

Programming Assignment 4

Student: 張翰東 / ID: 313513098

1. 資料處理 (Data Preprocessing)

原始資料來源：

- 檔案名稱：O-A0038-003.xml
- 格點大小： $67 \times 120 \rightarrow$ 共 8040 筆資料
- 格點代表意義：每個 (經度, 緯度) 上對應一個溫度值 ($^{\circ}\text{C}$)
- 無效值 (-999.0)：代表該位置沒有觀測數據 (例如海上或測站缺值)

轉換結果

1. 分類資料集 (Classification dataset)

- 格式：`(lon, lat, label)`
- 規則：
 - 溫度值 = -999 \rightarrow `label = 0`
 - 否則 `label = 1`
- 大小：8040 筆

2. 回歸資料集 (Regression dataset)

- 格式：`(lon, lat, value)`
- 規則：
 - 僅保留有效值
 - `value` = 對應的攝氏溫度
- 大小：3495 筆

2. 分類模型 (Classification Model)

模型選擇

- Logistic Regression (baseline)
- Random Forest Classifier (改進後)

訓練流程

1. 對 (lon, lat) 進行標準化 (StandardScaler)
2. 分訓練集與測試集 (80% / 20%)
3. Logistic Regression 與 Random Forest 分別訓練並比較

訓練結果

- Logistic Regression baseline : 準確率 57%
- Random Forest : 準確率提升至 90–97%

為什麼差這麼多？

1. 模型能力不同
 - Logistic Regression 是 線性分類器，它只能劃一條「直線」來分隔資料。
 - 但「台灣陸地 vs 海上」的分界是一個非常 複雜、不規則的曲線，線性模型無法正確捕捉。
 - Random Forest 由多棵決策樹組成，可以劃出許多非線性的分界，因此能精準分辨海陸邊界。
2. 資料特徵單一
 - 輸入只有 (lon, lat)，沒有額外資訊（例如地形或測站密度）。
 - Logistic Regression 只能學到「大概的趨勢」，因此準確率不高。
 - Random Forest 能自動切割特徵空間，學出更多細節，所以效果明顯更好。
3. 非線性 vs 線性
 - Logistic Regression → 嘗試用「直線」逼近「台灣複雜的海岸

線」。

- Random Forest → 可以用「拼圖狀的多段曲線」去逼近，貼合度自然更高。

總結

- Logistic Regression 在這個任務中只是 baseline，比較弱。
 - Random Forest 能有效處理非線性問題，因此準確率大幅提升。
-

3. 回歸模型 (Regression Model)

模型選擇

- Random Forest Regressor
- Input : (lon, lat)
- Output : 溫度值 ($^{\circ}\text{C}$)

訓練結果

- MSE (Mean Squared Error) : 8.2
- RMSE (Root Mean Squared Error) : 2.9°C
- R² (決定係數) : 約 0.9

討論

- 平均誤差約 $\pm 3^{\circ}\text{C}$ ，效果合理。
 - 但因為溫度還會受到海拔、季節、地形等影響，所以只用經緯度做特徵仍有限制。
-

4. 結果與討論 (Results & Discussion)

- 分類模型：
 - Logistic Regression baseline → 準確率僅 57%
 - Random Forest → 準確率提升至 95% 左右

- 原因：Random Forest 能處理非線性，Logistic Regression 只能劃直線。
 - 回歸模型：
 - 預測平均誤差 $RMSE \approx 2.9^{\circ}C$
 - 能學到經緯度與溫度的分布關係，但還有改進空間。
-

5. 結論 (Conclusion)

- 成功完成資料轉換，建立分類與回歸模型。
- Logistic Regression 作為 baseline，準確率偏低；Random Forest 在分類上有明顯優勢。
- Random Forest Regressor 預測誤差約 $\pm 3^{\circ}C$ ，具有一定解釋力。
- 改進方向：加入更多氣象特徵（如海拔、季節因子），或採用更強大的深度學習模型。