

Machine Learning Assignment 03

313652008 黃睿帆

September 29, 2025

Unanswered question in the class

- When using regression model to approximate periodic functions, will **the Fourier Series** give the best approximation? How complicated that the neuron network of Fourier Series looks like?

Programming assignments(本週作業)

利用 HackMD 上面所提供的氣象署資料觀測平台之”溫度分布-小時溫度觀測分析格點資料”(temperature.xml). 要進行資料轉換和模型訓練, 以下為資料說明:

1. 每個格點是一個溫度觀測值 (°C)
2. 資料無效值為 -999.。
3. 經緯度解析度: 經向、緯向各為 0.03 度。
4. 最左下角為第一個格點, 座標為: 東經 120.00 度; 北緯 21.88 度
5. 經向方向先遞增 (一列有 67 個數值), 再緯向方向遞增 (總共有 120 列), 因此資料形成一個 67×120 的數值網格。

1. 資料轉換

將原始資料轉換為兩個監督式學習資料集:

- (i) Classification: 將資料轉成輸出 data:(經度, 緯度, label), 若溫度觀測值為無效值 -999, 則 label = 0。若溫度觀測值為有效值, 則 label = 1。
- (ii) Regression: 將資料轉成輸出 data:(經度, 緯度, Value), 這裡僅保留有效的溫度觀測值 (剔除所有 -999.)。而 Value 為對應的攝氏溫度。

Python 程式 4-1 主要功能: 讀取 XML 檔, 並尋找特定內容 (<Content>) 取出其中的文字內容: 一大串數值字串 (就是我們所需的 data:(經度, 緯度, 溫度觀測值))。並利用 split 把數字字串切成一個一個的數值字元以及透過 float 把每個字轉換成浮點數。以文字轉數值, 最後分別輸出兩個 csv 檔案。(結果已放於 Github)

2. 模型訓練

使用上面整理出的兩個資料集, 分別訓練一個簡單的機器學習模型: (Note: Part of some Python codes are generated by ChatGPT, and some results part are discussing with ChatGPT.)

- (i) 分類模型 (classification model): 以 (精度, 緯度) 預測格點資料是否為有效值 (0 或 1)。
- (ii) 回歸模型 (regression model): 以 (精度, 緯度) 預測對應的溫度觀測值。

2-(i) Classification Model

這裡使用的是一個模型來判斷某個格點的數值是否為「有效」(Valid or invalid)。模型類型：Logistic regression，這裡採用 (logistic) sigmoid function 來做 Classification:

- (a) input $x = (x_1, x_2) = (\text{經度}, \text{緯度})$
- (b) output $y = (\text{label}) = 0$ or 1 , 0 means invalid point (-999), 1 means that it has (temperature) value.
- (c) hypothesis function: $h_{\theta}(x) = \sigma(w_0 + w_1 \cdot \text{lon} + w_2 \cdot \text{lat})$, where $\sigma(z) = \frac{1}{e^{-z} + 1}$ is sigmoid function.
- (d) Loss function= Binary cross-entropy loss= $Loss = -\frac{1}{N} \sum_{i=1}^N \{y_i \log(h_{\theta}(x_i)) + (1-y_i)(1-\log(h_{\theta}(x_i)))\}$

輸出的 $h_{\theta}(x) \in (0, 1)$, 即「格點為有效值的機率」。接著將 dataset 進行 Training/Validation split: 資料隨機拆成訓練集 (train set) 和驗證集 (validation set), 訓練集用來學習權重 w_0, w_1, w_2 , 驗證集用來監控 loss 下降, 避免 overfitting。Optimization 方法為 SGD(stochastic gradient descent)。

此處為訓練結果 (Class=Classification):

```
[Class] Epoch 20/100, train_loss=0.6936, val_loss=0.7004
[Class] Epoch 40/100, train_loss=0.6963, val_loss=0.6960
[Class] Epoch 60/100, train_loss=0.6867, val_loss=0.6938
[Class] Epoch 80/100, train_loss=0.6859, val_loss=0.6867
[Class] Epoch 100/100, train_loss=0.6844, val_loss=0.6868
```

Accuracy: 0.5100

Precision: 0.0000, Recall: 0.0000

Confusion matrix:

```
[[615  96]
 [495   0]]
```

Loss 曲線: Train loss 和 Validation loss 在起始時分別在 0.69 和 0.70 (為隨機猜的水準), 但到最後只有微幅下降, 大約接近 0.68, 代表模型沒有學到太多區分規則。

另外, Accuracy 約等於 0.51, 跟隨機猜是 0 是 1 差不多 (50%)。也因為 Precision 和 Recall 皆為 0, 表示這次的模型幾乎完全沒判斷出「有效值」。而混淆矩陣: $\begin{bmatrix} 615 & 96 \\ 495 & 0 \end{bmatrix}$; (左上)615: 表示正確預測的無效點; (右上)96: 把有效點錯當無效; (左下)495: 把無效點錯當有效; (右下)0: 完全沒有抓到任何有效點。

因此 Classification Model 基本上是不成功的。推斷可能的原因包含: (1) 地理座標 (lon, lat) 與是否為缺值幾乎無關 (缺值可能是觀測網格本身造成 (台灣的國土形狀), 形狀並非地理上能透過一般函數所預測)。(2) 無效 (-999) 資料點過多: 無效的和有效資料筆數相差太多 (2000 多筆), 無效資料筆數 5000 多筆占全部資料比例超過一半以上。(3) 模型過於簡單 (線性函數不足以表達地理分佈模式)。

2-(ii) Regression Model

這裡使用的是一個模型來以 (經度, 緯度) 預測溫度值 (排除 -999. 的無效點)。

- (a) input $x = (x_1, x_2) = (\text{經度}, \text{緯度})$
- (b) output $y =$ 對應格點的「實際溫度值」
- (c) hypothesis function: $h_{\theta}(x) = w_0 + w_1 \cdot \text{lon} + w_2 \cdot \text{lat}$.
- (d) Loss function= Binary cross-entropy loss= $Loss = \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)^2$

同樣像 2-(i), 將 dataset 進行 Training/Validation split, 也就是同樣分成訓練集/驗證集, 來監控訓練誤差與驗證誤差。也一樣使用 SGD。

訓練紀錄與結果測試如下 (Reg=Regression):

```
[Reg] Epoch 20/100, train_mse=55.45, val_mse=55.47
[Reg] Epoch 40/100, train_mse=39.55, val_mse=40.34
[Reg] Epoch 60/100, train_mse=35.61, val_mse=36.91
[Reg] Epoch 80/100, train_mse=35.95, val_mse=37.91
[Reg] Epoch 100/100, train_mse=36.72, val_mse=37.44
```

MSE (test): 31.618

MAE (test): 3.991

Max absolute error (test): 20.351

Loss 曲線的 MSE 從大約 55 降到大約 36, 也表示著 Regression Model 有學到一點點關聯。另外 MSE(平均平方誤差) 為 31.6, 取平方根後的 RMSE 約等於 5.6°C; MAE(平均誤差) 大約是 4.0°C; 最大誤差約 20.35°C, 也就是情況最差會差到 20°C (估計有可能出現在國土形狀邊界或某些外島的點 (此為推測))。

同時, Validation curve 顯示: 訓練誤差先下降再趨於平穩, 驗證誤差與訓練誤差相差不大, 沒有過度擬合, 但模型可能太簡單 (因為仍有一定誤差)。

3. 總結

Regression Model 比 Classification Model 成功一些, 能稍微捕捉到溫度與經緯度之間的分佈趨勢, 但精準度仍然較為有限。

(推測): 單純使用資料點 (經度, 緯度), 來去推估該點溫度會不夠精準, 實際的溫度受到地勢高度 (海拔)、地形、洋流 (靠不靠海)、測量時天氣 (下雨/太陽) 等因素影響, 而非單純看經緯度。另外可能也因為假設的模型是線性的, 如果換成非線性, 表現可能會更好。

因此, Classification Model 基本上幾乎失敗 (僅有隨機水準 (50%)), 原因可能是: 資料分布高度不平衡 (有效 vs. 無效點比例差太多), 可能只使用經緯度資料 (lon, lat) 無法區分有效性, 亦即是經緯度無法有效預測缺值, 因為缺值 (-999.) 是資料特性而非物理上或是函數上的規律。但是 Regression Model 有一點結果, (平均誤差僅 4°C), 可能需要: 加入更多特徵 (時間、地形高度、鄰近點溫度...) 或使用更強的模型, 來降低仍然算是較大的誤差, 但這也顯示了單純只靠地理座標不足以精確描述溫度分布。