Homework 1 Report - PM2.5 Prediction

學號：r06725028 系級：資管碩一 姓名：黃于真

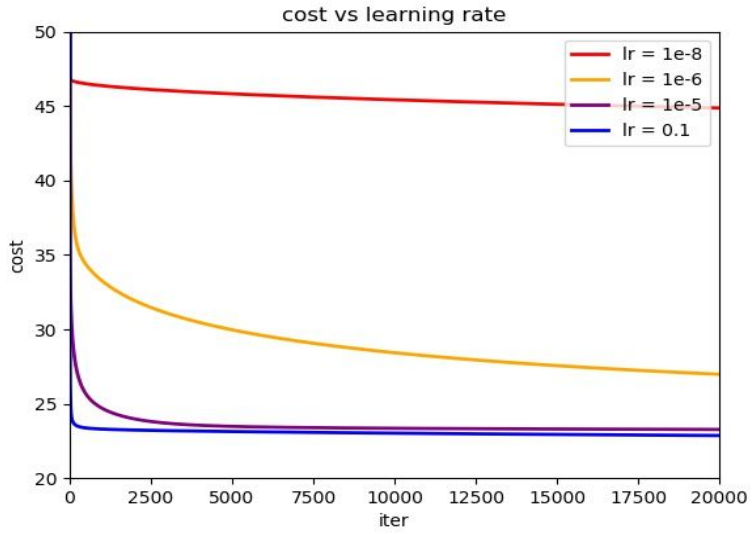
1. (1%) 請分別使用每筆data9小時內所有feature的一次項（含bias項）以及每筆data9小時內PM2.5的一次項（含bias項）進行training，比較並討論這兩種模型的root mean-square error（根據kaggle上的public/private score）。

feature數	9	162
public	9.94188	9.15385
private	10.03043	8.78979

依據public分數，在其他參數相同並訓練了五萬epoch後，發現使用全部特徵的預測效果會比只使用PM2.5特徵的模型結果好，RMSE值差了約0.8，不過這也顯示出只使用PM2.5特徵仍然有一定的預測能力，我想是因為我們要預測的對象就是PM2.5，而其他特徵對於PM2.5的影響其實就已隱含在訓練資料的PM2.5特徵中，而當其他特徵再加入時，影響力更加顯現，也增加模型預測力，所以使得RMSE減少。

在private分數上，發現兩者間的差距更大了，使用162個特徵的模型RMSE減少，但使用9個特徵的模型RMSE卻增加，顯見使用較多的特徵應該有助於減少overfitting，或者說更複雜的模型可以準確地預測出平均來說較少誤差的結果。

2. (2%) 請分別使用至少四種不同數值的learning rate進行training（其他參數需一致），作圖並且討論其收斂過程。



我選擇了le-8、le-6、le-5、0.1這四種learning rate，可以看到learning rate越小，收斂速度越慢，所以紅線幾乎是一直線，cost下降很緩慢，而黃線、紫線、藍線則是在最前幾個iter會有急遽的下降，接著趨於平緩，但learning rate越大，急劇下降的幅度就越大，像是藍線一開始幾乎是貼著y軸的，不過雖然藍線的急劇下降幅度大，但在

2500iter後，cost已降到一定程度，之後其實就和紫線差不多。另外，紅線和黃線的learning rate差了百倍，兩者曲線之間有著相當大的距離，但紫線和藍線learning rate差了萬倍卻相差不大，可見learning rate的增加對於收斂速度的增加，其影響力是邊際遞減的。

3. (1%) 請分別使用至少四種不同數值的regulization parameter λ進行training（其他參數需一至），討論其root mean-square error（根據kaggle上的public/private score）。

lambda	0.001	1.0	5.0	100
public	19.69968	21.77730	14.30385	19.44178
private	19.52329	20.26195	14.05523	19.68971

依據public的分數，在其他參數相同並訓練了五萬epoch後，可以看出不同的lambda值對於RMSE值是有明顯的影響，lambda值太大或太小都會導致RMSE的增加，有趣的是，lambda很大和很少時的結果其實是差不多的，可見lambda值的大小對於RMSE的影響曲線應該是兩端高，中間平坦，只要不取到極端值，結果應該都有一定的效果。當lambda值為100時，實際看輸出檔發現每筆資料預測出的結果幾乎都差不多(約25~30左右)，應該是規範項太大，使得大多特徵的係數都很低，連帶縮小了每筆資料間的差異，因此預測出的結果也就都差不多。而這次實驗中，表現最好的是lambda=5時，數值大於一，應該表示在這次實驗中，影顯曲線平坦的區段其lambda值是偏大的，可見模型中應該還有許多不需要的特徵，所以需要較大影響力的規範項，後續也要進一步做特徵篩選。另外，意外的是，表現最差的是lambda=1，和平常看到的結果不太相同，比lambda=100、0.001還差，實際看預測結果有正有負，預測出的範圍差距蠻大的，猜想是規範項對於參數的學習造成一定的影響，但又不足以去平衡掉不需要的參數，卡在中間才導致結果不好。

在private分數部分，基本上和public分數差不多，比較特別的是只有lambda=100時的分數是增加的，其他則都是減少的，但表現最好的仍是lambda=5時，最差的也同樣是lambda=1時。

4. (1%) 請這次作業你的best_hw1.sh是如何實作的？（e.g. 有無對Data做任何Preprocessing？Features的選用有無任何考量？訓練相關參數的選用有無任何依據？）

經過許多測試後，我使用助教範例程式的資料前處理方法，但不加上每個特徵的平方項和bias項，也不做feature scaling，使用sklearn套件中的f_regression函數來對於162個特徵進行選擇，找出前80個和回歸相關度較高的特徵，最後再使用sklearn套件中的RandomForestRegressor來跑回歸，至於參數選擇部分，則使用gridsearch方法，先切成5個fold來做cross_validation，並且以mse為分數衡量依據，tunning三個主要

參數，分別是min_samples_leaf(1～30)、max_depth(1~ 10)、n_estimators(1～30)，最後做出最好public結果的參數如下圖：

```
(5652, 80)
[15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 39, 40, 41, 42, 43,
45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 72, 73, 74, 75, 76, 77, 78, 79, 8
0, 81, 82, 83, 84, 85, 86, 87, 88, 89, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 1
21, 122, 123, 124, 125]
MSE6 train:460.727
R^2 train:0.667
(260, 80)
{'bootstrap': True, 'criterion': 'mse', 'max_depth': 9, 'max_features': 'auto', 'max_leaf_nodes': None, '
min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 5, 'min_samples_split': 2, '
min_weight_fraction_leaf': 0.0, 'n_estimators': 29, 'n_jobs': 5, 'oob_score': False, 'random_state': 42,
'verbose': 0, 'warm_start': False}
```