Machine Learning 2016 Fall
HW1 – PM2.5 prediction

b03902035 黃兆緯

1. Linear regression by Gradient descent

```
def func(w, x):
    # simply returns w dot x = w1x1 + w2x2 + ... + wnxn
    return np.dot(w, x)

def func_grad(w, x, y):
    # compute gradient of w
    diff = y[0]-func(w,x)
    wgrad = np.array([2*(-x[i])*diff for i in range(len(x))])
    return wgrad

# run several passes
for i in range(iteration+1):
    # compute gradients of w and sum over all training data
    wgrad = sum(func_grad(w, feature_train[j], ans_train[j]) for j in range(len(feature_train)))

    # compute summation of past gradients, for adagrad
    w_acc = w_acc + wgrad**2

    # update parameters, using adagrad
    w = w - rate*(1.0/(w_acc)**0.5)*wgrad

    # compute and print training error/validation error every 1 pass
    if i % 1 == 0:
        # E_in(training set error)
        train_ans = np.dot(feature_train, w)
        # use RMSE as error measurement
        train_error = np.sqrt(np.mean((ans_train[:,0] - train_ans)**2))

        # validation set error
        valid_ans = np.dot(feature_valid, w)
        valid_error = np.sqrt(np.mean((ans_valid[:,0] - valid_ans)**2))
        print('iteration %d,\ttrain error: %f,\t valid error: %f' % (i, train_error, valid_error))
```

2. 方法

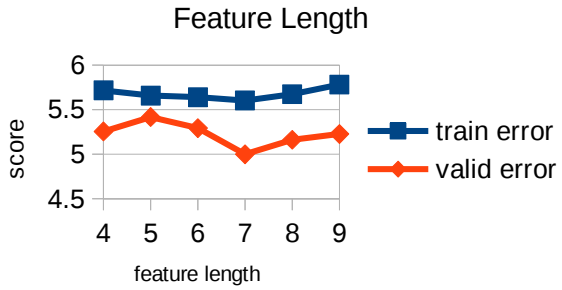
(1) Validation

取大約 20%的 training data 作為 validation set，經過幾次 submit 後找到性質和 public set 相近的 validation set，之後的實驗就以這個 validation set 作為一項指標

(2) Feature selection

I. Feature Length

一開始直接拿前九個小時的 feature 來訓練，不僅時間很久而且分數不好，所以實驗「取前幾個小時的 feature」。圖表 1 為實驗結果，可以看出取前七個小時 performance 最好，因此接下來都取前七小時 feature 來訓練

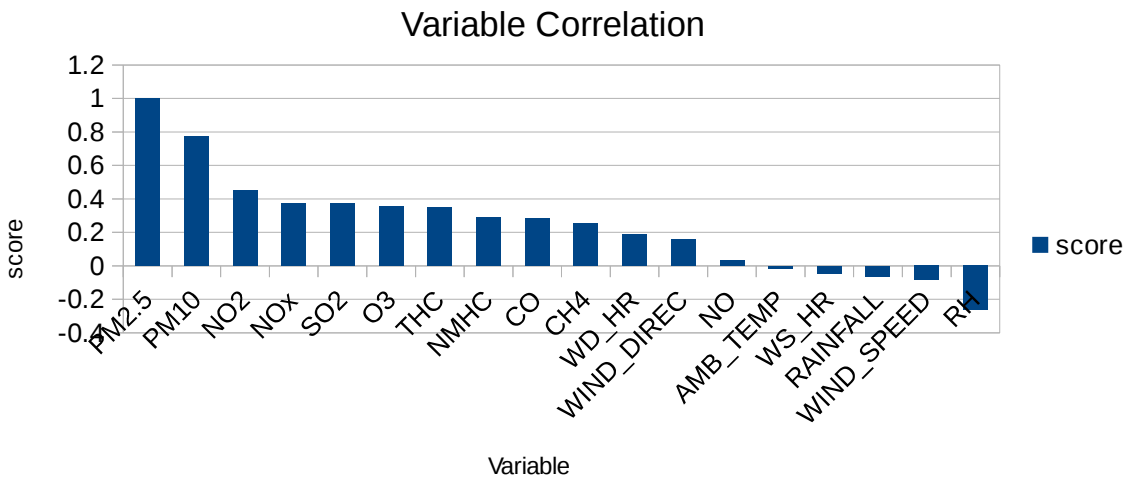


圖表 1

II. Feature selection

取前七小時所有 feature 仍然過多，而且有些 feature 其實是 noise，需要去除這些 feature，下面圖表 2 是每個變數對 PM2.5 的 correlation，對於 linear regression 來說，correlation 愈接近+1 或-1 的變數應該愈有用，因此我選擇 correlation 較高的 PM2.5、PM10、NO2、NOx、SO2、O3、THC、NMHC、CO、WD_HR(NOx=NO+NO2，所以選擇 NO 和 NO2)來進行實驗，實驗各種組合後，

PM2.5、PM10、NO2、NO、SO2、O3、NMHC、CO、WD_HR 這個組合是表現最好的。



圖表 2

III. 平方項

項目	一次項	一次項+二次項
Train error	5.772	5.601
Valid error	5.135	5.032

實驗後發現，加入二次項的 feature 後可以 fit 的更好，接下來實驗都使用一次項+二次項的 feature。

(3) Ensemble

將性質不同的模型的 output 統合起來，產生一組新的 output，通常能得到更好的 performance。因為每個模型都有一些缺陷，透過 ensembling 有機會把各個模型的缺陷消除。

	Model 1	Model 2	Model3	Ensemble result
1	Linear 5.58	Linear 5.62	Linear 5.65	5.52
2	NN 5.60	NN 5.68	Linear 5.52	5.27

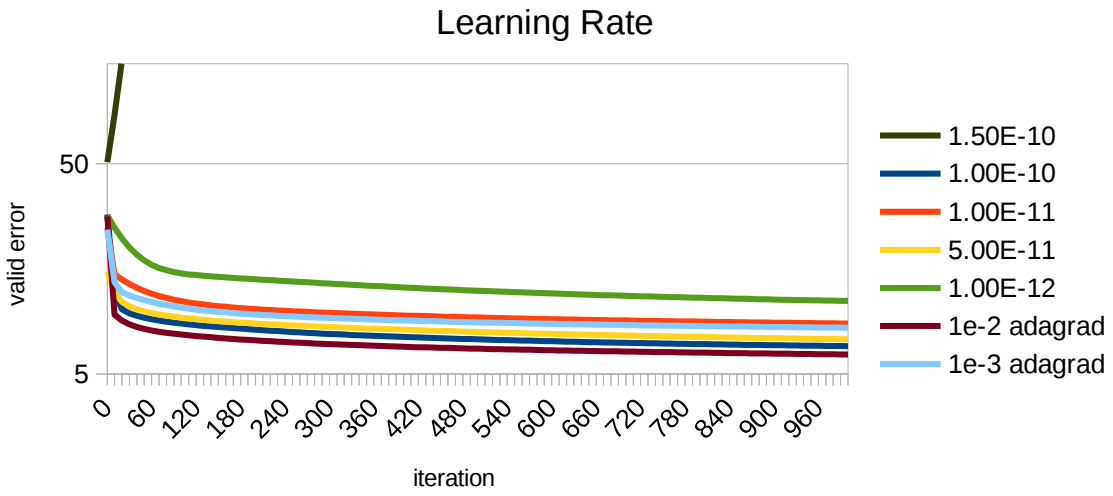
我的 ensemble 方法是單純對數個 output 進行平均，以上是兩次 ensemble 的結果。第一次使用三個 Linear Regression Model 的結果進行平均，結果只有些微進步，但是並不明顯，應該是因為這幾個模型的性質太相近了，因此第二次取了兩個 Neural Network Model 加進來 ensemble，可以看到這兩個模型原本的 performance 比 Linear Regression 還差，但是 ensemble 的結果有很大的進步。

3. Regularization

	L1 regularization		L2 regularization	
alpha	Training set error	Validation error	Training set error	Validation error
0.01	5.789	5.061	5.782	4.994
1	5.944	5.180	5.782	5.001
10	6.412	5.614	5.784	5.013
100	X	X	5.793	5.032
10000	X	X	5.860	5.130

分別做 L1 和 L2 regularization，上表可看出愈強的 regularization 並沒有比較好的效果，可能是由於 w 的值都很小(大部分都 e-2)，因此不太需要 regularization。

4. Learning Rate



圖表 3

見上圖表 3，以不同 learning rate 訓練時 valid error 對 iteration 作圖。可以看到 learning rate 愈大，valid error 下降的愈快，但太大的 learning rate 會讓 gradient descent 失效，如 1.50E-10 的曲線直接往上無法下降；使用 adagrad 時仍然需要調整 learning rate，調整幾次實驗後發現 learning rate=0.01 並使用 adagrad 可以收斂的最快

5. Multi-layer Perceptron

為了做 Model Ensembling，需要嘗試不同性質的模型，嘗試 Logistic Regression 後發現效果不彰，因此轉而實作 Multi-layer Perceptron(MLP)。以下針對 activation function 和 hidden layer size 進行實驗，但由於 Initialization 是隨機的，而 MLP 不像 Linear Regression 是 convex funtion，每次跑程式的結果不盡相同，因此實驗結果僅有約略值。

(1) Activation function

Activation function	Public score	Training set error
relu	5.6	5.2
tanh	~8	4.5

這裡選用 ReLU 和 tanh 函數進行實驗(sigmoid 函數因為 logistic regression 效果不太理想因此沒有進行實驗)。tanh 函數可以對 training set fit 的很好，但 public score 非常高，推測是有嚴重的 overfitting。ReLU 函數是取 max(x, 0)，其實也很接近 linear model，因此在 public score/training set error 都跟 Linear Regression 很相近，但性質比較不同，所以可以進行 ensemble。

(2) Hidden layer size

Hidden layer size	Public score	Training set error
(30,)	5.9	6.4
(30,30)	5.8	5.5
(30,10,5)	5.6	5.2
(30,5,5,3)	5.8	4.9

由於沒有做 neural network 的經驗，因此 hidden layer size 都只有隨機嘗試，經過幾次實驗選用(30,10,5)作為 hidden layer size