111652015 曹晉嘉 機器學習 HW6

(有使用 gpt 修飾與打希臘字母)

1. 解釋 GDA 模型

GDA 是生成式分類法。它假設$x \in \mathbb{R}^d$的特徵在給定類別下服從高斯分佈

$$x \mid y = k \sim \mathcal{N}(\mu_k, \Sigma_k) \quad , \Pr(y = k) = \pi_k, k \in \{0,1, \dots\}$$

參數 $\{\mu_k, \Sigma_k, \pi_k\}$ 用最大概似估計對資料估計

$$\mu_k = \frac{1}{n_k} \sum_{i:y_i=k} \quad x_i$$

$$\Sigma_k = \frac{1}{n_k} \sum_{i:y_i=k} \quad (x_i - \mu_k)(x_i - \mu_k)^\top$$

$$\pi_k = n_k / n$$

決策規則（ 貝氏法則 ）

$$\hat{y}(x) = \arg max_k \quad g_k(x)$$

$$g_k(x) = \log \pi_k - \frac{1}{2} \log \mid \Sigma_k \mid - \frac{1}{2}(x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k).$$

可用來分類的原因 ：

1. 貝氏最優：在滿足 GDA 假設的情況，對 0－1 分類損失是理論上最優的。
2. 可解釋且高效：訓練與推論皆為封閉式計算，穩定快速。
3. 低維連續特徵容易觀察：在低緯度 (如 2－3 維連續資料)，Σ_k 估計可靠，決策邊界可視化直觀。

可用在本題的原因 ：

1. 可以看做是用高斯來近似實際資料
2. 台灣經緯度在平面上近似橢圓形
3. 可視化 ： 易觀察模擬的好壞程度

可改進方向 ：

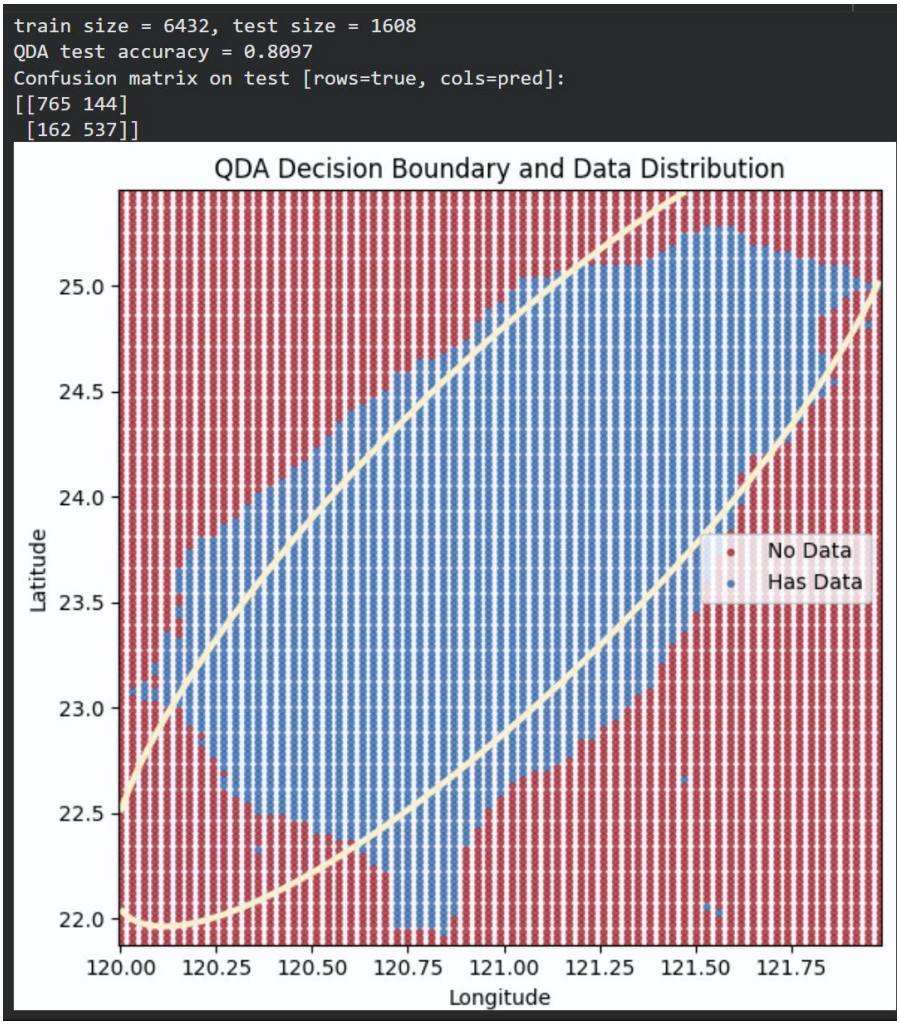
1. 改用混合高斯（GMM）每類多個分量或核密度估計，能更貼近海岸弧形

2. 訓練的結果

   準確率（ 測試集 ） ：80 . 97 %

   衡量方式 ： 將 20 % 的資料用做測試集，用以衡量模型的好壞

3. 決策邊界



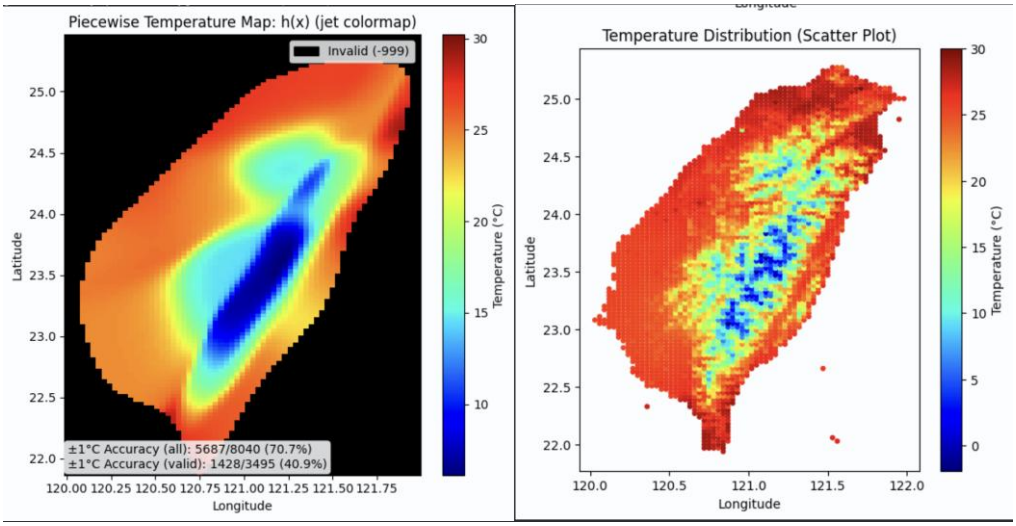
4. 如何建構這個合併函數

   先分別訓練分類模型與回歸模型（ 確保模型的準確率 ），再將它們按照題目的定義組合（ 先用分類函數判斷所有資料點，再將預測有值的資料點全部交給回歸函數預測該點的溫度 ）

5. 合併函數結果

   分類模型測試集準確率 ：96.83 %

   合併模型預測溫度與實際溫度誤差小於一度的準確率為 70.7 %



左圖為合併模型的預測值，右圖為真實資料值

6. Unanswered Questions

How badly do our class-conditional Gaussians deviate from reality, and how sensitive is QDA to this misspecification?

Do longitude/latitude need projection or scaling (e.g., local metric/UTM) to avoid distortion in distances?