學號：T05902136 系級： 資工一 姓名：Jenny Zhang(張臻凝)
母語：英語

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
答：
At first, I chose all features to be fed into my program. After looking at the output that it produced and comparing it with the result I envisioned, I realized that the difference is quite big. For this reason, I decided to choose selected features based on my trivial knowledge and the accuracy on Kaggle.
AMB_TEMP: From obeserving the training data, when the temperature is high, the PM2.5 value tends to be high too.
O3：From research online, I find out that O3 and particles are chemically coupled.
PM2.5：There is a direct relationship between PM2.5 in the previous hours and the current hour. When we have more PM2.5 accumulated, the PM2.5 in the next hour will usually be high.
PM10：Due to external factors, it is possible that PM10 will break down and thereby form PM2.5.
WIND_SPEED：When there is a higher wind speed, the particles are more likely to be blown apart or away.

2.請作圖比較不同訓練資料量對於PM2.5預測準確率的影響
答：
(以下是用Polynomial Regression跑出來的準確率)

資料量	準確率	變異量數
全12個月的	5.75784	413.42
前8個月	6.04036	410.05
前4個月	6.30732	419.83

From the above table, the amount of training data seems to be related to the accuracy of test results. With an extra of 4 months of data, the accuracy improves by approximately 0.3. Even though variance of all 12 months is a bit larger than variance of first 8 months, the accuracy still improves. This might be because there is a better bias-variance trade-off.

3. 請比較不同複雜度的模型對於PM2.5預測準確率的影響
答：

複雜度	準確率
Polynomial Regression(with quadratic terms)	5.75784

Polynomial Regression(without quadratic terms)	5.79880
Linear Regression	5.84572
Polynomial Regression (all features all hours, with square terms)	5.97192

From the above table, we see that the best result is from a polynomial regression model with quadratic terms. By adding in quadratic terms, it gives information such as the expected direction and change of slope with respect to changes in x. In the last row, even though a polynomial regression model is used, the accuracy is the lowest. This indicates that more features does not necessarily mean better result. With a greater variety of data introduced, more noises and unrelated information will be brought into the training process.

4. 請討論正規化(regularization)對於PM2.5預測準確率的影響

答：

(以下是用Polynomial Regression跑出來的準確率)

有無正規化？	準確率
無	5.75784
有 (Tikhonov regularization with α = 1)	5.73926

Regularization penalizes the loss function and is an attempt to solve overfitting. In this case, adding a regularization term stabilizes the problem and smooths the function which helps the accuracy of our result.

5. 在線性回歸問題中，假設有 N 筆訓練資料，每筆訓練資料的特徵 (feature) 為一向量 x^n，其標註(label)為一存量 y^n，模型參數為一向量w (此處忽略偏權值 b)，則線性回歸的損失函數(loss function)為 $\sum_{n=1}^{N}(y^n - w \cdot x^n)^2$ 。若將所有訓練資料的特徵值以矩陣 X = [x^1 x^2 … x^N] 表示，所有訓練資料的標註以向量 y = [y^1 y^2 … y^N]^{T}表示，請以 X 和 y 表示可以最小化損失函數的向量 w 。

答：

$$\sum_{n=1}^{N}(y^n - w \cdot x^n)^2 = (y - wX)^T(y - wX)$$

we need to find w such that the summation above is minimized, which is:

$$w = (X^T X)^{-1} X^T y \quad \text{(ordinary least squares estimator)}$$