ANN Homework 2 due 2021/11/29

2. 利用課程提供之環保署空氣品質監測站數據，完成類神經網路PM2.5預測模式，輸入資訊為前9小時觀測值，輸出第10小時PM2.5濃度值，詳細內容請參閱10/26投影片，並完成下列題目要求：

a. 將訓練資料(train_X.txt, train_Y.txt)分成3組

b. 以交叉驗證方式訓練BPNN模式，列出不同分組組合之RMSE與R^2值(表格)

c. 嘗試至少2種以上的輸入因子組合重作b小題，說明不同組合輸入因子挑選方式或理由，並比較其結果。

d. 從c小題結果中選擇你認為較佳的模式，以測試資料(test_X.txt, test_Y_real.txt)進行模式測試，比較其RMSE與R^2值與b小題交叉驗證結果之差異

e. 改以RBFNN模式進行訓練，重作b、d小題

f. 請比較BPNN與RBFNN之圖/表預測結果

備註：比較部分需有文字說明

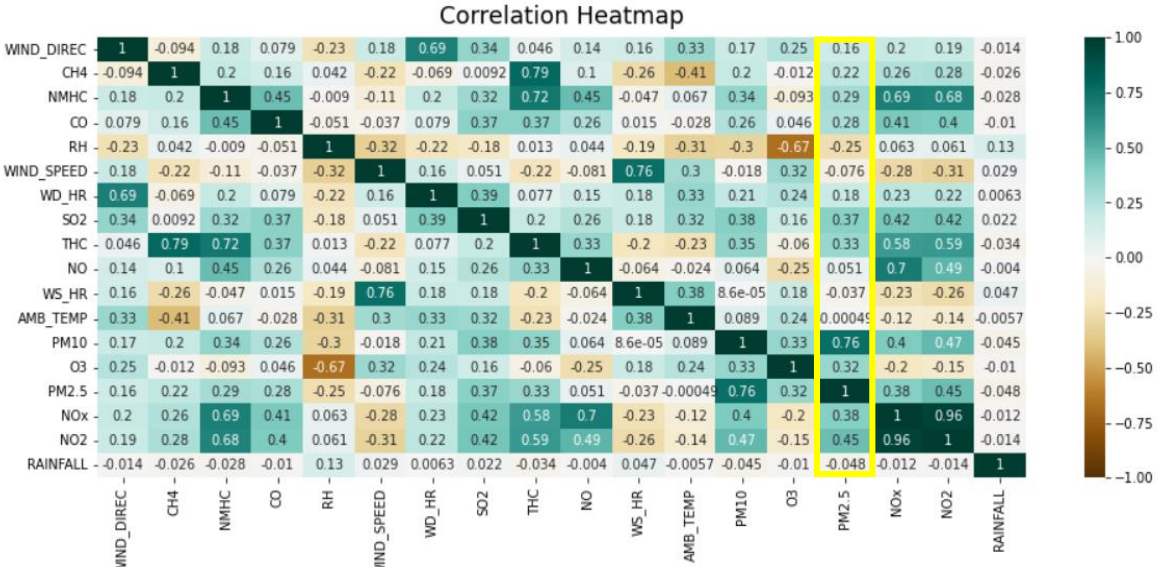
Q2

(a)- train 資料集->隨機分成 3 等分 (train01、train02、train03)

(b)

	G1			G2			G3		
BPNN	Training		validation	Training		validation	Training		validation
	train 01	train 02	train 03	train 01	train 03	train 02	train 02	train 03	train 01
18	'WIND_DIREC','CH4','NMHC','CO','RH','WIND_SPEED','WD_HR', 'SO2','THC','NO','WS_HR','AMB_TEMP','PM10','O3','PM2.5', 'NOx','NO2','RAINFALL'								
RMSE	5.9248		6.8917	6.2165		6.2498	5.9496		6.3761
R^2	0.8685		0.8261	0.85824		0.8626	0.86729		0.8526

(c) 將 18 個變數因子繪製成相關係數圖，觀察相關係數圖，選擇和 PM2.5 相關係數較大的因子，並按照不程度的相關性(0.1、0.3、0.4)，來篩選變數因子。如下表，因子個數由原本的 18 個，分別減少成 3、7、12 個，來訓練模型。
其實用這三種因子組合所訓練出來的模型的模型表現都差不多，都有不錯的結果，但是若使用 3 個因子，可能會因為因子數太少，表現會稍微比較浮動、不穩定。若使用 12、18 個因子數會造成太多時間成本。因此在 BPNN 中，選擇和 PM2.5 相關係數絕對值大於 0.3 的=> NO2、NOx、PM10、O3、THC、SO2、PM2.5 ，並利用 G2 ，是最好的組合。



	G1			G2			G3		
BPNN	Training		validation	Training		validation	Training		validation
	train 01	train 02	train 03	train 01	train 03	train 02	train 02	train 03	train 01
	和 PM2.5 相關係數絕對值大於 0.4 的=> NO2、PM10、PM2.5								
RMSE	5.6512		6.5836	6.1248		5.6539	5.8860		5.9179
R^2	0.8798		0.8406	0.8578		0.8721	0.8678		0.8673
	和 PM2.5 相關係數絕對值大於 0.3 的=> NO2、NOx、PM10、O3、THC、SO2、PM2.5								
RMSE	5.668		6.443	6.111		5.784	5.9858		6.0234
R^2	0.878		0.8416	0.8592		0.8729	0.8700		0.8640
	和 PM2.5 相關係數絕對值大於 0.1 的=> WIND_DIREC、CH4、NMHC、CO、RH、WD_HR、 NO2、NOx、PM10、O3、THC、SO2、PM2.5								
RMSE	5.7846		6.6961	6.0541		5.9978	5.8215		6.2331
R^2	0.87314		0.82881	0.8605		0.8670	0.8720		0.8570

(d) 挑選上題最佳的組合的模型，對 Test 資料做測試。可以發現依然有不錯的測試結果，RMSE 和 R 平方，都和交叉驗證後最佳組合的結果差不多，沒有 overfitting、underfitting 的情形發生。

BPNN	Training		validation	Testing
	train 01	train 03	train 02	(test_X.txt,test_Y_real.txt)
7	和 PM2.5 相關係數絕對值大於 0.3 的=> NO2、NOx、PM10、O3、THC、SO2、PM2.5			
RMSE	6.111		5.784	6.8723
R^2	0.8592		0.8729	0.8664

e-(b)

	G1			G2			G3		
RBF	Training		validation	Training		validation	Training		validation
	train 01	train 02	train 03	train 01	train 03	train 02	train 02	train 03	train 01
18	'WIND_DIREC','CH4','NMHC','CO','RH','WIND_SPEED','WD_HR', 'SO2','THC','NO','WS_HR','AMB_TEMP','PM10','O3','PM2.5', 'NOx','NO2','RAINFALL'								
RMSE	8.1571		8.897	8.272		8.48	8.12		8.30
R^2	0.745		0.706	0.743		0.737	0.751		0.739

e-(c)

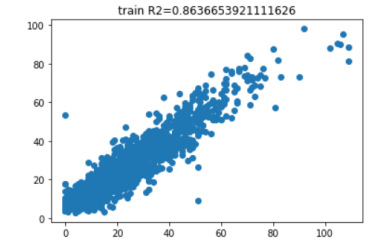
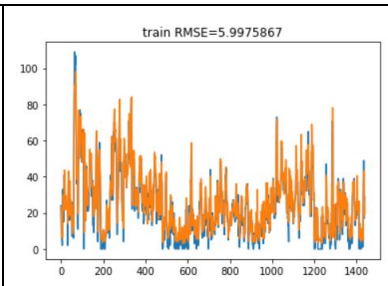
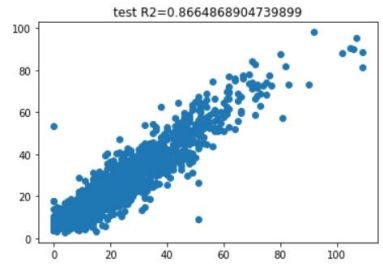
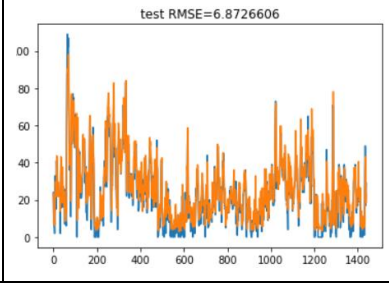
在 RBFNN 中，若加入太多變數，反而模型表現比較差，如(b)小題的 18 個因子，和下表的 12 個因子。，選擇 和 PM2.5 相關係數絕對值大於 0.4 的=> NO2、PM10、PM2.5 ，並利用 G2 ，是最好的組合。

	G1			G2			G3		
RBF	Training		validation	Training		validation	Training		validation
	train 01	train 02	train 03	train 01	train 03	train 02	train 02	train 03	train 01
	和 PM2.5 相關係數絕對值大於 0.4 的=> NO2、PM10、PM2.5								
RMSE	6.53		7.57	6.82		6.51	6.678		6.51
R^2	8.384		0.784	0.823		0.843	0.830		0.842
和 PM2.5 相關係數絕對值大於 0.3 的=> NO2、NOx、PM10、O3、THC、SO2、PM2.5									
RMSE	6.437		7.43	6.81		6.67	6.773		6.709
R^2	0.844		0.795	0.824		0.836	0.825		0.830
和 PM2.5 相關係數絕對值大於 0.1 的=> WIND_DIREC、CH4、NMHC、CO、RH、WD_HR、 NO2、NOx、PM10、O3、THC、SO2、PM2.5									
RMSE	7.817		8.571	8.172		8.237	7.980		8.551
R^2	0.770		0.729	0.748		0.749	0.7580		0.721

e-(d) 挑選上題最佳的組合的模型，對 Test 資料做測試。可以發現依然有不錯的測試結果，RMSE 和 R 平方，都和交叉驗證後最佳組合的結果差不多，沒有 overfitting、underfitting 的情形發生。

RBF	Training		validation	Testing
	train 01	train 03	train 02	(test_X.txt,test_Y_real.txt)
7	和 PM2.5 相關係數絕對值大於 0.4 的=> NO2、PM10、PM2.5			
RMSE	6.82		6.51	7.852
R^2	0.823		0.843	0.800

(f) BPNN 在訓練模型時，比較不會受到不同因子組合的影響，造成模型表現有太大的浮動。反之在 RBFNN 中太多沒有明顯影響 PM2.5 的因子加入模型，會讓模型訓練的比較不好。在各自選擇最好的組合，並對同一筆 TEST 資料做測試時，BRNN 的表現似乎比較好。R 平方比較大，Rmse 比較小。

BPNN		
		

RBFNN