111652041 蔡翔宇 HW4 的程式報告

1 我們首先先來做 Classification：

1.1 我們將格點資料的分類問題定義為一個 二元分類 (binary classification) 任務：

$$f:(\text{longlist, latitude}) \mapsto \{0,1\}$$

其中，Label 標籤$y \in \{0,1\}$：$y = 1$表示該格點溫度值有效（非 -999）。$y = 0$表示該格點為缺測或無效值（-999）。

1.2 而我們的資料來源是從中央氣象局提供的檔案 (O-A0038-003.xml)中取<Content>，解析為 $67 \times 120 = 8040$ 個浮點數。

1.3 格點重建：將解析結果 reshape 成二維陣列 (120,67)，並生成對應的經緯度網格：

$$lon_{\mathrm{i}} = lon_0 + i \cdot \Delta ln, latj = lat_0 + j \cdot \Delta lat$$

$\Delta lon = \Delta lat = 0.03$(格點網格中，每個格點的經度與緯度都是等間距分布的)

1.4 然後我們開始建立我們的 Label:

$$y_{\mathrm{ij}} \begin{cases} = 1, \text{if value} \neq \text{-999} \\ = 0, \text{else} \end{cases}$$

1.5 我們的 input : $X = \begin{bmatrix} lon \\ lat \end{bmatrix}$，我們的 output : $y \in \{0,1\}$

1.6 選擇 Logistic Regression 為我們的模型選擇

$$P(y = 1 \mid X) = \sigma(w_0 + w_1 \cdot lon + w_2 \cdot lat)$$

(1) σ是取 sigmoid function

(2) 參數 w_0, w_1, w_2 透過最大似然估計 (Maximum Likelihood Estimation) 訓練。

(3) 為避免經緯度範圍差異影響，先進行標準化

$$X' = \frac{X - \mu}{\sigma}$$

1.7 (1)準確率：

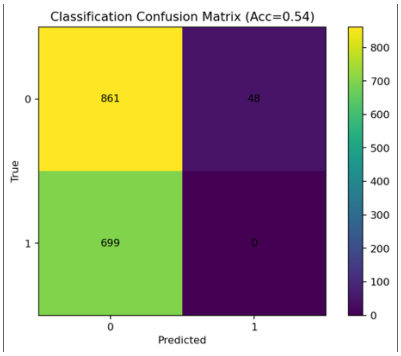
$$\mathrm{ACC} = \frac{\textit{預測成功的比數}}{\textit{總測試的筆數}}$$

(2) Confusion Matrix:

$$\begin{bmatrix} \mathrm{TN} & \mathrm{FP} \\ \mathrm{FN} & \mathrm{TP} \end{bmatrix}$$

Where, TN = True Negative, FP = False Positive, FN = False Negative, TP = True Positive。

2 我們再先來看我們最後跑出來的 Classification 的預測：



2.1 這是一個 Confusion Matrix，這是用來檢查我們分類模型在 test 上的一些表現，其中數字代表模型在各類別上的分析結果。而我們現在著重看右下角的 Block，發現真實是「有效(1)」，模型也正確預測為「有效(1)」的判斷結果遽然是 0!因此我認為我們做的分類模型失敗了，它沒有學到經緯度與是否有效之間的關聯。而這是為什麼呢？在氣象格點資料中，我們取得了很多的-999 值，但其實他並不是一個有物理意義的地理現象，而只是缺測對於海上無效的標記。

2.2 分類模型拿到的輸入只有 (longitude, latitude)，讓它試著學哪個座標是有效，哪個是無效，但如果在同一區域裡既有有效值，也有無效值，我們的模型就無法靠位置做出穩定判斷。

2.3 在我們的資料裡，無效值 (0) 比例遠大於有效值 (1)。導致了 Class Imbalance，對於 Logistic Regression 這種簡單分類器，它會傾向於預測大宗類別 (0)來最大化準確率。所以其實模型不是在「判斷是否有效」，而是在利用大宗類別取巧。

2.4 結論：準確率約為大概很可憐的只有 **54%**。模型無法學到經緯度與「是否有效」之間的真正關聯，最後就退化成幾乎都猜成 0 了。

3 我們第二個來解釋 Regression:

3.1 我們先來重新整理我們的 Data:

也就是有效值過濾：僅保留 $value_{ij} \neq -999$ 的格點作為回歸樣本：

$$\mathrm{D_{reg}} = \{(x_\mathrm{k}, t_\mathrm{k})\}_{\mathrm{k}=1}^{\mathrm{N}}, x_\mathrm{k} = [lon_\mathrm{k}, \mathrm{lat_k}]^\mathrm{T}, t_\mathrm{k} = valuek$$

3.2 我們的目標：使用特徵 $\boldsymbol{X}$：使用地理座標 $[lon, lat]$。得到目標 $\boldsymbol{y}$：對應格點的實測溫度 t（℃）。

3.3 我們使用的策略採線性回歸（Linear Regression）作為簡單基線模型，並以標準化搭配管線化訓練：

$$\hat{t} = \beta_0 + \beta_1 \cdot lon + \beta_2 \cdot lat$$

(1) 先以 StandardScaler 對 lon, lat 進行零均值、單位變異的轉換，避免量級影響

(2) 再用最小平方法估計 β 參數（scikit-learn 的 Linear Regression）

(3) 使用 train_test_split(test_size=0.2, random_state=42)：80% 訓練 / 20% 測試。在訓練集上擬合參數，在未看過的測試集上評估泛化能力。

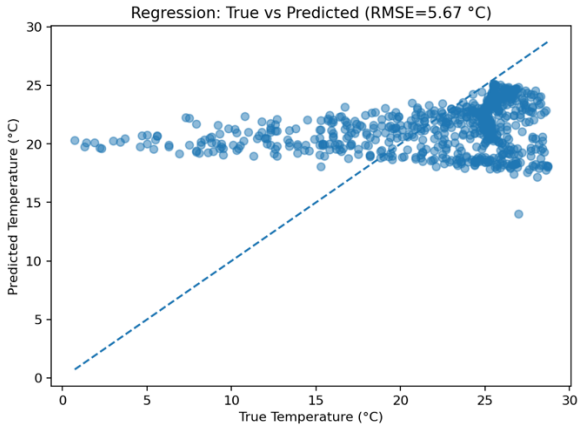
3.4 我們的評估指標使用 RMSE（Root Mean Squared Error）

(1) RMSE 的單位為 °C，數值越小越好，代表平均預測誤差越低。

(2) RMSE = $\sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{t}_i - t_i)^2}$

3.5 使用 True vs Predicted 散點圖看我們的預測模型

4 我們再先來看我們最後跑出來的 Regression 的預測：



4.1 問題一：誤差大 RMSE 偏高，我們得到的 RMSE 大約 **5–6 °C**，代表模型平均預測會差到 5 度以上，在氣象應用裡，這樣的誤差很大，表示模型只能抓到「大方向」而非「精準溫度」。

4.2 問題二：輸入特徵太少！模型只用了經度、緯度兩個數字，但氣溫受到很多很多的因素影響：地形 (高山、平地)、海洋、城市熱島效應、等等

4.3 問題三：線性假設過於簡單，模型假設溫度和經緯度的關係是線性的。然而實際上，氣溫分布通常是非線性的，所以在圖上看到：點雲雖然跟趨勢有關，但在高溫區明顯偏離，而我們的線性模型無法捕捉這些非線性模式。

4.4 結論：所以這個模型只能作為基準模型 (baseline)，未來需要更複雜的模型或更多特徵來改善我們所需要的預測。

(＃以上許多不理解的做法為詢問 ChatGPT，但報告都是自己理解完打出來的)