GDA Classification 與 Regression 整合模型

資料來源：第 4 週整理之 classification/regression 資料集（經緯度座標）。

Dataset split -> Training set: 6432 samples, Test set: 1608 samples

一、Classification using GDA（高斯判別分析）

(a) 自行實作 GDA 演算法

程式中以兩類（Sea=0、Land=1）為前提，自行估計：

• 先驗機率 $\varphi = P(y = 1)$

• 各類別條件平均向量 u_0 、u_1

• 各類別條件共變異矩陣 $\sum_0$ 、$\sum_1$

並以多變量常態對數機率密度 $logN(x|\mu_k, \sum_k)$（含數值穩定化處理）計算後驗對數分數，決策函數為 $g(x) = \log(x|y = 1) + log\varphi - [log\, p(x|y = 0) + log(1 - \varphi)]$，以 $g(x) > 0$判為 Land。

(b) 為何可用於此分類情境

經緯度 (lon, lat) 的地理點在小區域內常可近似為高斯團塊分布；本題將「台灣陸地」與「周邊海域」視為兩群，以類別特有的均值與共變異矩陣捉住群集形狀與方向。當$\sum_0$ 、$\sum_1$不同時，決策邊界為二次曲線，能貼合較複雜的海陸分界；若 $\sum_0 = \sum_1$，邊界則為線性超平面。

(c) 模型訓練與效能

使用 80/20 訓練-測試切分，測試集準確率為：0.8526。

另計算混淆矩陣以檢視類別錯誤型態，並繪製決策邊界與資料點散佈作為視覺化佐證。

(d) 決策邊界視覺化

下圖示範以網格取樣計算 $g(x)$ 並以等高線/等值區塊繪出預測類別，藍色為海域、橘色為陸域。

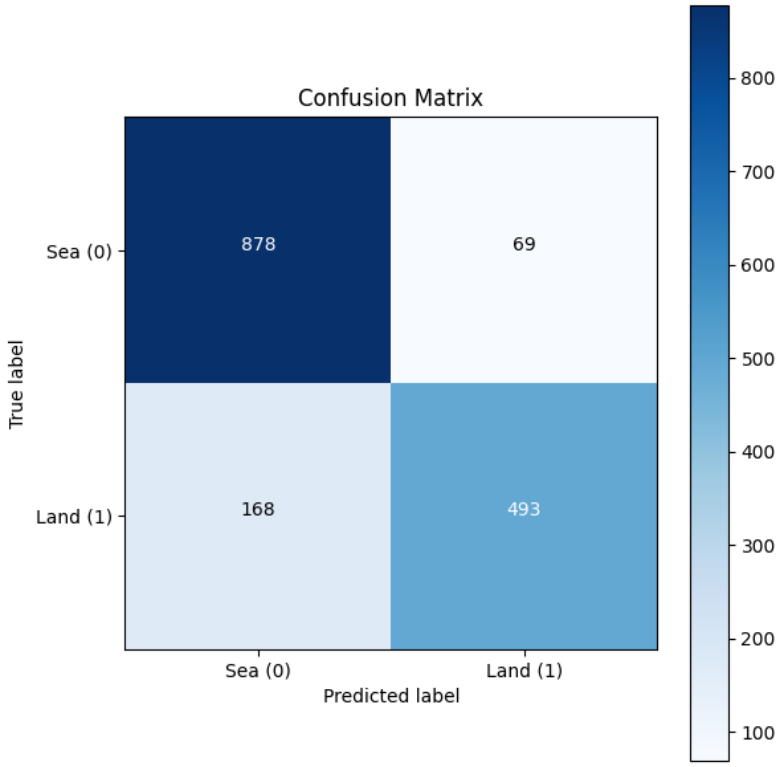

圖 1：模型視覺化（混淆矩陣）。

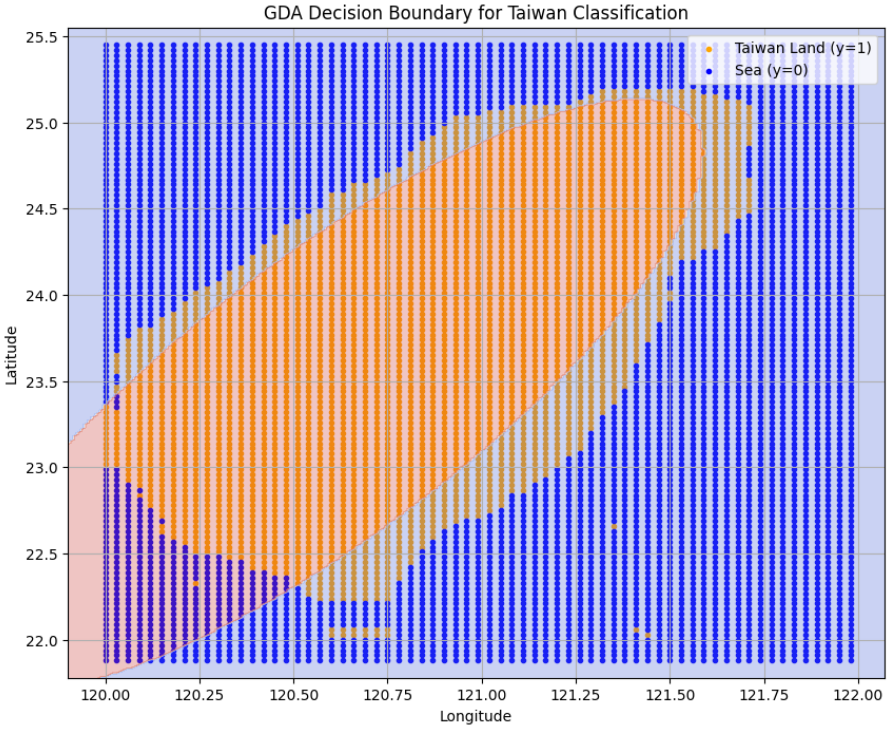

圖 2：模型視覺化（決策邊界）。

二、Regression：分段平滑函數 $h(x)$ 建模

建立迴歸模型 $R(x)$ 以第 4 次作業的回歸資料為目標值（例如溫度 value），特徵為 (lon, lat)。

程式使用 PolynomialFeatures(degree=4) 擴增特徵，並以線性迴歸擬合，得到平滑的地理變化面。

接著結合第 (一) 題之分類器 $C(x)$，定義：

$h(x) = R(x)$（若 $C(x) = 1$，屬於陸地），$h(x) = -999$h（若 $C(x) = 0$，屬於海域）。

(a) 結合模型實作

實作上先以 GDA 對輸入座標產生分類預測，再以四次多項式回歸輸出連續值；最後透過 np.where(C(x)=1, R(x), -999) 完成分段函數。

(b) 驗證分段定義

從真實海域點中挑選 5 個，經 C(x) 預測確為 0，h(x) 皆輸出 -999：

Longitude (lon)	Latitude (lat)	h(x) Predicted Output
120.12	21.88	-999.0
120.15	21.88	-999.0
120.18	21.88	-999.0
120.21	21.88	-999.0
120.24	21.88	-999.0

(c) 如何建構 $h(x)$

1) 以清理過的回歸資料訓練 $R(x)$；2) 以第 (一) 題的 GDA 訓練 $C(x)$；3) 以條件選擇（np.where）將兩者拼接。此設計可確保海域不會被迴歸模型外插成不合理數值，而陸地區域則維持連續且可微的平滑面。

(d) 視覺化/表格

下列圖例展示 $h(x)$ 於地理網格上的行為，含海陸邊界與迴歸面/散點對照。

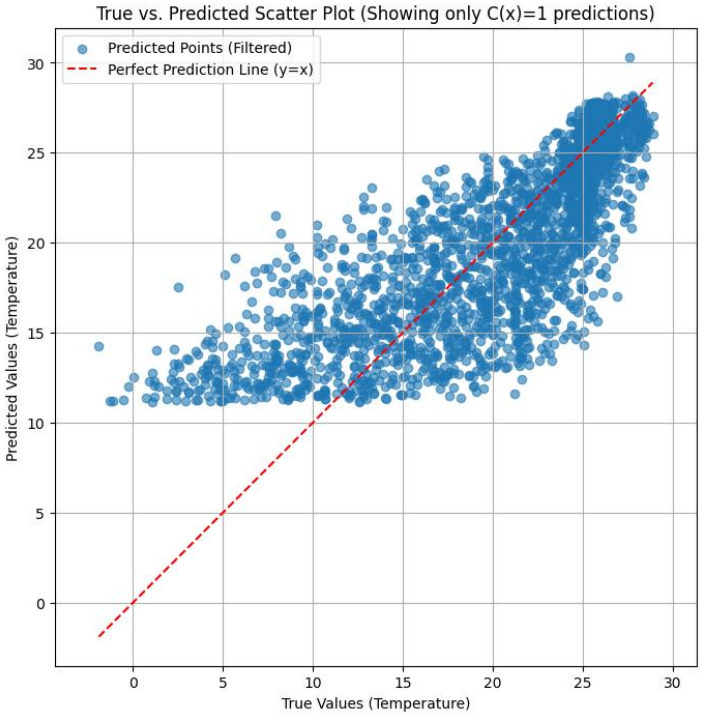

圖 3：真實 vs. 預測散佈圖（只顯示 $C(x) = 1$）。

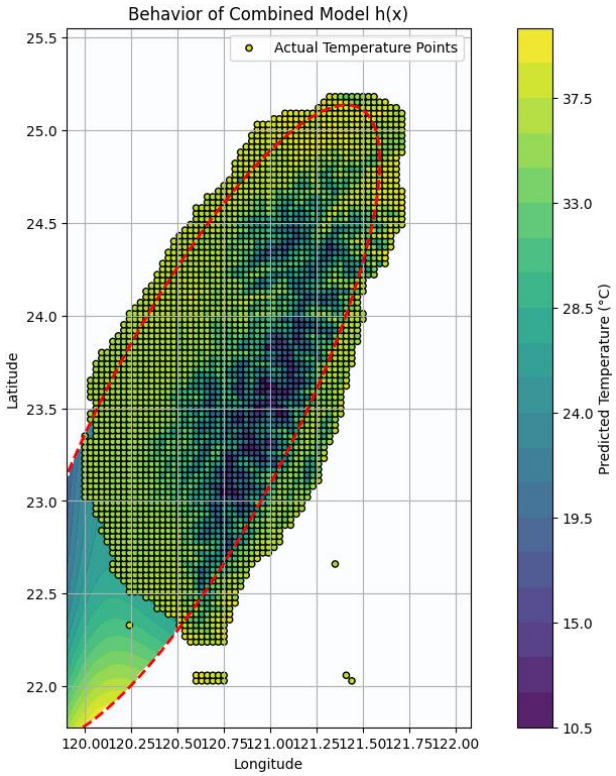

圖 4：h(x) 行為視覺化。

三、討論與結論

• GDA 能在$\sum_0 \neq \sum_1$的情況下提供二次邊界，對海陸彎曲邊緣有較好擬合能力。

• 實驗中測試集準確率約為 0.8526，代表在未見資料上仍有穩健性。

• 在海域以常數 -999 表示缺意義的回歸值，可避免以 $R(x)$ 外推至無意義區域；在陸地以四次多項式確保足夠的靈活度。若出現過擬合，可透過降低多項式階數、正則化、或交叉驗證選模。