1141ML-Week 4 Programming Assignment 01

Taiwan Weather Dataset Prediction

313652018 王宣瑋

September 2025

Introduction

此測試爲以氣象資料觀測平台之" 溫度分布-小時溫度觀測分析格點資料" 作爲機器學習之模型測試。測試分兩類：島內資料預估，以及經位度與溫度回歸預測。

1. 分類原始 dataset

原始資料集爲.xml 檔案，將其利用 pandas 的套件轉換成.csv 檔以後較有利於資料操作。

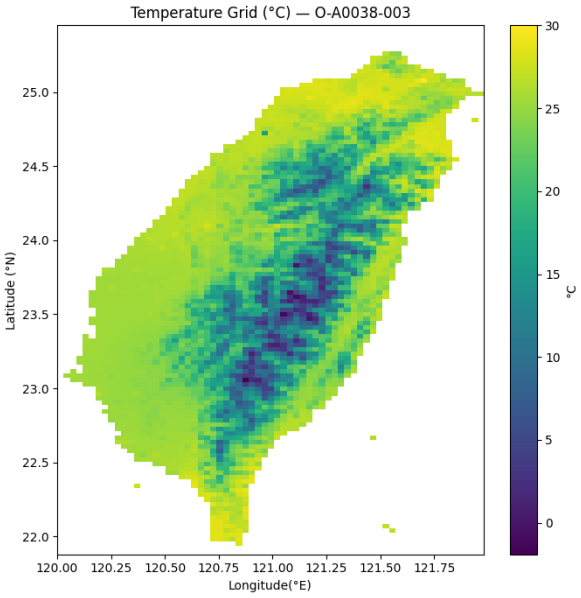

Figure 1: 以.xml 檔案來畫熱圖，得到類似於台灣地圖的分布

2.(a) 分類模型訓練

資料處理與目標

先將資料貼 label：

- 格式：(精度, 緯度, label)
- 規則：
  - 若溫度觀測值爲無效值 −999，則 label = 0。
  - 若溫度觀測值爲有效值，則 label = 1。

目標：訓練一機器模型**以 (精度, 緯度) 預測格點資料是否爲有效值 (0 或 1)**

使用方法

先使用 Logistic regression 再使用 random forest。

邏輯斯迴歸 (Logistic regression) 是一種線性模型，假設特徵與機率之間存在線性關係。模型簡單、可解釋性高，但在非線性或複雜資料上表現有限。

而隨機森林 (Random forest) 由多棵決策樹組成，透過隨機取樣與投票提升泛化能力。能捕捉非線性與特徵交互作用，表現通常比單純線性模型更強。

數據結果及 visualization

```
=== Claddification Model => Logistic Regression ===
              precision    recall  f1-score   support

           0      0.565     1.000     0.722       909
           1      0.000     0.000     0.000       699

    accuracy                         0.565      1608
   macro avg      0.283     0.500     0.361      1608
weighted avg      0.320     0.565     0.408      1608
```

(a) Data result of Logistic regression.

```
=== Claddification Model => Random Forest ===
              precision    recall  f1-score   support

           0      0.990     0.991     0.991       909
           1      0.989     0.987     0.988       699

    accuracy                         0.989      1608
   macro avg      0.989     0.989     0.989      1608
weighted avg      0.989     0.989     0.989      1608

混淆矩陣：
 [[901   8]
 [  9 690]]
```

(b) Data result of random forest

Figure 2: Logistic Regression 是線性模型，它只能畫出一條直線（或一個超平面）來切分資料。結果模型學不出好的 decision boundary，乾脆把所有點都歸類成「0」(無效)，導致 recall=0、precision=0 對 class 1 完全失敗；相反地，Random Forest 適合處理非線性、區域性強的分布，所以表現優秀，accuracy 都能在 98.9%。其中混淆矩陣的意思是表示模型分類結果中，眞實標籤與預測標籤的對照情形，清楚顯示正確與錯誤分類的數量。這裡代表僅有 17 筆的分類錯誤資料

「經緯度 vs 有效/無效」的分佈其實很複雜（有效值分布在島內，無效值分布在海上），不是單純的線性邊界，所以想當然利用線性方法解釋，效果奇差無比。所以合理推測：在地理座標類問題中，非線性模型（Random Forest、KNN、SVM）會比單純線性模型更合適。

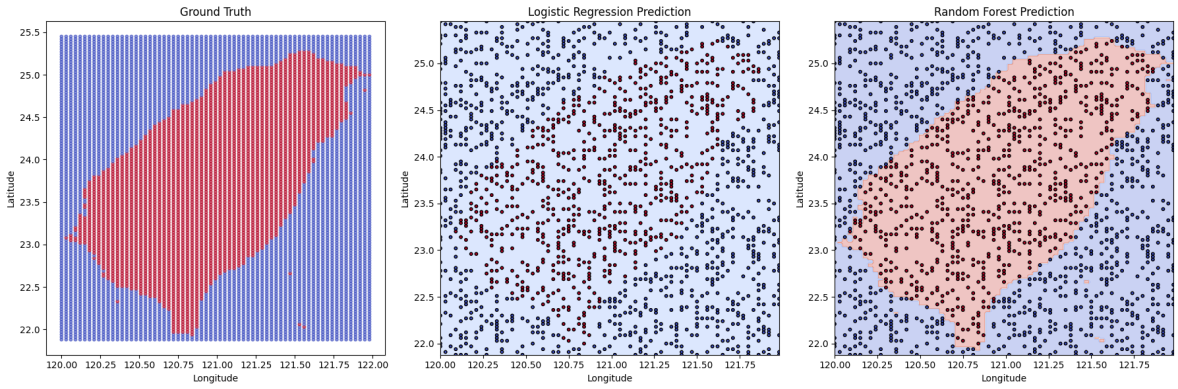

Figure 3: 從左到右分別爲：眞實資料分布圖 (ground truth)、Logistic Regression Prediction，以及 Random Forest Prediction。可以看見中間 logistic 回歸似乎都隨機判別點的 label，甚至都當成「0」，所以台灣的輪廓並不明顯。相對地，random forest 分類邊界效果尤佳，透過多棵決策樹的投票，能夠捕捉複雜的非線性邊界，分類幾乎有效。

Logistic Regression 無法學到台灣島的「不規則輪廓」，因此表現失敗。而 Random Forest 能捕捉到非線性邊界，準確率接近 99%。所以上圖結果顯示：非線性模型更適合地理空間分類問題，因爲資料的分布通常不會是單純線性可分的。

2.(b) 溫度預測模型訓練

資料處理與目標

先將資料與溫度值 (Value) 整理：

- 格式：(精度, 緯度, value)
- 規則：
  - 僅保留有效的溫度觀測值（剔除所有 -999.）。
  - value 爲對應的攝氏溫度。

目標：訓練一機器模型**以 (精度, 緯度) 預測對應的溫度觀測值。**

訓練方法

這裡也會先用 Linear regression 跟 random forest 的方法。不同的是，這次加入了 RNN 以及 SVR 方法。

循環神經網路 RNN (Recurrent Neural Network) 能處理序列資料，透過隱藏狀態記錄前後文資訊。常用於語音 (speech)、文字 (text)、時間序列 (time series) 等任務。

而支持向量迴歸 SVR (Support Vector Regression) 是一種利用超平面逼近資料，並允許一定誤差範圍。適合處理高維特徵下的迴歸問題，對異常值 (outliers) 有一定的穩健性。

數據結果及 visualization

```
Linear        MAE=4.399  RMSE=5.669
RandomForest  MAE=1.517  RMSE=2.281
KNN           MAE=1.463  RMSE=2.178
SVR(RBF)      MAE=2.170  RMSE=3.244
```

Figure 4: 四個方法分別的 MAE 跟 RMASE 之值。

在這次的實驗中，目標值 `value` 的範圍約落在 0 到 30 之間，平均值約爲 21.6。因此，可以將各模型的誤差與此範圍相比較。

- Linear 的 MAE 爲 4.399，RMSE 爲 5.669，誤差相對偏大，說明線性模型不足以捕捉資料特徵
- Random Forest 與 K 最近鄰 (KNN) 的表現最佳，MAE 約爲 1.5，相當於整體範圍的 5%，預測相當精準
- 而 SVR 則有 MAE = 2.170、RMSE = 3.244，雖稍遜於前兩者，但整體仍在可接受範圍內

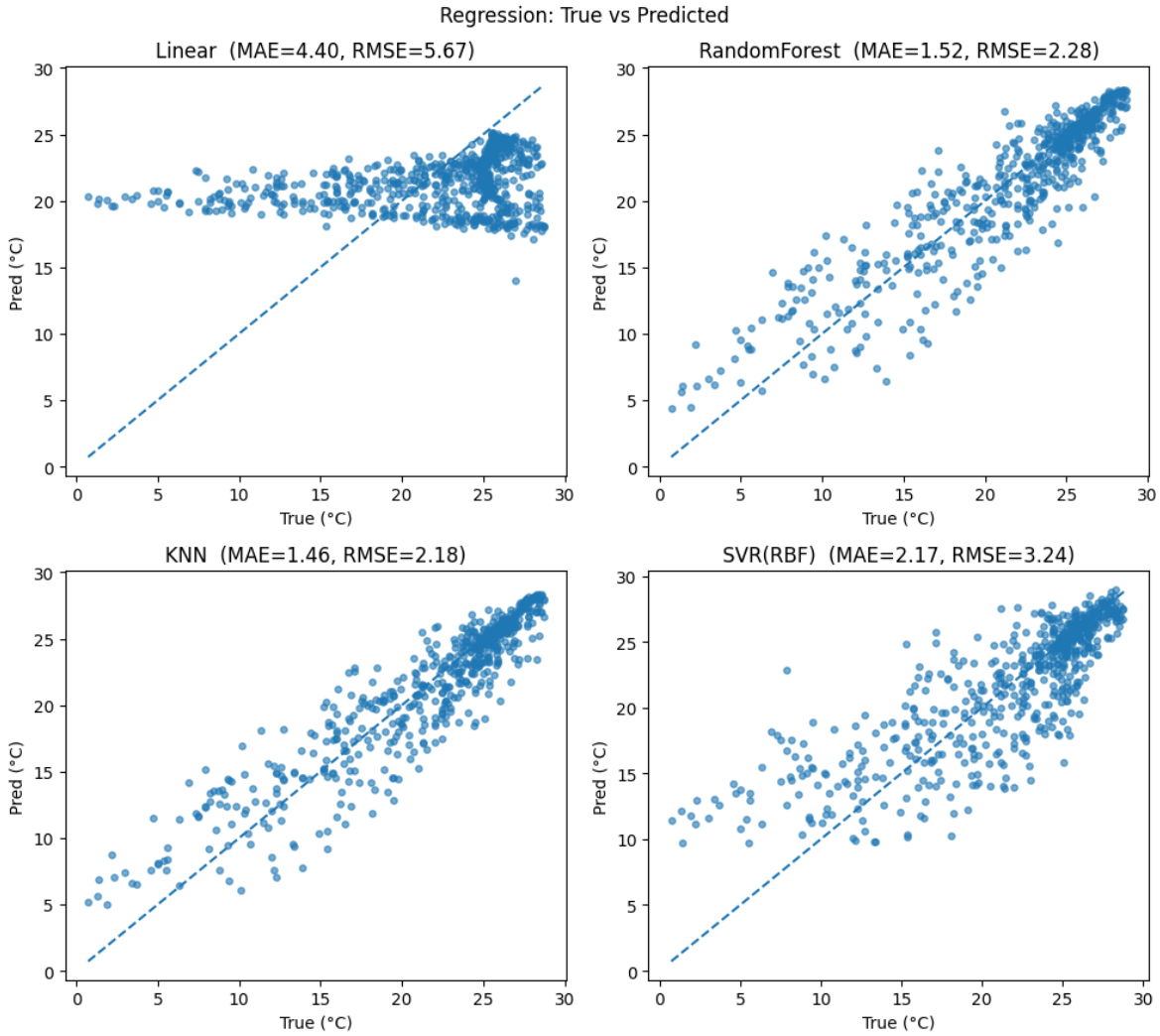

Figure 5: 這是眞實資料與預測資料的散點圖。意思是點 x 座標爲眞實值；y 座標爲預測值。如果回歸得好，那散點越接近圖中的虛線 (斜率爲 1)。所以可以看見 Linear 幾乎只給一個接近平均的水平線，無法區分高低溫區。RandomForest 以及 KNN 的點雲接近對角線，說明預測與眞實值吻合。SVR(RBF) 介於兩者之間，有一定擬合但仍有偏差。

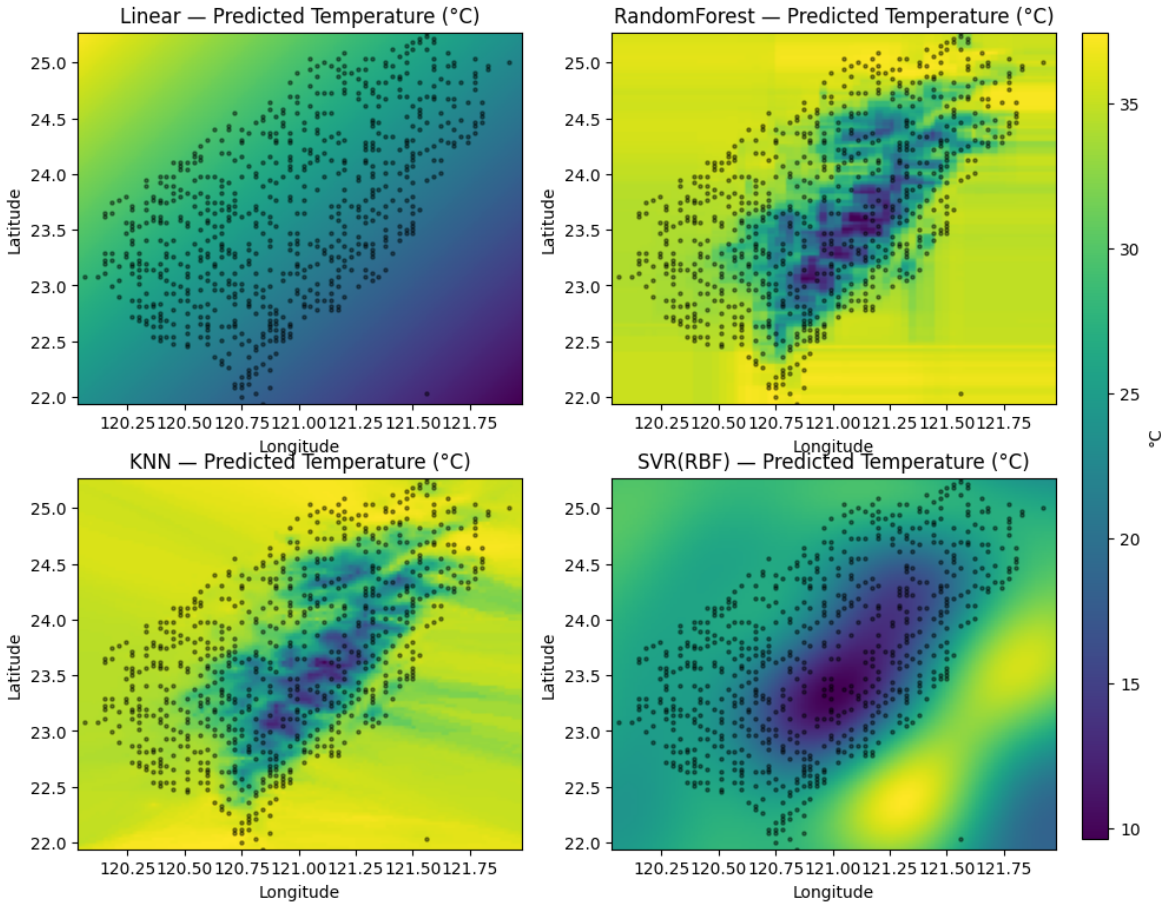

Figure 6: 四種迴歸模型的溫度預測分布。背景色彩表示模型預測的溫度場（單位：°C），黑點爲測試樣本的位置。線性迴歸僅能捕捉大致的東西向梯度，無法反映區域差異；隨機森林與 KNN 能較精確地重現溫度的空間分布，與眞實情況更爲接近；而 SVR（RBF）則產生較平滑的分布，但在部分區域有低估或高估的情形。

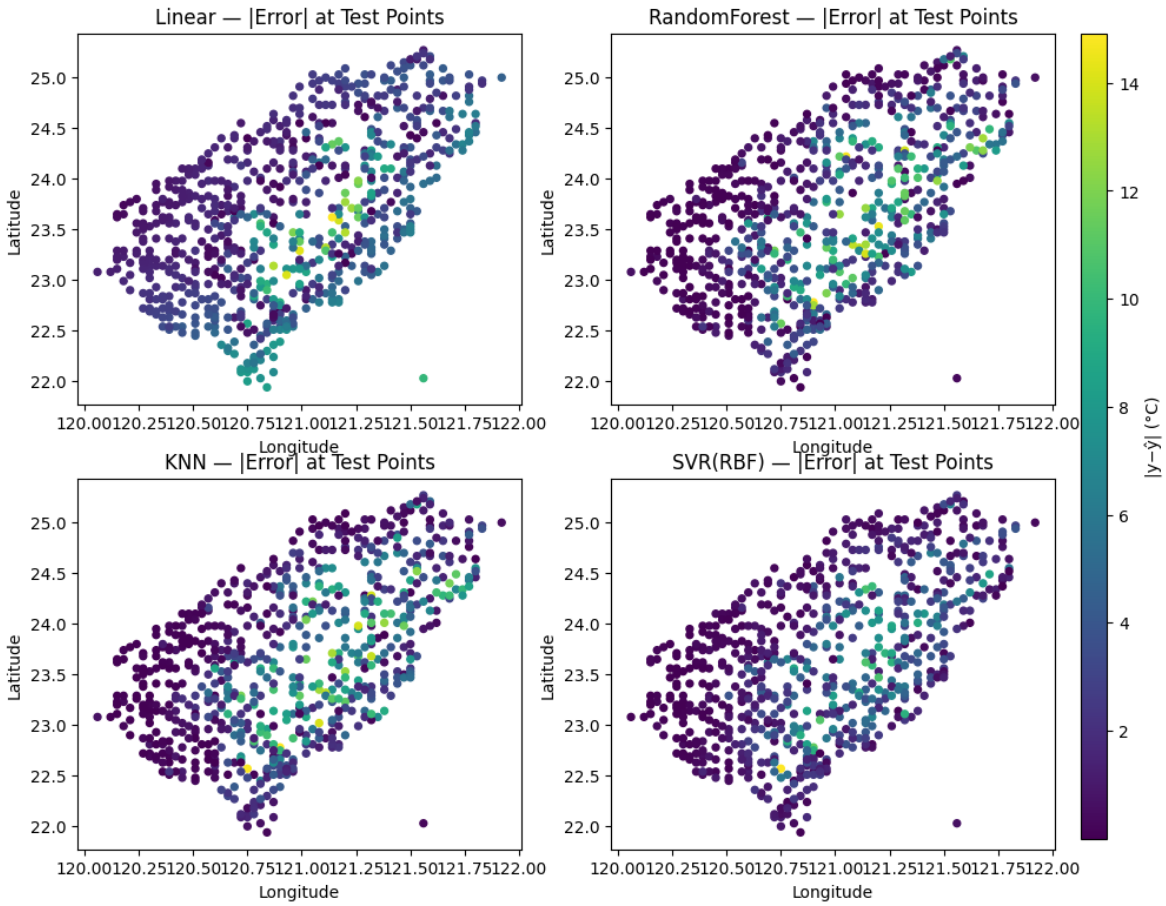

Figure 7: 四種迴歸模型在測試資料點的絕對誤差分布（$|y-\hat{y}|$）。點的顏色代表預測誤差大小（單位：°C），顏色越亮表示誤差越大。線性迴歸在整體上誤差普遍偏高；隨機森林與 KNN 的誤差相對較小，僅在部分邊界區域出現偏差；SVR（RBF）雖能捕捉部分非線性結構，但誤差仍明顯高於隨機森林與 KNN。