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))

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。

	AMB_TEMP	CH4	CO	NMHC	NO	NO2	NOx	O3	PM10	PM2.5	RAINFALL	RH	SO2	THC	WD_HR	WIND_DIREC	WIND_SPEED	WS_HR
0	11.1	2.01	0.31	0.10	1.5	11.9	13.5	21.6	38	25	0.0	64	0.0	2.11	38	53	3.0	2.6
1	11.2	1.99	0.28	0.10	1.4	10.4	11.9	25.1	29	24	0.0	65	2.1	2.09	41	46	3.4	2.4
2	11.4	2.00	0.28	0.08	1.4	9.8	11.2	25.6	27	13	0.0	63	2.1	2.08	49	43	2.7	2.5
3	11.5	2.02	0.33	0.09	1.5	12.1	13.7	22.4	24	14	0.0	63	1.8	2.11	54	54	3.0	2.5
4	11.6	2.03	0.32	0.10	1.4	12.4	13.9	21.1	29	15	0.0	63	1.1	2.13	50	50	2.6	2.1

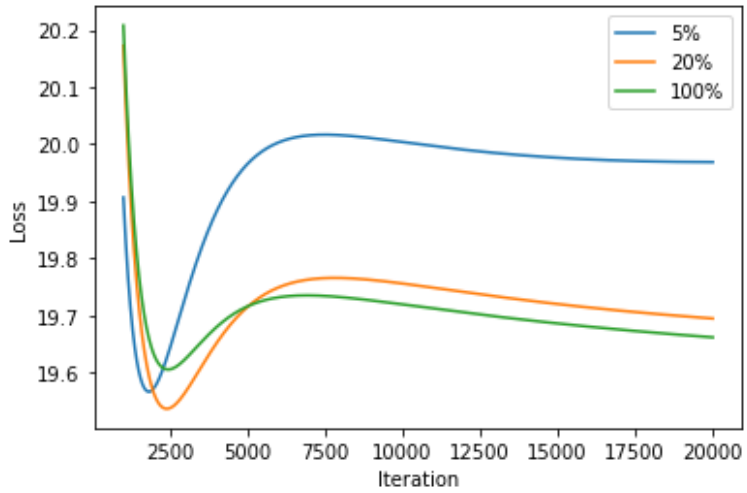
而且因為有些欄位的數值是 NaN，我們還需要用 fillna 來將這些欄位填上 0，避免我們的 model 遇到無法處理的資料。接下來將每小時的數據當作 X，並擷取下一個小時的 PM2.5 值當作 y 來處理。也因為我只需要使用前一小時的資料，在 testing data 的部分我也只擷取最後一小時的資料就能進行預測。

擷取合適的 features 後剩下 6 個 columns 如下。

	AMB_TEMP	NOx	PM10	PM2.5	RH	WIND_SPEED
0	11.1	13.5	38	25	64	3.0
1	11.2	11.9	29	24	65	3.4
2	11.4	11.2	27	13	63	2.7
3	11.5	13.7	24	14	63	3.0
4	11.6	13.9	29	15	63	2.6

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。

	AMB_TEMP	CH4	CO	NMHC	NO	NO2	NOx	O3	PM10	PM2.5	RAINFALL	RH	SO2	THC	WD_HR	WIND_DIREC	WIND_SPEED	WS_HR	Bias
0	-0.0313	-0.1421	0.4524	0.1375	-0.0470	0.0519	0.0303	0.0076	0.1267	0.6652	-0.2012	0.0183	-0.2603	-0.0040	0.0002	0.0020	-0.1559	-0.1030	0.2974
100	-0.0323	-0.0718	0.3008	0.0000	-0.0140	0.0829	-0.0000	0.0071	0.1261	0.6666	-0.1811	0.0172	-0.2464	0.0000	0.0002	0.0019	-0.1600	-0.0930	0.2821
1000	-0.0306	-0.0000	0.0000	0.0000	-0.0049	0.0762	0.0000	0.0036	0.1235	0.6730	-0.0232	0.0145	-0.1110	0.0000	0.0001	0.0017	-0.1832	-0.0000	0.2356
10000	-0.0112	-0.0001	-0.0000	-0.0000	-0.0001	0.0624	0.0000	-0.0017	0.1174	0.6796	-0.0000	0.0091	-0.0000	-0.0001	-0.0029	-0.0015	-0.0001	-0.0001	-0.0058
100000	-0.0085	-0.0002	-0.0002	-0.0000	-0.0003	-0.0002	0.0004	-0.0096	0.1531	0.5534	-0.0001	0.0003	-0.0001	-0.0004	-0.0321	-0.0302	-0.0006	-0.0002	1.1918

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。