科目：海量資料分析導論　　系所組：生物醫學海量資料分析碩士學位學程

第一部分 選擇題 (8 題，每題 5 分，共 40 分)

1. 甲數列：X1, X2, X3, X4, X5；乙數列：Y1, Y2, Y3, Y4, Y5；丙數列：Y1/2, Y2/2, Y3/2, Y4/2, Y5/2。已知：甲數列與乙數列之相關係數＝r，則甲數列與丙數列之相關係數為何？
   (A) r
   (B) r+2
   (C) r^2
   (D) 2r
2. 下列敘述何者錯誤？
   (A) 樣本平均數的值 易受極端(Outlier)影響
   (B) 相較於平均數，中位數較不易受極端值 (Outlier)影響
   (C) 只要樣本數大於 30，則樣本平均數就會服從常態分佈
   (D) 若平均數大於中位數，則該分佈較可能是屬於右偏的分佈
3. 下列方法在都適用的情形下，何者無法有效提升資料運算速度?
   (A) 使用平行運算
   (B) 優化演算法
   (C) 使用圖形處理器 GPU(Graphics Processing Unit)協助運算
   (D) 加大硬碟空間
4. 在做資料分析時，我們常需要把不同格式的項次轉成相同格式。如果一個資料裡的某個項次有 83 個不同的數值，要將該項次轉換成二進位數值，需要使用多少個二進位變數?
   (A) 4
   (B) 5
   (C) 6
   (D) 7
5. 下列何者不是海量資料的特性?
   (A) 數量大
   (B) 產生速度快
   (C) 需使用超級電腦
   (D) 可能存有誤差資料

※ 注意：1.考生須在「彌封答案卷」上作答。
2.本試題紙空白部份可當稿紙使用。
3.考生於作答時可否使用計算機、法典、字典或其他資料或工具，以簡章之規定為準。

6. 一般來說，下列何種分類器(classifier)所得出的結果較易解讀?
   (A) Support Vector Machine (SVM)
   (B) Linear Regression
   (C) Decision Tree
   (D) k-Nearest Neighbor
7. 下列何者是非監督式(unsupervised)學習法?
   (A) k-means Clustering
   (B) Decision Tree Induction
   (C) Linear Regression
   (D) Naive Bayesian
8. 下列關係何者為誤?
   (A) True negative = correctly rejected
   (B) True positive rate (TPR) = Sensitivity = Recall
   (C) False negative rate (FNR) = Specificity
   (D) False Positive = Type I error

第二部分 計算/簡答題 (6 題，每題 10 分，共 60 分)

1. 盒形圖(或盒鬚圖，Box Plot) 又稱做「五指標摘要圖」(five-number summary plot)。若要從下列的 19 個觀測值 [28 21 21 3 22 31 35 26 27 33 36 35 23 24 43 31 30 34 48] 畫出其盒形圖,請計算以下數值
   (A) 五指標綜合(five-number summary)
   (B) 四分位數間距(interquartile range，IQR)及極端值(outliers)
2. 請利用表一中 6 個點(p1~p6)的座標及表二中的距離矩陣，根據以下指定的兩種方法劃出系統樹圖(或稱樹形結構關係圖，dendrogram)。
   (A) 單一聯結法 (single-link)
   (B) 完全聯結法 (complete-link)

表一

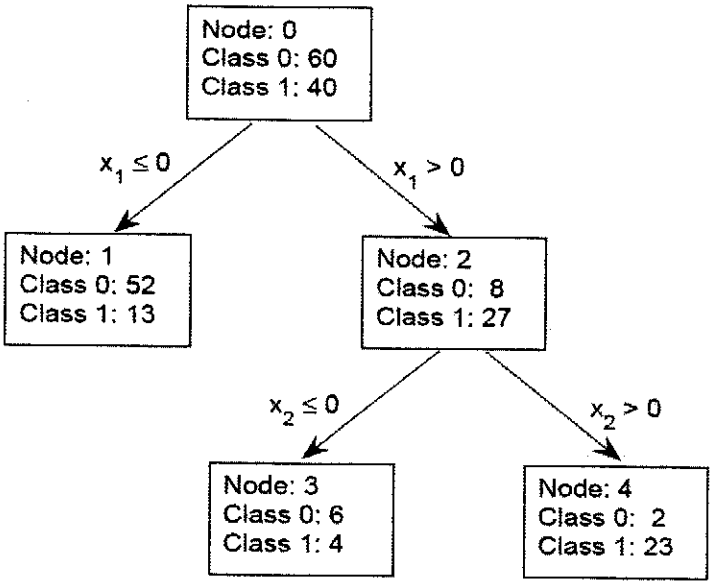
Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

表二

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

※ 注意：1.考生須在「彌封答案卷」上作答。
2.本試題紙空白部份可當稿紙使用。
3.考生於作答時可否使用計算機、法典、字典或其他資料或工具，以簡章之規定為準。

3. 下圖顯示的決策樹中，要將 100 筆資料，根據兩個屬性(x_1, x_2)，區分成類別 0(Class 0)及類別 1(Class 1)。請舉出一種指標來評估該決策樹分類結果的好壞，並根據以下決策樹計算該指標。



4. 根據世界衛生組織(WHO)的資料，從 2003 年到 2009 年這段期間，H5N1 禽流感在全球造成 468 個人類病例，其中 282 人死亡。請利用以上資訊，在 $\alpha = 0.05$ 水準下，檢定「H5N1 禽流感人類病例之死亡率至少有 2/3」的假設。(你或許會用到以下全部或部分的臨界值表：$F_{11,55,0.05} = 1.97$，$F_{3,40,0.05} = 2.84$，$Z_{0.05} = 1.645$，$Z_{0.025} = 1.96$，$Z_{0.005} = 2.57$)

5. 許多資料分析法都假設資料為常態分佈(normal distribution)。若不知資料的實際分佈形式而貿然用之，可能導致錯誤。請舉出一種你覺得最有說服力的方法來鑑定你手上的資料是否為常態分佈(鑑定是常態分佈或不是常態分佈其中一種即可)，並簡單描述其作法。

6. 請解釋下列名詞：

   (A) 海量資料(或稱大數據，Big Data)

   (B) 中央極限定理(Central Limit Theorem)

   (C) 自助抽樣法(Bootstrap Method，Bootstrapping 或靴拔重抽法)

   (D) 熵(Entropy，在資訊理論(Information Theory)中的定義及意義)

   (E) 物聯網(Internet of Things，IoT)

※ 注意：1.考生須在「彌封答案卷」上作答。
2.本試題紙空白部份可當稿紙使用。
3.考生於作答時可否使用計算機、法典、字典或其他資料或工具，以簡章之規定為準。