Homework 1 Report - PM2.5 Prediction

學號：b04705026 系級：資管三 姓名：林彥廷

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）。

Submission and Description	Private Score	Public Score
report_1_pm25.csv just now by Adam add submission details	8.39272	8.45642
report_1_all.csv a few seconds ago by Adam add submission details	7.39257	7.58944

兩個模型皆使用”Linear Regression”，包含”所有”feature的模型在Public或是Private的分數都優於”只包含PM2.5”的模型。

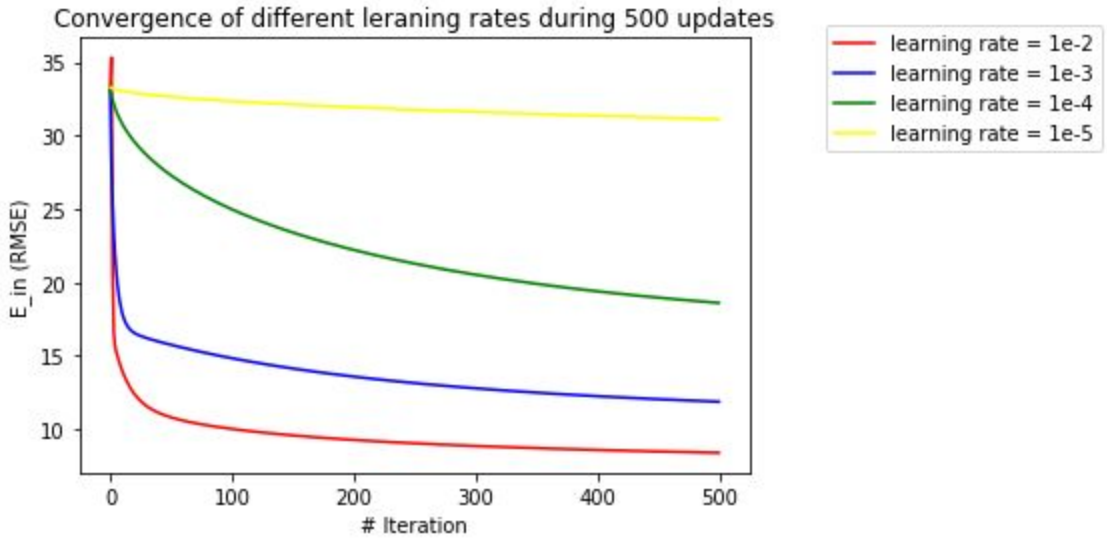
包含所有feature的模型，feature數目也只有163個，遠遠小於data的數量，並不會造成overdetermined的情形，以至於模型表現奇差。所以直覺上來說，feature數量越多越好，但實際上還是要觀察feature之間的統計相關性、以及空氣品質預測的專業知識，來增進模型表現。

以下是觀察資料後探討出的原因。

第一個因素：**只利用PM2.5的歷史資料無法掌握所有控制因素的變異**（下雨、風向、風速）。例如可能前九個小時濃度很高，但是突然下了一場雨，濃度就劇烈下降了。

第二個因素：**PM2.5的資料中有一些極端值**。這是測量儀器的讀數，發生錯誤是很正常的事情，但如果沒有把幾個七百多、六百多的資料點做處理（刪除、改成前後平均），會巨大地影響”只包含PM2.5”的模型的準確度，反之，包含全部Feature的模型，因為參數眾多，相對影響小。

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。



以上模型皆使用Linear regression with all features and adagrad to optimize

Learning rate 越大收斂的越快。雖然各learning rate個別差10倍，但是更新完一定次數的E_in 並不成比例。Learning rate越大，learning rate除以微分值總和就會越大，更新一次會更新相對多、相對"大步"。

因為Linear regression是一個convex的問題，不會有local minimum，朝著gradient的反方向可以走向最佳解，並且藉由adagrad的幫助，可以動態調控走的步伐（更新次數越多步伐越小），不用擔心在谷底跳動，最終可以很接近minimum。

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）。

Submission and Description	Private Score	Public Score
report_3_alpha=10.0.csv 11 minutes ago by Adam add submission details	7.39389	7.57983
report_3_alpha=1.0.csv 11 minutes ago by Adam add submission details	7.39278	7.58063
report_3_alpha=0.001.csv 11 minutes ago by Adam add submission details	7.39266	7.58073
report_3_alpha=0.1.csv 11 minutes ago by Adam add submission details	7.39267	7.58072

分別使用λ= 10 , 1, 0.1, 0.001來進行四次比較（linear regression with all features）

Public score與private score的排序結果完全相反！Public的結果是 λ 越大越好，但是**Private的分數則是 λ "越小越好"**。

以最終結果(Private Score)來說，控制模型複雜度的λ卻是是希望模型比較"複雜"。這個模型雖然包含所有的features但是都只有一次項，並沒有平方項或是交叉項。**對於這個dataset來說，hypothesis set太小了，並不包含optimal model，所以增加regulizor λ 只是讓hypothesis set更小，表現會更差。**

由此可以推論，如果排除一些相關性小的feature，增加模型的hypothesis set（增加平方項、交叉項、高度轉換、rbf等等...），搭配regulizor控制模型複雜度，使用Cross Validation選擇 λ，可能才會優於原本單純的線性模型。

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？）

Data Preprocessing:

這次的資料是氣象局儀器的讀數，可能會有儀器故障、檢修種種錯誤。

最重要的欄位：PM2.5

PM2.5這個欄位有許多outliers，把超過200的資料都改成前後的平均。

其他欄位：

統一把在10個標準差以外的measures改成前後的平均。

Features Transformation:

WS_HR、WD_HR、WIND_DIREC、WIND_SPEED 分別紀錄每小時平均或是每小時最後10分鐘的風速、風向。

風向是一個"方向性"的資料，如果單純丟入regression等同於認為359度 以及0度差異很大的迥異邏輯。

處理方式： 拆解風速成兩個方向

Wind_Speed_X=	cosine(WIND_DIREC) x WIND_SPEED
Wind_Speed_Y=	sine(WIND_DIREC) x WIND_SPEED
Wind_Speed_X_hr=	cosine(WD_HR) x WS_HR
Wind_Speed_Y_hr=	sine(WD_HR) x WS_HR

Model Selection: **Linear Regression with adagrad**

λ Selection: K-fold Cross Validation

一開始跑Ridge Regression用了1e-12 ~ 1e1中100個數當作 λ 進行cross validation，選一個E_val最小的 λ，最終選出了一個數值非常小的 λ，所以乾脆直接使用Linear Regression。