HW01

PM2.5 prediction

313581038 智能系統 蒲品憶

1. **How do you select features for your model input, and what preprocessing did you perform?**

**選擇特徵:**

每項特徵的與PM2.5的相關係數

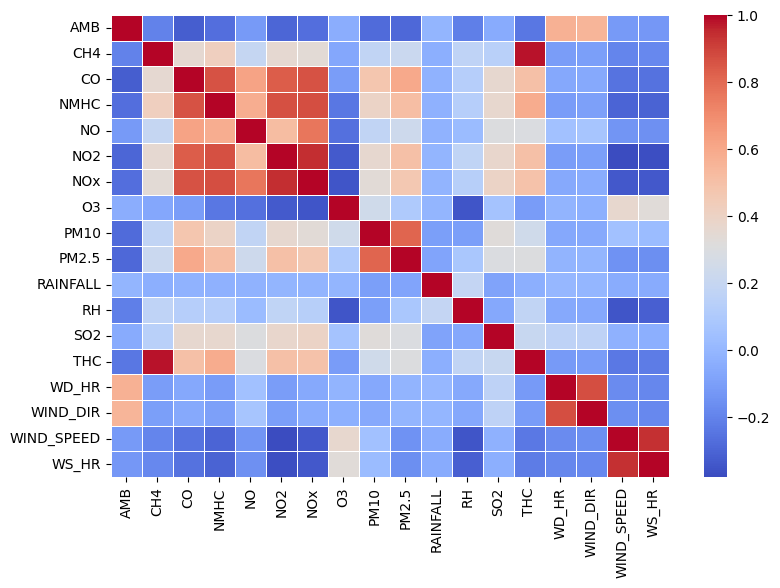


Figure1: 相關係數圖

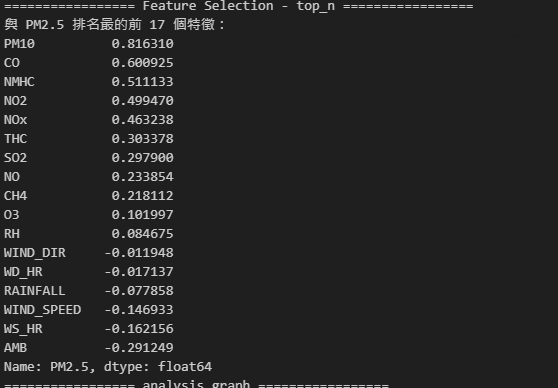


Figure2: 各項數值

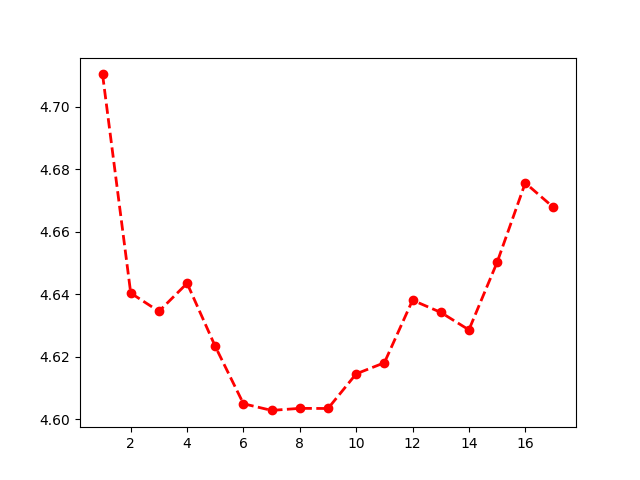


Figure3: 選取不同特徵對valid loss的影響

從Figure1可以觀察出PM10對於PM2.5的關係最接近、CO次之、NMHC、NO2、NO並列、後面接著是SO2、THC……。

在Figure2能夠清楚知道每項特徵與PM2.5對應的相關係數的詳細數值，其中以PM10對於PM2.5的關係最接近。最初挑選特徵，我是先將小於0以下的特徵刪除，效果明顯比使用全部特徵valid loss下降許多；後來又嘗試使用top 6的相關features作為特徵，效果有稍微好一些。

之後，我又想說試試全部不同特徵對於valid loss的影響，於是產生Figure3。從Figure3可以看出，選擇top 6 -top 9的features對於valid loss下降效果較佳，經過丟上kaggle測試後，最後認定取top6 features對於目前任務效果最佳。

**前處理:**

**數據清理**：將#,\*,x,A改成用零替換。在實作時發現test data還有個不是數值的值'WIND\_DIR+D2070EC'也將之進行替換，其實也可以用平均值替換但是比較費工夫，所以這裡我只用補零處理。做數據清理好處是，可以避免丟失大量有價值的信息，從而保留更多的數據以供分析。而且要做要做linear regression之前要先將數據都轉換成數值。

**數據標準化**：將數據縮放到相同範圍，標準化（z-score）或最小-最大縮放。

**公式: x = (x – mean\_x) / std\_x**

標準化的好處:

1. **加快收斂速度:** 梯度下降法會受到數據範圍的影響。標準化可以讓所有特徵處於相同的尺度，這樣梯度下降等優化算法會收斂得更快。
2. **提高模型性能:** 標準化能夠讓各個特徵對模型訓練的影響更加均衡，使得每個特徵都在相似的範圍內，這有助於模型更好地學習每個特徵的影響，從而提高模型的性能。
3. **避免數據偏差:** 如果數據集中的某些特徵的數值範圍過大，未經標準化可能會導致模型訓練過程中的不均衡，這可能會引起模型的偏差。標準化有助於減少這種偏差，讓模型在所有特徵上學習的影響更為均衡。

1. **Compare the impact of different amounts of training data on the PM2.5 prediction accuracy. Visualize the results and explain them.**

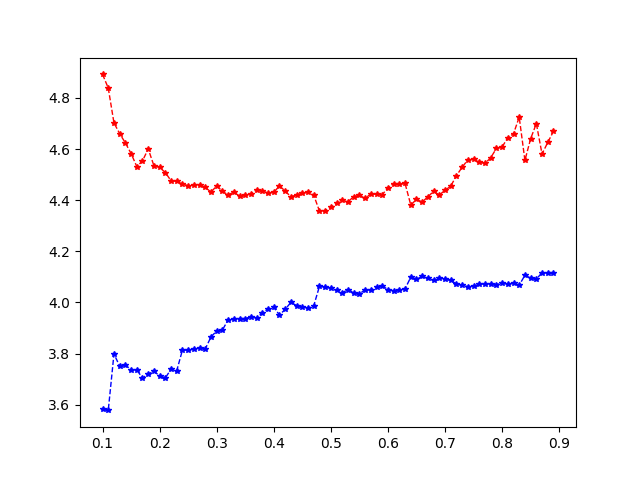


Figure3: 不同訓練集之下valid and train loss

紅色: valid loss、藍色: train loss、X: training Data、Y: predict

**訓練數據的影響**：藍色線條代表模型在訓練損失。

隨著訓練資料量的增加，訓練損失穩步下降，這是因為模型有更多資料來學習。

**測試數據的影響**：紅色代表的是測試損失。

驗證損失初期下降，但在某個點之後開始上升，顯示出過擬合(overfitting)的情形。這意味著當訓練資料過多時，模型開始學習過於專注於訓練資料中的細節，無法更好地處理新的資料。

總結:

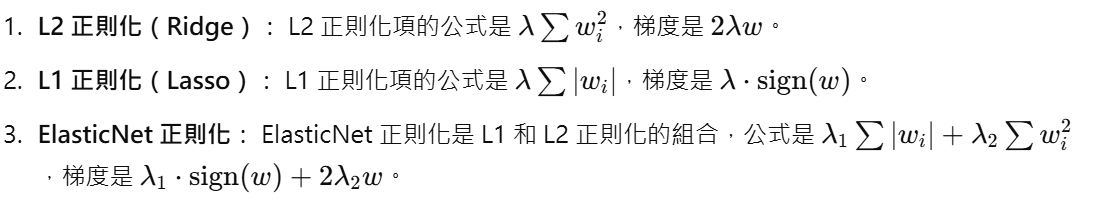
訓練及太多或太少都會造成效果不佳，使用training data雖然圖上，0.5-0.7區間valid loss最低，但是考慮到test Data的特徵值可能與valid Data不這麼相似，所以我選擇讓training data調整在0.7-0.8區間進行訓練，讓更多的特徵進行學習，有機會在test Data有更好的表現。

1. **Discuss the impact of regularization on PM2.5 prediction accuracy.**

(Regularization）在機器學習中主要是用來防止模型過擬合的技術，對於PM2.5預測準確度有著重要影響。當模型訓練時，尤其是在資料較為複雜或者訓練資料量較少的情況下，過擬合的風險會增高，這會導致模型在驗證集或測試集上的表現不佳。正則化技術通過在損失函數中加入額外的項來限制模型的複雜度，從而提高其泛化能力。

以下是我使用的Regularization方法，將會提及公式、三種方法說明以及三種方法，最終使用何種Regularization。

**正則化公式：**

****

**三種方法說明:**

1. L2 正則化（Ridge）

* **L2 正則化的作用**：對模型的權重參數進行懲罰，特別是權重過大時。它的正則化項是權重的平方和，這會使模型更加平滑，從而減少過擬合的風險。
* **λ 的影響**：
  + 當 **λ 较小** 时，正則化的效果較弱，模型的權重較大，可能容易過擬合。
  + 當 **λ 较大** 时，正則化會較強，會使模型的權重變得較小，進而降低過擬合的風險。但如果 **λ 太大**，模型的表現會變得過於簡單，可能會欠擬合（underfitting）。

2. L1 正則化（Lasso）

* **L1 正則化的作用**：L1 正則化會對模型的權重進行懲罰，並且會推動某些權重為零，這樣可以實現**特徵選擇**。L1 正則化的正則化項是權重的絕對值和，這樣某些特徵的權重會被壓縮到零，從而消除不重要的特徵。
* **λ 的影響**：
  + 當 **λ 较小** 时，L1 正則化的效果較弱，模型會包含更多的特徵，可能會導致過擬合。
  + 當 **λ 较大** 时，L1 正則化會強烈地懲罰權重，使得更多的特徵被壓縮到零，模型變得更簡單，避免過擬合，但也有可能造成欠擬合。

3. ElasticNet 正則化

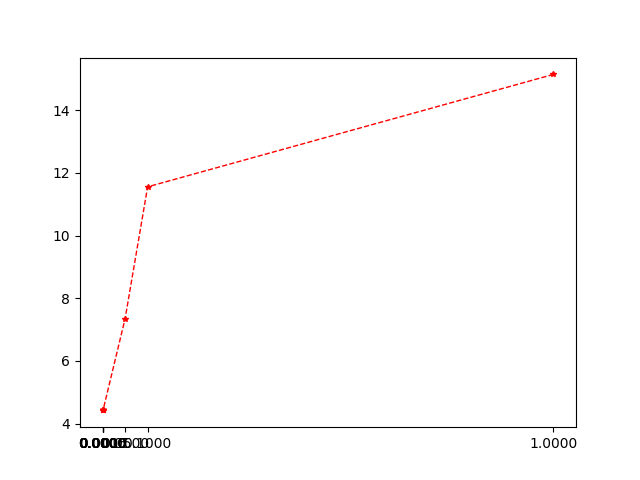
* **ElasticNet 正則化的作用**：ElasticNet 正則化結合了 L1 和 L2 正則化的特點，既能進行**特徵選擇，又能控制模型的複雜度**。它的正則化項包含 L1 和 L2 的加權和，因此既能保持 L1 的特徵選擇能力，又能避免 L2 正則化過度平滑的問題。
* **λ₁ 和 λ₂ 的影響**：
  + **λ₁** 控制 L1 正則化的強度（特徵選擇的效果）。
  + **λ₂** 控制 L2 正則化的強度（控制模型複雜度）。
  + 當 **λ₁ 较大** 時，更多的特徵會被壓縮為零，並且特徵選擇的效果會更強。
  + 當 **λ₂ 较大** 時，模型會更加平滑，權重會被壓縮到較小的值，減少過擬合的風險。

總結:

* L2 正則化（Ridge）：適用於所有特徵都對預測有貢獻的情況，並且 λ 值過大會導致欠擬合。
* L1 正則化（Lasso）：適用於特徵數量很大且懷疑某些特徵不重要的情況，並且 λ 值過大會過度縮小權重，導致過度簡化。
* ElasticNet 正則化：當數據中有冗餘特徵時，ElasticNet 是一個很好的選

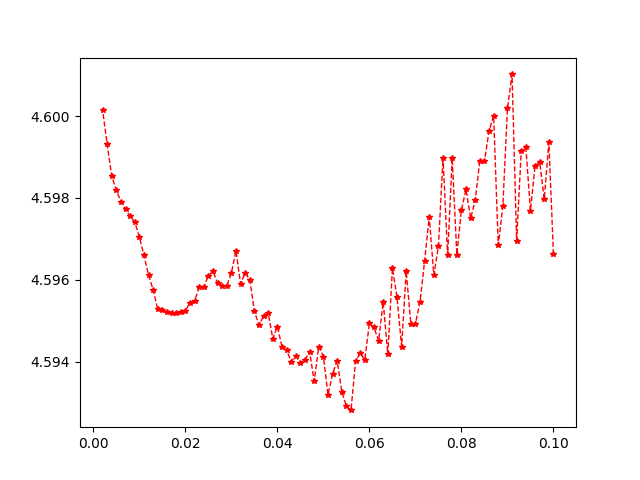
**實作比較:**

1. **L2 正則化（Ridge）**



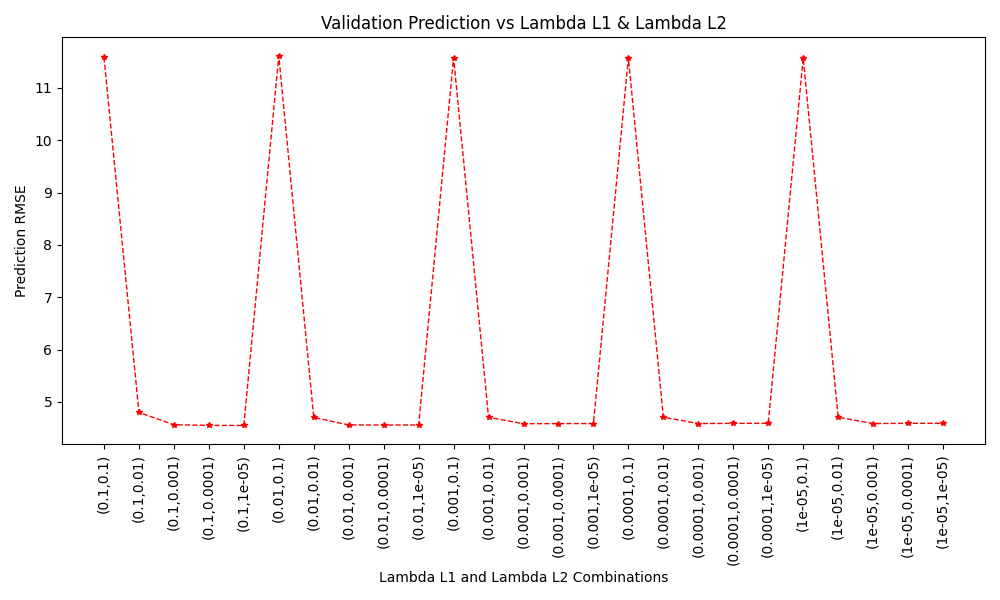
測試lambda\_reg=0.00001~1，測試結果可以看出在lambda\_reg=0.01時，valid loss數值最低，valid loss=4.6535841555

1. **L1 正則化（Lasso）**

****

測試lambda\_reg=0.01~0.1，測試結果可以看出在lambda\_reg=0.56，valid loss值最低，valid loss=4.587490810568106

1. **ElasticNet 正則化**



測試 lambda\_reg\_L2 = [0.1, 0.01, 0.001, 0.0001, 0.00001]，

lambda\_reg\_L1 = [0.1, 0.01, 0.001, 0.0001, 0.00001]，

測試結果可以看出在(lambda\_reg\_L1, lambda\_reg\_L2) =(0.1, 1e-05) 時，valid loss數值最低，valid loss=4.6972905330245

**結論:**

並沒有一定是哪個方法好壞，只是在此任務中時做測試的結果下，我使用L1 正則化（Lasso），在valid loss的效果最佳因此使用此方法。推測是因為L2 正則化（Ridge）的平方損失方式對於簡單的梯度下模型效果太強烈，導致成效不好；ElasticNet是L1+L2的混合版本，因此可能也有相同問題。