Machine Learning Week 6 Programming Assignment Report

111652042 顏友君

P1 (Classification using GDA)

(B) Clearly explain how the GDA model works and why it can be used for classification, in particular this data set.

模型工作原理

Gaussian Discriminant Analysis (GDA)是一種生成式模型，它假設每個類別的資料都服從具有**共享變異數矩陣**的**多變量常態分佈**。訓練目標是估計分布參數。預測時，GDA 利用貝氏定理計算後驗機率來進行分類。

適用於此資料集的原因

1. **分布特性吻合:** 有效值(陸地)與無效值(海洋)在經緯度空間上形成了兩個明顯的群集，適合用常態分佈來建模。
2. **非線性邊界處理:** 為了精確擬合出台灣**非線性的輪廓**，我在經緯度特徵中加入了二次多項式特徵，這讓 GDA 能夠學習出一個**二次決策邊界**，將陸地與海洋空間分隔。

(C) Train your model on the given dataset and report its accuracy. Be explicit about how you measure performance (e.g., accuracy on a test set, cross-validation, etc.).

性能衡量方法

我將資料劃分為 80%的**訓練集**和 20%的**測試集**，性能評估採用在測試集上的**準確率**(Accuracy)和**混淆矩陣**(Confusion Matrix)。

訓練結果

該模型在測試集上的準確率為 **88.50%**，混淆矩陣如下：

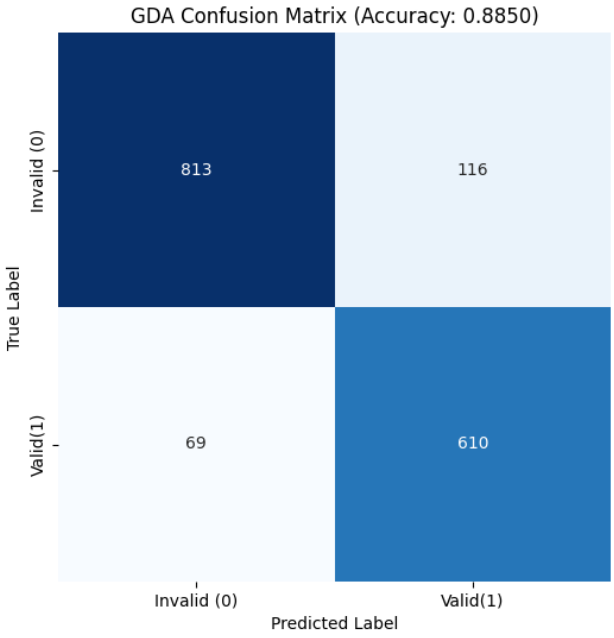

Fig.1 混淆矩陣

此外，根據該混淆矩陣，可以進一步算出 Recall、Precision、F1-score:

- **Recall** : 0.89
- **Precision** : 0.88
- **F1-score** : 0.88

(D) Plot the decision boundary of your model and include the visualization in your report.

Decision Boundary

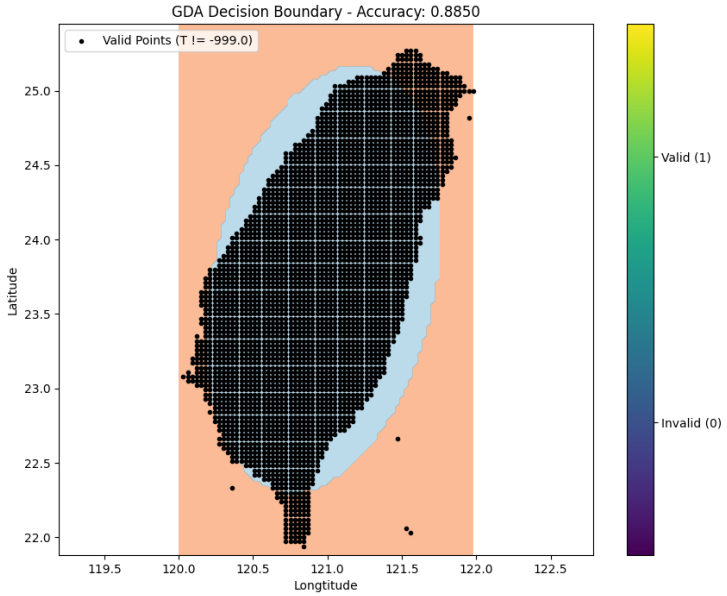

Fig.2 Decision Boundary

說明：橘色代表預測為海洋的範圍，藍色代表預測為陸地的範圍。

P2 (Regression)

模型目標

構建一個分段函數，結合分類和回歸功能，使其在陸地 (有效點) 上輸出平滑溫度預測，在海洋 (無效點) 上輸出缺失值 。

(C) Briefly explain how you built the combined function.

我定義組合模型 h(x)如下：

$$h(\vec{x}) = \begin{cases} R(\vec{x}), & \text{if } C(\vec{x}) = 1 \\ -999, & \text{if } C(\vec{x}) = 0 \end{cases}$$

- **分類模型**: 在此實作中，我延續作業四，使用 **RandomForestClassifier**，負責判斷輸入點是否為有效點 (1)。

- **回歸模型**: 使用 **RandomForestRegressor**，僅在有效點的訓練集上訓練，負責預測溫度。

此函數的運作邏輯是：先對所有輸入點進行分類預測，然後對於被分類為陸地的點(1)，套用回歸模型的預測結果，對於被分類為海洋的點(0)，則強制賦值為-999。

(B) Apply your model to the dataset and verify that the piecewise definition works as expected.

實作驗證結果

檢查項目	驗證結果	說明
預測 C(x)=0 的數量	4531	模型預測為無效點的數量
h(x)在 C(x)=0 的輸出	全為-999	成功將所有預測為無效點的輸出設為-999
預測 C(x)=1 的數量	3509	模型預測為有效點的數量
h(x)在 C(x)=1 的輸出	無-999	成功對所有預測為有效點的區域應用 R(x)預測

Table.1 實作驗證結果

(D) Include plots or tables that demonstrate the behavior of your model.

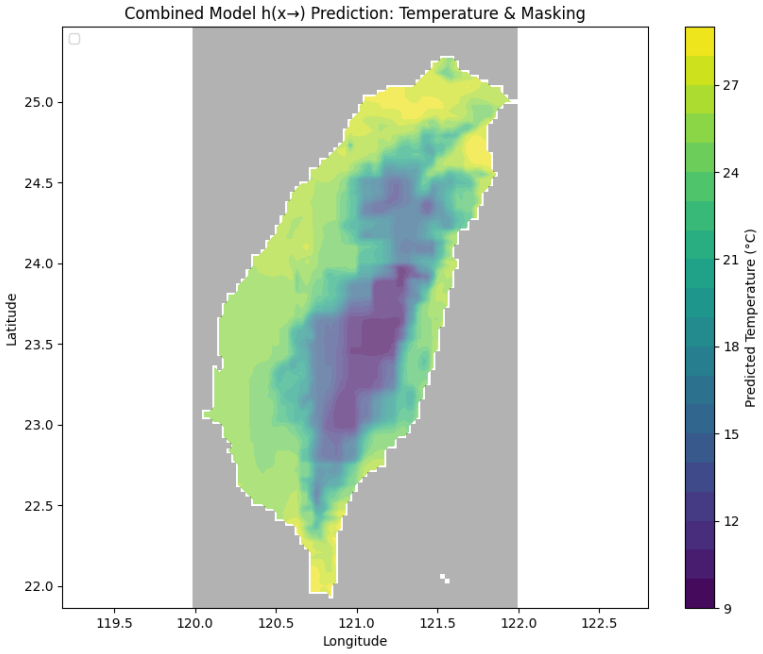

Fig.3 溫度分布圖

- **陸地區域：**顏色呈現**平滑的漸變**，這是 Random Forest 回歸模型預測出的連續溫度分佈 。
- **海洋區域：**顏色為單一的**灰色**，代表統一輸出的缺失值-999。