學號：R05944041 系級： 網媒所碩二 姓名：戴長昕

1.記錄誤差值 (RMSE)，討論兩種特徵抽取方法的影響：

特徵選擇	抽全部9小時內的污染源feature的一次項	抽全部9小時內pm2.5的一次項當作feature
Kaggle Public Score	7.26672	7.00600

為了取得足夠訓練資料，將一個月的前 20 天串接，並以 9 小時為單位切割資料，並以下個小時的 PM2.5 為正解，意即全部的訓練資料，扣掉隔月的串接點，共有 5652 筆。

使用 $J(\theta) = \frac{1}{2}\sum_{i}((h_\theta(x^{(i)}) - y^{(i)})^2) + \lambda \sum w^2$ 做為 loss function，使用 gradient descent 計算最佳的參數組 θ 使 $J(\theta)$ 最小，比較抽取9個小時內的18項特徵共 18 * 9 = 162 個一次項特徵加上 bias，和僅只抽取前9個小時的一項 PM2.5 的一次項作為特徵加上 bias，在 learning rate η 0.0001 並經過 10000 個 iteration 後，特徵較少的抽取方法反而在 training set 和 testing set 都得到更佳的結果。

由此可見，訓練數據包含許多冗餘或無關的特徵，這些特徵有可能無助於訓練，甚至造成負面的作用，適度的特徵選擇可以強化模型的學習能力，並簡化模型和縮短訓練時間，甚至能降低造成 overfitting 的風險。

2.將特徵從抽前9小時改成抽前5小時，討論其變化：

特徵選擇	抽全部5小時內的污染源特徵的一次項	抽全部9小時內的污染源特徵的一次項
Kaggle Public Score	7.30521	7.26672

在 learning rate η 0.0001 並經過 10000 個 iteration 後，特徵較多的抽取方法在 training set 和 testing set 都得到更佳的結果，與題

目1得到的結果相反，卻沒有差距太多，推測可能是某些特徵具有較長時間性的影像。

3.Regularization on all the weight：

λ	0.1	0.01	0.001	0.0001
Kaggle Public Score	7.0060	7.06137	7.19834	7.11648

使用前9個小時的PM2.5一次項作為特徵，並使用 L2 正規化，可以發現 regularization lambda 的變化對於預測結果的影響不大，在training set的 cross validation 也沒有特別突出。

4.Linear Regression Normal Equation的求解與推導：

最終目的是求解 $X \cdot \theta = y$，假設常量 y 不屬於線性變換X的值域 $Column(X)$，即 $X \cdot \theta = y$ 無解，則求最佳近似解 $\hat{\theta}$ ，並使得誤差向量 $e = y - X\hat{\theta}$ 有最小的長度平方，意即 $Minimize_{\hat{\theta}} \| y - X\hat{\theta} \|$。

$\hat{\theta} \xrightarrow{X} p$，若最小誤差發生在$e = y - p$與 p 正交 ，p 即為 y 在 $Column(X)$ 的正交投影。令 P 為正交投影至 $Column(X)$ 的轉換矩陣，是 $n \times n$ 且滿足$P^2 = P = P^T$，要求出投影矩陣P，以投影矩陣性質可知 $P^2 y = Py = p$，則 $P(y - p) = p - p = 0$。 從上面可以知道 $y - p$ 屬於 $N(P)$ ，則 $e = y - p = y - X\hat{\theta}$ 屬於 $Column(X)$ 的正交補$Column(X)^{\perp} = N(X^T)$ ，則可以得到 $X^T e = X^T(y - X\hat{\theta}) = 0$，能夠改寫成 $X^T X\hat{\theta} = x^T y$ 。

Normal Equation中的normal即為垂直，指e正交於 $Column(X)$，假設 $X^T X$可逆，即 $Rank(X) = Rank(X^T X)$，故存在最小平方近似解 $\hat{\theta} = (X^T X)^{-1} X^T y$，答案為 © 。