Week 6 assignment

1 問題一：使用 GDA 進行分類

1.1 GDA 模型原理說明

高斯判別分析（GDA）是一種生成式學習演算法（Generative Learning Algorithm）。其核心假設是不同類別的特徵資料 $P(\vec{x}|y)$ 遵循多變量高斯分佈。在本次二元分類問題中，我們假設代表海洋的類別（$y = 0$）和代表陸地的類別（$y = 1$）的資料分佈如下：

$$P(\vec{x}|y = 0) \sim \mathcal{N}(\vec{\mu}_0, \Sigma_0)$$
$$P(\vec{x}|y = 1) \sim \mathcal{N}(\vec{\mu}_1, \Sigma_1)$$

同時，類別本身的分佈 $P(y)$ 則服從伯努利分佈（Bernoulli Distribution），其中 $P(y = 1) = \phi$，$P(y = 0) = 1 - \phi$。

演算法的目標是透過最大化對數概似函數（Log-Likelihood）來從訓練資料中估計出以下參數：

- ϕ：類別為 1（陸地）的先驗機率。
- $\vec{\mu}_0, \vec{\mu}_1$：兩個類別的平均向量。
- Σ_0, Σ_1：兩個類別的共變異數矩陣。

當這兩個共變異數矩陣不相等時（$\Sigma_0 \neq \Sigma_1$），此模型稱為二次判別分析（Quadratic Discriminant Analysis, QDA），其決策邊界為二次曲線。若假設 $\Sigma_0 = \Sigma_1$，則為線性判別分析（LDA），決策邊界為直線。本次作業中，我們未做此假設，因此實作的是 QDA。

在預測階段，我們使用貝氏定理（Bayes' Theorem）來計算給定特徵 $\vec{x}$ 時的後驗機率 $P(y|\vec{x})$，並選擇機率較大的類別作為預測結果：

$$\arg\max_y P(y|\vec{x}) = \arg\max_y \frac{P(\vec{x}|y)P(y)}{P(\vec{x})} = \arg\max_y P(\vec{x}|y)P(y)$$

1.2 模型訓練與參數估計

我們將 'classification_dataset.csv' 資料集以 80% 作為訓練集，20% 作為測試集。經過模型訓練後，得到的參數估計值如下：

ϕ: 0.4347

$\vec{\mu}_0$: $\begin{pmatrix} 121.0098 \\ 23.6054 \end{pmatrix}$

$\vec{\mu}_1$: $\begin{pmatrix} 120.9719 \\ 23.7427 \end{pmatrix}$

Σ_0: $\begin{pmatrix} 0.4542 & -0.1351 \\ -0.1351 & 1.4648 \end{pmatrix}$

Σ_1: $\begin{pmatrix} 0.1792 & 0.1857 \\ 0.1857 & 0.5815 \end{pmatrix}$

1.3 模型效能評估

為了評估模型的泛化能力，我們在從未見過的測試集上進行預測。評估結果顯示，模型的準確率（Accuracy）達到了 **83.02%**。

詳細的結果如下：

- 精確率 **(Precision)**: 在所有被預測為某類別的樣本中，實際也為該類別的比例。海洋（類別 0）和陸地（類別 1）的精確率均為 0.83，表示模型預測的結果具有不錯的可信度。
- 召回率 **(Recall)**: 在所有實際為某類別的樣本中，被模型成功預測出來的比例。海洋的召回率為 0.88，陸地的召回率為 0.77。這表示模型對於識別海洋樣本的能力稍優於陸地樣本。
- **F1-score**: 精確率與召回率的調和平均數，是評估模型整體效能的綜合指標。海洋與陸地分別為 0.85、0.80。

整體而言，模型在兩個類別上都達到了均衡且良好的預測表現。

1.4 決策邊界視覺化

為了更直觀地理解 GDA 模型如何區分陸地與海洋，我們將其決策邊界繪製出來，如圖 1 所示。圖中，不同顏色的點代表原始資料的兩個類別，背景色則表示模型對該區域的預測類別。分隔兩種背景色的曲線即為模型的決策邊界。由於我們使用的是 QDA，決策邊界呈現非線性的二次曲線形式，這能更好地貼合台灣島嶼的輪廓。

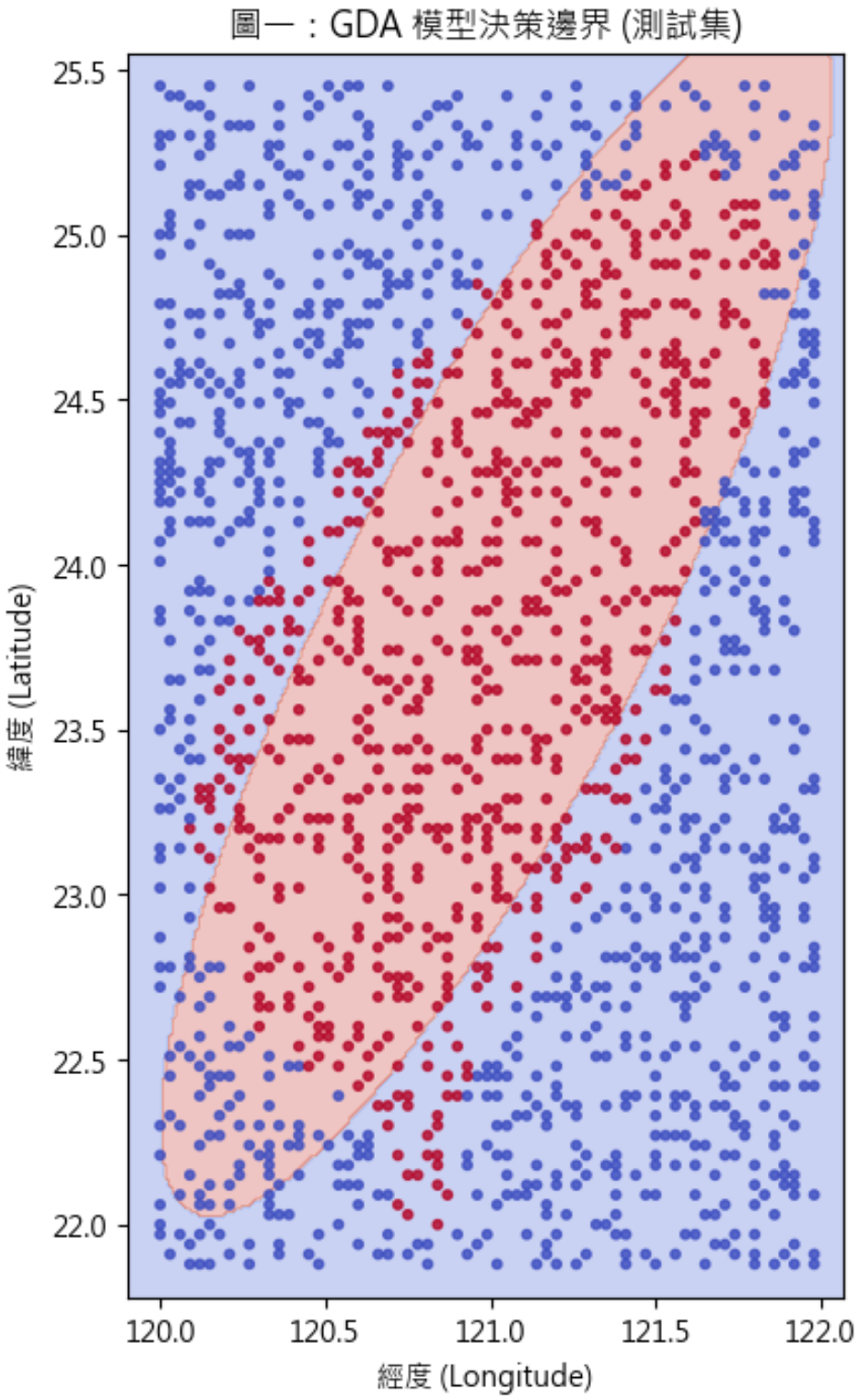

Figure 1: GDA 模型決策邊界與資料分佈圖

2 問題二：分段平滑迴歸模型

2.1 模型定義與實作

第二個問題要求我們結合分類與迴歸模型，建立一個新的函數 $h(\vec{x})$。我們將問題一的 GDA 分類模型定義為 $C(\vec{x})$，並使用第四週作業的線性迴歸模型作為 $R(\vec{x})$。此新模型的定義如下：

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

這個模型的概念是利用分類器 $C(\vec{x})$ 作為一個「 開關 」。對於任何一個經緯度座標點 $\vec{x}$，我們首先判斷它屬於陸地還是海洋。如果被判定為陸地，則使用迴歸模型 $R(\vec{x})$ 預測其溫度；如果被判定為海洋，則直接賦予無效值 -999。

在實作上，我們建立了一個 Python 函數，該函數接收一個座標點陣列，對每個點執行上述的判斷邏輯，並回傳對應的預測值。

2.2 模型應用與視覺化

我們將此組合模型 $h(\vec{x})$ 應用於整個台灣地區的網格點資料上，以生成一個完整的溫度分佈圖。視覺化結果如圖 2 所示。

圖中，彩色的部分對應 $C(\vec{x}) = 1$ 的區域，其顏色深淺代表由迴歸模型 $R(\vec{x})$ 預測的溫度高低。而淺灰色的部分則對應 $C(\vec{x}) = 0$ 的區域，表示被模型判定為海洋的無效資料區。

從圖中可以清晰地看到，溫度預測只在被 GDA 模型識別為台灣陸地的範圍內進行，而周圍的海洋區域則被成功地 mask 掉。這驗證了我們建構的分段函數模型已成功地達成了預期目標。

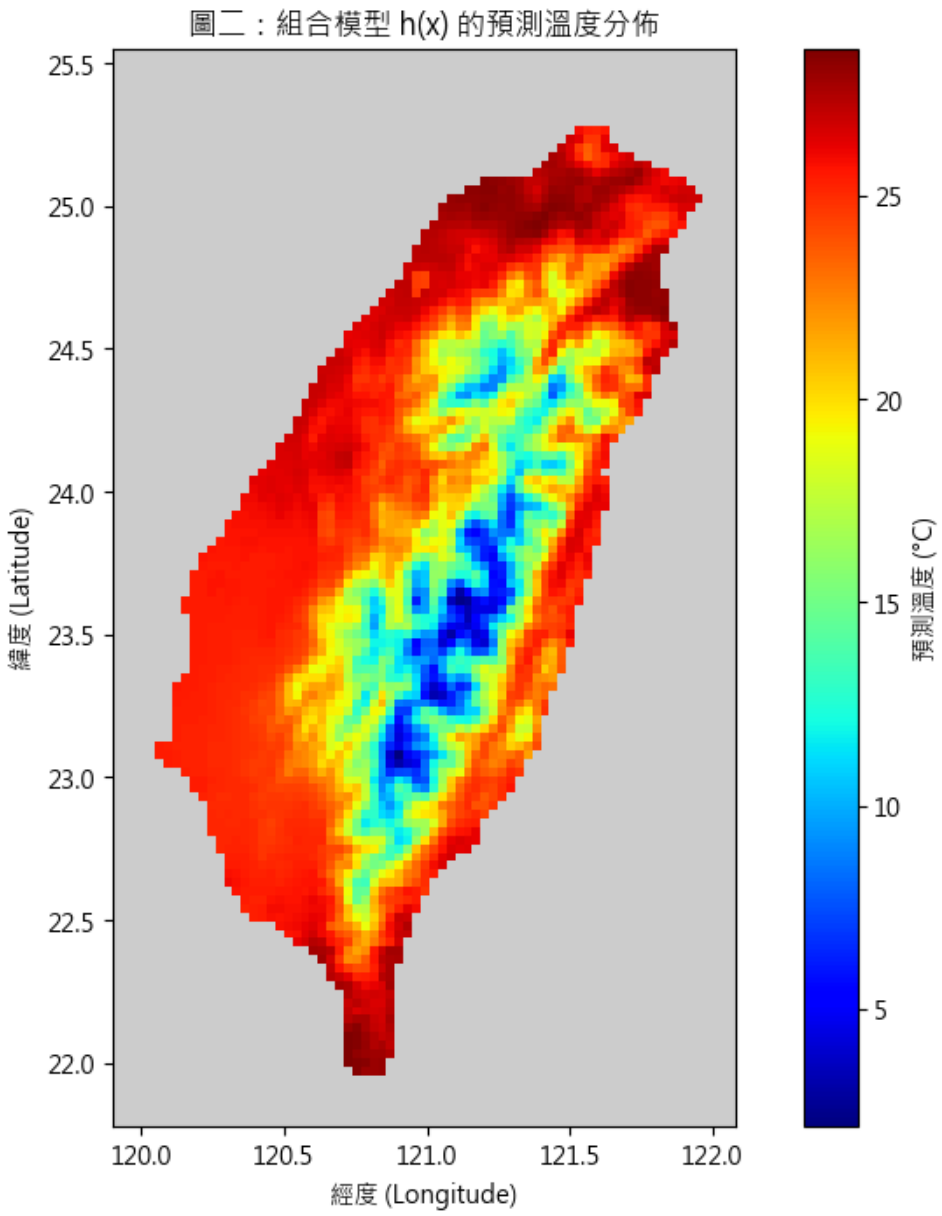

Figure 2: 分段迴歸模型預測之溫度分佈圖

3 結論

本次作業中，我們成功地從零開始實作了高斯判別分析（GDA）模型，並在台灣陸地/海洋分類任務上取得了 83.02% 的高準確率。透過視覺化決策邊界，我們驗證了 QDA 模型能有效地學習到資料的非線性分界。

接著，我們將 GDA 分類器與前期建立的線性迴歸模型相結合，創造了一個分段函數模型。此模型能夠智能地判斷地理位置，並僅針對陸地區域進行溫度預測，成功地生成了台灣地區的陸地溫度分佈圖。這項練習不僅加深了我們對 GDA 原理的理解，也展示了如何將不同的機器學習模型組合起來，以解決更複雜的實際問題。