ProgrammingAssignmentReport_111652026劉馥瑞

摘要

本報告旨在分析中央氣象署提供的格點溫度資料，並建立機器學習模型來處理數據完整性（分類問題）與溫度預測（回歸問題）。資料分析結果顯示，在 8040 個格點中，約 56.5% 為無效值。分類模型採用 **Logistic Regression**，但由於類別高度不平衡，模型性能極差。回歸模型採用 **Linear Regression**，雖然捕獲了地理梯度效應，但 RMSE 達 5.85◦C，顯示模型過於簡化。

1.程式碼解釋

1.1 資料設定與 XML 解析 (`parse_xml_and_extract_data`)

- 參數設定區：定義了檔案名稱 (`O-A0038-003.xml`)、XML 命名空間 (NAMESPACE)、網格尺寸 (經向 67, 緯向 120)、解析度 (0.03) 和起始經緯度 (120.00, 21.88)。
- **XML** 解析核心：使用 `xml.etree.ElementTree` 解析 XML 文件。
- 資料提取：程式碼修正了之前可能遇到的問題，使用正規表達式 (`re.findall`) 匹配所有浮點數 (包括科學記號格式如 `-999.0E+00`)，將所有資料點提取為一維陣列。
- 重塑網格：將一維資料重塑成 120×67 的二維 NumPy 陣列，代表整個地理網格的溫度數據。

1.2 資料集建立 (`create_datasets`)

此函數根據解析後的網格資料生成兩個 Pandas 資料框：

- 地理特徵生成：根據設定的起始經緯度和解析度，為網格中的每個點計算其精確的 (Longitude, Latitude) 座標。
- 分類資料集 **(Classification Data)**：所有 8040 個格點都包含在內。
  - 目標變數 **(Label):** 如果溫度值為 −999.0，則標記為 **Invalid (0)**；否則標記為 **Valid (1)**。
- 回歸資料集 **(Regression Data)**：僅包含 3495 個有效值的格點。
  - 目標變數 **(Value):** 實際的攝氏溫度值。

1.3 模型訓練與評估 (`train_and_evaluate_models`)

- 訓練/測試集分割：兩個資料集都使用 70% 訓練/30% 測試的比例分割。分類資料集使用了 stratify 參數來確保訓練集和測試集中的類別比例一致。
- 分類模型 **(Logistic Regression)**：
  - 模型：**Logistic Regression**（邏輯迴歸，用於預測 0/1）。
  - 特徵：(Longitude, Latitude)。
  - 評估：輸出 Classification Report（包含 Precision、Recall、F1-score 和 Accuracy）。

- 回歸模型 **(Linear Regression)**:
    - 模型: **Linear Regression**(線性迴歸, 用於數值預測)。
    - 特徵: (Longitude, Latitude)。
    - 評估: 輸出模型係數 (Intercept, Coefficients)、均方誤差 **(MSE)** 和 均方根誤差 **(RMSE)**。

2. 任務與資料轉換

2.1 任務目標

本次分為兩個主要目標:

1. 資料完整性預測(分類): 根據格點的地理位置 (Longitude, Latitude), 預測其數據是否為有效值(1)或無效值(0)。
2. 溫度數值預測(回歸): 對於有效數據點, 根據其地理位置, 預測實際的攝氏溫度值。

2.2 資料集建立結果

原始資料集來自 CWA 的 XML 檔案, 包含 67×120=8040 個格點。

項目	數值	說明
總格點數	8040	涵蓋經度 120.00◦ 至 121.98◦; 緯度 21.88◦ 至 25.45◦。
分類資料集筆數	8040	輸入特徵: (Longitude, Latitude)。目標變數: (Invalid=0 / Valid=1)。
回歸資料集筆數	3495	輸入特徵: (Longitude, Latitude)。目標變數: 溫度值 (Value); 僅包含有效數據。
無效值比例	4545 (~56.5%)	數據存在嚴重的類別不平衡現象, 無效值(主要為海域格點)為多數類。

3. 分類模型訓練與分析

3.1 模型與訓練過程

- 模型：邏輯迴歸 (Logistic Regression)。
- 訓練/測試集：70%/30% 分割，使用 stratify 確保類別比例分配。
- 目標：預測格點數據是否為有效值（1）。

3.2 評估結果分析

模型評估報告如下：

類別	**Precision (**精確度**)**	**Recall (**召回率**)**	**F1-score**	**Support (**樣本數**)**
Invalid (0)	0.57	1.00	0.72	1363
Valid (1)	0.00	0.00	0.00	1049
總體 **Accuracy**				0.57
Weighted Avg	0.32	0.57	0.41	2412

結果解讀：

1. 極度偏向多數類：Valid (1) 類別的 Recall 為 0.00，這表示 Logistic Regression 模型未能成功識別測試集中的任何一個有效溫度格點。
2. 模型策略失敗：由於訓練資料中 Invalid (0) 佔多數 (~56.5%)，簡單的 Logistic Regression 模型選擇了最「安全」的策略：將所有樣本都預測為 Invalid (0)。
3. **Accuracy** 具有誤導性：雖然 Accuracy 達到 0.57，但這僅是因為多數類佔 56.5%(24121363≈0.565)。模型透過盲猜多數類別來達到這個準確度，在實際應用中完全不可用。

4. 回歸模型訓練與分析

4.1 模型與訓練過程

- 模型：線性迴歸 (Linear Regression)。
- 訓練/測試集：70%/30% 分割。
- 目標：預測有效格點的溫度值 (Value)

4.2 評估結果分析

模型評估結果如下:

指標	數值	單位	說明
Longitude Coef (經度係數)	−4.7341	◦C/degree	經度每增加 1◦, 溫度下降 4.73◦C。
Latitude Coef (緯度係數)	−2.6565	◦C/degree	緯度每增加 1◦, 溫度下降 2.66◦C。
均方根誤差 (RMSE)	5.8492	◦C	模型的平均預測誤差。
測試集實際溫度平均	21.71	◦C	
測試集預測溫度平均	21.40	◦C	

結果分析:

1. 地理梯度效應: 係數皆為負值, 這符合基本的地理學原理, 即溫度隨經度向東(接近海岸與山區)和緯度向北而降低。經度對溫度的影響(斜率 |−4.73|) 比緯度更強烈。
2. 模型精度不足: RMSE 達到 5.85◦C, 這對於天氣預測任務而言是非常大的誤差。這表示僅使用經度和緯度作為線性特徵, 無法捕捉影響溫度的關鍵因素(如海拔高度、與海岸線的距離、地形遮蔽效應等)。

註:本次程式及報告有使用chatgpt作為輔助工具