Least Square & Gradient Descent

Loss/Cost Function: $J = \Sigma(\hat{y}_i - y_i)^2$ >> $J(\theta_i) = \sum(h(\theta_i) - y_i)^2$

>> to optimize by minimizing loss function

Least Square >> find min J

Gradient Descent >> $J(\theta_0, \theta_1, \cdots, \theta_n) = ½m \sum_{j=0}^{m} \left(h_\theta\left(x_0^{(j)}, x_1^{(j)}, \ldots, x_n^{(j)}\right) - y_i\right)^2$

Take partial derivative >> $\frac{\partial}{\partial_{\theta_i}} J(\theta_0, \theta_1 \ldots, \theta_n)$

>> set a step length α

Then $\alpha \frac{\partial}{\partial \theta_i} J(\theta_0, \Theta_1, .., \Theta_n)$

$$\alpha \frac{\partial}{\partial \theta_i} J(\theta_0, \Theta_1, .., \Theta_n) = \alpha \frac{1}{m} \Sigma \left(h_\theta\left(x_0^{(j)}, x_1^{(j)}, \ldots, x_n^{(j)}\right) - y_i\right) x_i^{(j)}$$

Each time after a step, we need to update our θ_i

$$\theta_i = \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta_0, \Theta_1, .., \Theta_n)$$

That is $\theta_i = \theta_i - \alpha \frac{1}{m} \Sigma \left(h_\theta\left(x_0^{(j)}, x_1^{(j)}, \ldots, x_n^{(j)}\right) - y_i\right) x_i^{(j)}$

Which means, the current step direction depends on the samples, and $\alpha \frac{1}{m}$

Can also be seen as a constant

In machine learning algorithms, we usually use matrix method to do calculations

Therefore, here we regard the former functions as a matrix function with

Y.hats as a matrix transformed by sample matrix **X** and parameter vector **θ**, that is **Y.hats** = $h_{\boldsymbol{\theta}}$(**X**) = **Xθ**

Then, the loss function becomes $J(\boldsymbol{\theta}) = (\boldsymbol{X}\theta - \boldsymbol{Y})^T(\boldsymbol{X\theta} - \boldsymbol{Y})$

We take the gradients/ partial derivatives for the J(**θ**), then get: $\frac{\partial}{\partial \theta_i} J(\theta) = \boldsymbol{X}^T(\boldsymbol{X}\theta - \boldsymbol{Y})$

With the update of θ, the matrix expression function can be written as $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \boldsymbol{X}^T(\boldsymbol{X\theta} - \boldsymbol{Y})$

Notes for Gradient Descent to Optimize models

- The choice of step length
- The choice of initial value
- Regularization

Gradient Descent Family

>> Batch GD, Stochastic GD, Mini Batch GD