3.本試題為單選題共50題，每題各2分，共100分，須用2B鉛筆在專業科目答案卡畫記作答，於本試題、英文答案卡或其他紙張作答者不予計分。 4.測驗式試題均為單選題，每題選項應有4個，以(A)(B)(C)(D)標示，請就各題選項中選出最適當者為答案；各題答對得該題所配分數，答錯不倒扣；畫記多於1個選項或未作答者，該題不予計分。 5.考試結束前離場者，試題須隨答案卡繳回，俟本節考試結束後，始得至原試場索取。 6.考試時間：與英文合併一節考試，共150分鐘。

1.甲袋有2個白球8個紅球，乙袋有7個白球3個紅球，某人從甲乙兩袋中任選一袋，假設他選中甲袋的機率為 1/3，選中乙袋的機率為 2/3，接著再從選出的袋子中隨機取一球，請問此球顏色是白球的機率為何？

(A) 2/9　(B) 8/15　(C) 9/50　(D) 以上皆非

2.颱風正接近台灣，某地方首長需要決定明天是否停班停課，他設立 2 個假設：H_0：颱風會登陸、H_1：颱風不會登陸。若「該放假而不放假」，則犯下列何者型態錯誤？

(A) 型 I 錯誤　(B) 型 II 錯誤　(C) 型 III 錯誤　(D) 型 IV 錯誤

3.承第 2 題，若「不該放假而放假」，則犯下列何者型態錯誤？

(A) 型 I 錯誤　(B) 型 II 錯誤　(C) 型 III 錯誤　(D) 型 IV 錯誤

4.假設 X 為間斷隨機變數，其 E(X) = 5，Var(X)=2，試求 E($x^2 + 7x - 3$) 為何？

(A) 59　(B) 17　(C) 57　(D) 14

5.在觀察分類模型之分類結果，通常會使用下列何種矩陣？

(A) 稀疏矩陣　(B) 相似度矩陣　(C) 對角矩陣　(D) 混淆矩陣

6.欲比較兩公司員工薪資之離散程度，可採用下列何種統計量？

(A) 變異數　(B) 變異係數　(C) 全距　(D) 平均數

7.下列哪種圖表最適合用來展示資料中各類型數據所佔比例？

(A) 散點圖(Scatter plot)(B) 長條圖(Bar chart)　(C) 折線圖(Line chart)　(D) 圓餅圖(Pie chart)

8.下列哪種圖表最適合用來顯示資料隨著時間的變化趨勢？

(A) 散點圖(Scatter plot)(B) 長條圖(Bar chart)　(C) 折線圖(Line chart)　(D) 圓餅圖(Pie chart)

9.下列何者是將時間序列資料轉換到頻域空間的方法？

(A) 傅立葉轉換　(B) 特徵值加權　(C) 資料降維　(D) 隨機抽樣

10.下列何者為非結構化的資料？

(A) 戶政單位的居民資料　(B) 報稅網站中的 XML 資料

(C) 警察局收到的報案電話紀錄　(D) 市政網站的滑鼠點擊紀錄

【請翻頁繼續作答】

11.當 A、B 的共變異數(Covariance)Cov[A,B]=5 時，則 Cov[2A,B+1]=？
(A) 11 (B) 10 (C) 6 (D) 5

12.某地區住宅用戶每月用電度數平均值為 240 度，標準差為 20 度。請利用柴比雪夫定理(Chebyshev Theorem)，求至少有多少比例的住宅用戶每月用電度數會落在 200 至 280 度之間？
(A) 55% (B) 65% (C) 75% (D) 85%

13.均方根誤差(RMSE)是藉以衡量下列何者？
(A) 樣本量大小 (B) 預測的準確性 (C) 移動平均週期 (D) 指數平滑

14.下列何者非 Python 所支援的預設資料型態？
(A) list (B) map (C) string (D) char

15.在 Python 語言中，假設 var 值是 1，執行完指令 var＝'Hello'[var]+ 'var'，下列何者為 var 的結果？
(A) evar (B) li (C) Hvar (D) Hello

16.在 Python 語言中，已知「a=numpy.array([[1,3],[2,4]]」，則 3*a 意義為何？
(A) a*a*a (B) 3 乘以第 1 行元素 (C) 3 乘以第 3 行元素 (D) 3 乘以每個元素

17.考慮 R 語言運算，使用 read.table 函數匯入文字檔的資料結構為何？
(A) 資料框(Data Frame)(B) 向量(Vector) (C) 矩陣(Matrix) (D) 串列(List)

18.下列關於 NoSQL 資料庫的敘述，何者有誤？
(A) MongoDB 是一種 NoSQL 的資料庫 (B) 使用記憶體方式建立分散資料庫
(C)可採用 Key-DM 資料架構來建立資料庫 (D) 各種 NoSQL 資料庫所支援的語言可能不同

19.下列關於虛擬電廠(Virtual Power Plant)的敘述，何者有誤？
(A) 是將眾多小型、分散的發電與儲能設施，整合在同一個管理平台下
(B) 可有效調整電網的供電量以維繫電網平衡
(C) 通常由聚合商進行資源整合
(D) 需量反應不屬於虛擬電廠之範疇

20.經過網路爬蟲蒐集的網頁資料(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，此過程與下列何者較為相符？
(A) 資料擴增 (B) 資料組織 (C) 資訊分類 (D) 模型預測

21.下列關於 MapReduce 框架的敘述，何者有誤？
(A) Mapper 的輸出需要是鍵值組(key-value pair)的結構
(B) 實現 Mapper，通常是定義如何處理個別鍵值下的值集合
(C) Reducer 的輸出值通常也是鍵值組(key-value pair)的結構
(D) 資料在進入 Map 階段之前會經過整理階段(shuffle)

22.下列關於 Python 變數管理的敘述，何者有誤？
(A) 變數不須宣告資料型態 (B) 變數不須事先宣告
(C) 變數不須先建立和給值而直接使用 (D) 變數可以使用 del 釋放資源

23.下列關於巨量資料的敘述，何者有誤？
(A) 巨量資料的資料來源皆為主動產生且有規律的資料
(B) 物聯網加速巨量資料的發展
(C) 真實性(Veracity)亦為巨量資料的特點
(D) 巨量資料不適合使用關聯式資料庫

24.下列何者非屬需量反應輔助服務的優點？
(A) 達成節電目標 (B) 降低電網投資成本 (C) 引導用戶管理用電 (D) 調控再生能源發電

【請另頁繼續作答】

25.下列何者不是資料進行變異數分析時，所需之假設？

(A) 資料呈常態分配　(B) 各組母體平均數相等(C) 各組母體變異數相等 (D) 各組資料間獨立

26.下列何者非屬特徵選擇(Feature-Selection)的標準方法？

(A) 嵌入方法(Embedded)　(B) 過濾方法(Filter)

(C) 包裝方法(Wrapper)　(D) 抽樣方法(Sampling)

27.下列關於資料特徵的敘述，何者有誤？

(A) 資料特徵個數愈多，該模型所需的運算時間也就愈短

(B) 資料特徵個數愈多，容易引起維度災難，模型也會愈複雜

(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

(D) 可透過模型計算資料特徵重要程度，例如：Random Forest

28.當資料科學家建模時，下列何者為過度配適(Over-fitting)的狀況？

(A) 訓練誤差低，測試誤差低　(B) 訓練誤差低，測試誤差高

(C) 訓練誤差高，測試誤差低　(D) 訓練誤差高，測試誤差高

29.當模型發生過度配適(Over-fitting)的情形時，下列何種方法無法緩解？

(A) 蒐集更多資料　(B) 減少模型複雜度　(C) 增加模型訓練的時間(D) 使用正則化

30.機器學習模型中，下列關於模型的偏差(Bias)與變異(Variance)的敘述，何者正確？

(A) 高偏差代表模型過於複雜　(B) 高變異代表模型過於簡單

(C) 希望訓練好的模型能是高變異、低偏差　(D) 偏差與變異之間存在一平衡(Trade-off)關係

31.下列何者不是資料進行屬性轉換的主要目的？

(A) 能夠讓資料的可讀性更高

(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

(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常態分布

(D) 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

32.資料進行屬性轉換通常可以降低量綱尺度(Scale)對模型的影響。下列何種類型的模型方法，不需要做屬性轉換？

(A) k-means 集群　(B) 支援向量機　(C) 類神經網路　(D) 樹狀模型

33.下列何者不是用來處理連續值的預測問題？

(A) 簡單線性迴歸(simple linear regression)　(B) 多元迴歸分析(multiple regression analysis)

(C) 羅吉斯迴歸(logistic regression)　(D) 支援向量迴歸(support vector regression)

34.二元分類問題中，如果資料存在類別極度不平衡的問題，建立模型後在測試集達到了 99%的準確度(Accuracy)，下列敘述何者正確？

(A) 模型有足夠高的準確度，可上線運行

(B) 準確率(Accuracy)不適合用來評估二元分類問題

(C) 應使用其他指標來評估不平衡的二元分類問題

(D) 可能有過度配適(Over-fitting)的風險，應更換更簡單的模型

35.下列何者不是處理分類問題時，不同類的樣本數不平衡時的方式？

(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

36.下列關於 HTTPS 與 HTTP 差異的敘述，何者正確？
(A) HTTPS 使用了類似多執行緒的技術來建立多個連線，以加快資料交換的速度
(B) HTTPS 使用了加密技術以提升安全性
(C) HTTPS 使用網際網路協定第六版(IPv6)以解決位址枯竭的問題
(D) HTTPS 是 HTTP 之舊稱

37.下列關於機器學習的敘述，何者正確？
(A) 可以採用監督式、非監督式、半監督式或強化式共 4 種學習模式
(B) 監督式學習之演算法有羅吉斯迴歸和 K-means 等
(C) 非監督式學習可協助我們辨別出照片上的動物是貓還是狗
(D) 用人力對訓練資料做特徵標籤，嘗試錯誤的學習方法，是強化學習的特色

38.下列何者為衡量「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

39.下列關於 K-means 演算法的敘述，何者正確？
(A) 對異常值、極值的資料敏感
(B) 當集群中心不再變動，就達到全局最佳解(global optimum)
(C) 群集的界線可以是曲線和折線以達到全局最佳解
(D) 屬於不斷切割群集的一種演算法

40.下列關於隨機森林(Random Forest)建立過程的敘述，何者有誤？
(A) 因為隨機採樣的關係，就算不剪枝，也較不會出現過度配適(Over-fitting)的現象
(B) 隨機森林是對決策樹(Decision Tree)的一種改進，森林中的每棵樹具有不同的分佈
(C) 當隨機森林中的決策樹個數很多時，進行資料訓練需要的空間和時間會比較大
(D) 隨機森林能處理很高維度的資料，並且不用做特徵篩選

41.下列關於深度學習的敘述，何者有誤？
(A) 是機器學習(Machine learning)的一個分支
(B) 可應用於自然語言處理、推薦系統、生醫資訊
(C) 透過由人力撰寫的演算法產生特徵
(D) 越多層的模型效果應該不會比較少層的模型差

42.下列關於類神經網路的敘述，何者有誤？
(A) 卷積神經網路、遞歸神經網路均屬於類神經網路的一種
(B) 是一種模仿生物神經系統的數學模型
(C) 活化函數通常是一種非線性的轉換，如 Sigmoid 函數
(D) 池化層的降維採樣可以用隨機梯度下降法處理

43.下列關於支援向量機(Support Vector Machine, SVM)模型超參數(hyperparameters)的敘述，何者有誤？
(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)

44.下列關於 k 近鄰法(KNN)的敘述，何者正確？
(A) 基本運作是以樣本間的距離為基礎　(B) 是非監督式學習的一種
(C) k 值小，容易配適不足　(D) k 近鄰法需要有適配模型才能進行

【請另頁繼續作答】

45.下列關於訓練機器學習(Machine Learning)模型的敘述，何者有誤？

(A) 機器學習是實現人工智慧的其中一種方式

(B) 為資料貼標是機器學習的必要方法

(C) 資料清理、特徵萃取、特徵選擇都是重要的過程

(D) 特徵萃取(Feature Extraction)是從資料中挖出可以用的特徵

46.經常被用於分析巨量資料之統計學習(Statistical Learning)方法中，下列何者所應用之領域問題與其他不同？

(A) 隨機森林(Random Forest)　(B)主成分分析(Principal Component Analysis)

(C) 彈性網路(Elastic Net)　(D)分類迴歸樹(Classification and Regression Tree)

47.下列何者不是資料前處理的步驟？

(A) 資料清理(Cleaning)　(B) 資料操弄(Manipulation)

(C) 資料建模(Modeling)　(D) 資料變形(Reshaping)

48.在物聯網(IoT)佈建中，下列何種技術能同時滿足長距離、低耗電、低成本的要求？

(A) Wi-Fi　(B) Bluetooth　(C) Zigbee　(D) LoRa

49.當輸入資料數量越來越多時，下列何種時間複雜度之演算法會有最差速度？

(A) O(1)　(B) O(n)　(C) O(2^n)　(D) O(n^3)

50.集成式分類方法是將弱分類器(weak classifiers)集合起來，用以增強分類的準確率與穩定度，下列何者不屬集成式分類方法？

(A) AdaBoost　(B) Gradient Boosted Trees

(C) K-Nearest Neighbor　(D) Random Forest