**問題一：How do you select features for your model input, and what preprocessing did you perform?（如何選擇模型輸入的特徵，以及執行了哪些預處理？）**

本專案的特徵選擇與預處理過程經過了多次迭代，最終採用了「先用領域知識精簡，再以模型驗證」的策略。整個流程如下：

**第一部分：數據預處理(Data Preprocessing)**

為了將原始的 `train.csv` 轉換為乾淨、可用的格式，執行了以下預處理步驟：

1. **數據清洗**

原始數據中的無效值（如 `'A`, `\*`, `x` 等）會被統一替換，以便後續進行數值計算

1. **插值填補**

對於缺失的數據點，採用**線性插值法**進行填補。

鑒於訓練數據為每個月的前20天，為避免不同月份數據的相互干擾（橫跨天數），插值是在每個月份內部獨立進行的

1. **數據重塑**

將時序數據轉換為監督式學習的樣本格式

採用大小為9的Sliding Window，將連續9個小時的所有氣象數據作為模型輸入特徵 (X)，並以第10個小時的PM2.5數值作為預測目標 (y)

**第二部分：特徵工程與選擇(Feature Engineering & Selection)**

**初版策略（已廢棄）：**

最初嘗試先利用相關性篩選出8個核心特徵，並為其生成**二次項**和**交互項的**組合特徵。希望透過Lasso (L1正規化) 從中自動篩選出有效特徵。然而實驗證明，在本次的小型資料集上，過於複雜的特徵不僅沒有帶來提升，反而引入了大量噪音，導致模型過擬合，個人認為，這屬於維度的詛咒

**最終策略：**

經過數次的嘗試，最終採用了相對簡單卻成功的策略

1. **基於領域知識的核心特徵選擇**

手動指定了在環境科學中最具相關性的**4個核心特徵**：`PM2.5` (自身歷史值), PM10, O3, CO

1. **簡化的特徵組合**

僅使用了這4個核心特徵的9小時歷史值及其**平方項**，總共4 \* 9 \* 2 = 72個特徵

這個數量的特徵既捕捉了數據的非線性關係，又避免了維度詛咒

1. **Lasso篩選**

此處依然使用Lasso (alpha=0.05) 對這72個特徵進行篩選，它最終保留了71個特徵

總結來說，最終特徵策略是：**以領域知識為基礎，進行小範圍、有根據的特徵擴展，得到少量卻實用的特徵**

**問題二：比較不同數量的訓練數據對PM2.5預測準確度的影響，並視覺化呈現（Compare the impact of different amounts of training data on the PM2.5 prediction accuracy. Visualize the results and explain them.）**

**實驗方法**

1.  固定驗證集不變

2.  從完整訓練集中，依序抽取 20%, 40%, 60%, 80%, 100% 的數據作為不同規模的訓練子集

3.  使用每個訓練子集分別訓練一個新的模型

4.  在固定的驗證集上評估各個模型的 RMSE，並將結果繪製成圖

**結果與分析**

一張含有 文字, 行, 繪圖, 圖表 的圖片

AI 產生的內容可能不正確。

上圖展示的學習曲線清晰地揭示了數據量與模型性能的關係

從圖中可以觀察到，隨著訓練數據量的增加（從20%到100%），模型在驗證集上的RMSE呈現出**穩定下降**的趨勢，這表明：

1. **模型並未飽和**

即使使用了100%的訓練數據，曲線仍未完全走平，這意代表如果能獲得更多的數據，模型的性能很可能還會繼續提升

1. **數據的重要性**

這個實驗有力地證明了數據量對模型性能的正面影響，更多的數據能讓模型學習到更普適、更穩健的規律，從而提高其泛化能力（不過這裡的數據量本身就不多，當數據量多到一個程度，RMSE應該會再次出現不穩定，有點像盲人摸象，當數據量極大，我們就很難看清數據的規律）

**問題三：討論正規化對 PM2.5 預測準確度的影響（Discuss the impact of regularization on PM2.5 prediction accuracy.）**

正規化在模型中扮演了至關重要的角色，它能有效防止模型過擬合，提升泛化能力。此次作業，在兩個階段分別使用了不同目的的正規化

**第一階段：使用L1 (Lasso)進行特徵篩選**

**目的**：

在面對上百個候選特徵時，利用L1正規化的稀疏性（Sparsity）特點，將不重要特徵的權重壓縮至零，從而自動篩選出一個更精簡、更高效的特徵子集

**影響：**

雖然在最終方案中，由於手動選擇的核心特徵質量很高，Lasso 只排除了極少數特徵，但在早期的複雜模型探索中，它是對抗維度詛咒、分析特徵重要性的重要工具

**第二階段：使用L2 (Ridge)提升模型穩定性**

在最終模型的訓練時，加入了L2正規化項，並透過最佳正規化參數搜索尋找最佳的

**實驗方法：**

遍歷一組預設的值（如[0, 0.001, 0.01, 0.1, 1]），訓練多個模型並在驗證集上評估其RMSE

**結果與分析：**

一張含有 文字, 行, 繪圖, 圖表 的圖片

AI 產生的內容可能不正確。

**當過小（趨近於0）時**：

模型接近於一個普通的線性迴歸，沒有足夠的懲罰項來约束權重，容易過度擬合訓練數據中的噪音，導致在驗證集上表現不佳

**當過大時**：

對權重的懲罰過於嚴厲，導致模型欠擬合，模型更加專注於壓制權重，表現同樣會變差

**結論：**

根據訓練最終模型時，最佳正規化參數搜索的結果顯示，是當下最佳選擇

不過先前訓練時根據特徵的不同也會出現或其他值的情況下有更好的表現

這表明一個**適度**的L2正規化是必需的，只有在擬合參數與最小化權重間取得平衡，才能在使模型更平滑（增加泛化能力）的同時確保其準確性

Github：<https://github.com/megrez33281/Linear-Regression-PM2.5-Prediction>