

# Machine Learning

## Homework #1 - Predict PM2.5

B02901080 電機四 董皓文

### Linear Regression Function

Linear Regression Model

$$y = b + \theta \cdot X$$

$X$  為前九小時所有 data (9x18)

$b$  為 bias

$y$  為第十個小時的 PM2.5

$\theta$  為 9x18 維變數

Method of updating coefficient

SGD:

$$\theta_t = w_{t-1} - \eta g_t$$

AdaGrad:

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\sum_{i=0}^t (g_i)^2 + \varepsilon}} g_t$$

AdaDelta:

$$\begin{aligned} E[g^2]_t &= \gamma E[g^2]_{t-1} + (1 - \gamma) \\ E[\Delta\theta^2]_t &= \gamma E[\Delta\theta^2]_{t-1} + (1 - \gamma) \Delta\theta_t^2 \\ \theta_t &= \theta_{t-1} - \frac{\sqrt{E[\Delta\theta^2]_{t-1} + \varepsilon}}{\sqrt{E[g^2]_t + \varepsilon}} g_t \end{aligned}$$

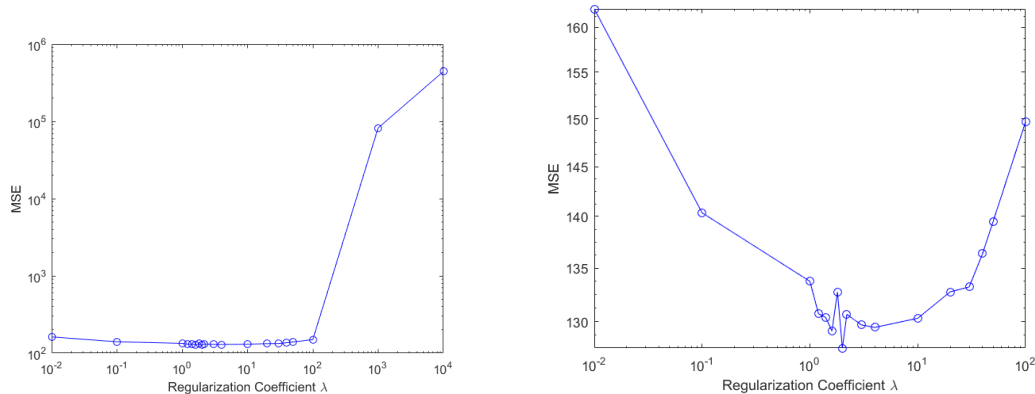
### Method

Feature: 前九小時所有 data 作為  $X$ ，第十個小時的 PM2.5 作為  $y$

Training data: train.csv 中所有連續十小時的資料(共 5652 組)

比較 VanillaGD、adaGrad、adaDelta 後，發現 adaGrad 收斂最快，以 adaGrad 進行 Cross Validation 找出最佳的 regulation coefficient  $\lambda$ 。

## Discussion on Regularization



上圖為不同  $\lambda$  時的 cross validation 的 MSE，可見  $\lambda$  太小時，有 underfitting 的問題；但  $\lambda$  太大時，有 overfitting 的問題；可發現  $\lambda$  在  $10 \sim 100$  的數量級有最好的 fitting 結果。

## Discussion on Learning Rate

在 SGD 中，Learning rate 較難調整，微微調高就可能會爆掉，調低收斂的速度又太慢。

在 AdaGrad 中，Learning rate 的起始值可以調得較大，AdaGrad 的演算法會自動調整其 learning rate，使 learning rate 逐漸 decay。此方法在 training 初期可以迅速找到局部極值附近，但在 training 後期受到 AdaGrad 設計的懲罰機制，更新得較慢。

AdaDelta 是 Adagrad 的改良，改善 AdaGrad 在 training 後期因懲罰機制而更新較慢的情況

## Other

```
0 MSE: 0.178413652215
100 MSE: 0.114696579144
200 MSE: 0.0528363075007
300 MSE: 0.0528363158056
400 MSE: 0.0528363260494
500 MSE: 0.0528363261336
600 MSE: 0.0528363261342
700 MSE: 0.0528363261342
800 MSE: 0.0528363261342
```

有觀察到 Loss function 在更新後變大的情況~~~