Statistical Learning and Data Mining, Midterm Exam 2025/10/29

1. **題目一：使用"Data1" 資料集進行迴歸分析**

   請使用"**Data1**" 資料集，其中應變數為y，自變數為$V_1, V_2, V_3, \ldots, V_9$。

   請回答下列問題：

   - 當使用具隨機性的演算法（例如隨機抽樣）時，請設定`random_state` $= 2025$ 以確保結果可重現。
   - 若所使用的模型**不具尺度不變性（scale-invariance）**，請在建模前**先對輸入的自變數進行標準化處理（standardization）**。

   (a) 計算變數間的交叉相關係數(cross-correlation)，並將相關係數矩陣以熱圖（heatmap）方式繪製出來。請描述你的觀察結果。(6%)

   (b) 使用`train_test_split` 隨機選取四分之三的觀測資料作為訓練集（training set)，其餘四分之一作為測試集（test set)。(5%)

   (c) 使用訓練集擬合一個多元線性迴歸模型。是否至少有一個自變數與y 具有統計上顯著的關聯？根據p 值，哪些變數是顯著的？(10%)

   (d) 繪製studentized 殘差（studentized residuals）對擬合值（fitted values）的散佈圖(scatter plot)。根據圖形判斷，是否存在離群值（outliers）？(5%)

   (e) 使用訓練集，透過逐步向前選取法（forward stepwise selection），根據赤池資訊量準則（AIC criterion）找出最佳線性迴歸模型。此方法選出了哪些變數？(10%)

   (f) 在訓練集上擬合一個Lasso regression 模型，並繪製九個預測變數在標準化後的迴歸係數（standardized coefficient estimates）隨著λ（或α）變化的圖形。接著，找出能使五折交叉驗證誤差（5-fold cross-validation error）最小的λ（或α）數值。在進行交叉驗證與繪圖時，請從10^4 掃描到10^{-4}，並在對數刻度下均勻取樣100 點。最後請回答是否有估計出來的迴歸係數為0？(15%)

   (g) 利用訓練集與測試集，計算(c)、(e)、(f) 三個模型在訓練資料與測試資料上的均方誤差（MSE，Mean Squared Error）。(15%)

2. **題目二：使用"Data2" 資料集進行分類分析**

請使用**"Data2"** 資料集，其中應變數為y，自變數為$V_1, V_2, \ldots, V_{20}$。

請回答下列問題：

- 請以接收者操作特徵曲線下面積（ROC AUC，Area Under the Receiver Operating Characteristic Curve）作為模型評估指標。可以參考`sklearn.metrics`模組與分類模型中的`predict_proba` 方法，從預測分數計算ROC AUC。
- 當使用具隨機性的演算法時，請設定`random_state` = 2025 以確保結果可重現。
- 若所選用的模型**不具尺度不變性（scale-invariance）**，請在建模前對輸入的自變數進行標準化處理。

(a) 解釋真陽性率（True Positive Rate）與假陽性率（False Positive Rate）的意義。什麼是ROC 曲線（ROC curve）？它是如何構建的？(9%)

(b) 將資料集等比例分割為訓練集與測試集，並對類別變數進行獨熱編碼（one-hot encoding）。(5%)

(c) 使用訓練集建立線性判別分析（LDA，Linear Discriminant Analysis）分類器，並回答訓練集與測試集上的ROC AUC 分數。(8%)

(d) 使用訓練集建立二次判別分析（QDA，Quadratic Discriminant Analysis）分類器，並回答訓練集與測試集上的ROC AUC 分數。(8%)

(e) 使用訓練資料集建立K 最近鄰分類器（KNN，K-Nearest Neighbors classifier）。透過10 折交叉驗證（10-fold cross-validation）選出最佳的鄰居數K。哪一個K 對應到最高的**準確率（accuracy score）**？請回報該模型在訓練集與測試集上的**ROC AUC** 分數。(10%)

(f) 根據題組(b) 至(e) 的結果，你會建議使用哪一個模型來進行預測？請說明對應的理由。(4%)