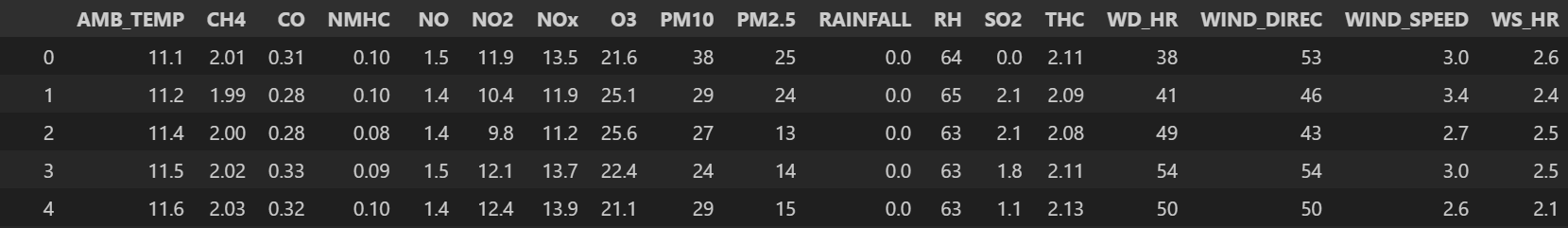
Data Mining HW1 Report 0816077 陳柏安

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

Features的選擇上，我一開始使用每小時除了PM2.5外的其他features作為input來訓練，output該小時的PM2.5預測值，雖然能通過simple baseline，但效果有限。所以我嘗試使用前一小時的所有features來預測下一小時的PM2.5值，成果大幅改善。最後我參考環境部官方網站的資訊，選擇了幾個跟PM2.5濃度相關的features，包括溫度、PM2.5、PM10、NOx、風速和相對溼度來進行訓練，防止較不相關的數值影響訓練結果。

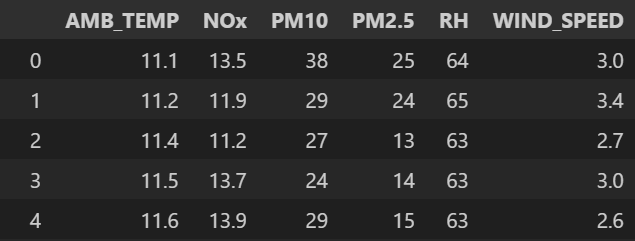
(Reference：[細懸浮微粒一日變化特徵 - 空氣品質監測網 (moenv.gov.tw)](https://airtw.moenv.gov.tw/cht/Encyclopedia/pedia09/pedia9_3.aspx))

資料處理的部分，首先我將每小時的共18個features轉成columns的欄位，每個row的資料會是每小時的各feature數值，如下圖。



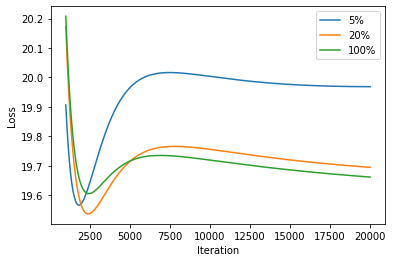
而且因為有些欄位的數值是NaN，我們還需要用fillna來將這些欄位填上0，避免我們的model遇到無法處理的資料。接下來將每小時的數據當作X，並擷取下一個小時的PM2.5值當作y來處理。也因為我只需要使用前一小時的資料，在testing data的部分我也只擷取最後一小時的資料就能進行預測。

擷取合適的features後剩下6個columns如下。



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

為了觀察訓練資料量對於訓練結果的差異，我先將training data和validation data以90%和10%的比例分組，接著分別用training set的100%、20%、5%資料以相同參數進行訓練，並將訓練過程在validation data上的loss變化畫成圖表。



從圖中我們能看出，資料量最少的5%在剛開始訓練時loss下降速度是最快的，但時間一旦拉長，就會因為資料集中的偏差樣本而出現overfit導致loss突然上升，在訓練後期也未能達到理想的數值，而反觀資料量大的20%及100%，雖然起初的loss下降沒有那麼快，但overfit的狀況也相對輕微，最終的loss值也好上許多。

3. Discuss the impact of regularization on PM2.5 prediction accuracy.

這個部分我用numpy實作了lasso regularization，並比較和實驗了不同係數對於模型預測的影響，lambda分別從0到100000，在epoch和learning rate固定的情況下，可以看到訓練完成後weight的結果如下。



圖中在lambda值稍大的情況許多係數都被降為0或趨近於0，達到類似feature selection的效果，但當lambda值過大（超過10000）連一些重要參數，如PM2.5的weight都被影響，也造成預測結果的偏差，視覺化如下圖。

