Linear Regression

our dataset format is $(x^{(i)}, y^{(i)})$, and we have dataset size m, so where i is ith data, and for $x^{(i)}$, it may have n features, that to say, for any x,

$$x = \{x_0, x_1, ..., x_n\}$$

Note that dimension of x is $n + 1$, that is to simplify the calculation, x_0 often equal to 1, to make θ_0 a constant term.
To fit LR on our dataset, we assume:

$$h_\theta = \theta_0 x_0 + \theta_1 x_1 + .., + \theta_n x_n = \theta^T x$$

and $h_\theta^{(i)}$ is LR predicted output for $x^{(i)}$, there $x_0 = 1$.

So how do we know our predition is reliable? we introduce cost function here.

$$J(\theta) = \frac{1}{m} \sum_{i=0}^{m} (h_\theta(x)^{(i)} - y^{(i)})^2$$

the larger the $J(\theta)$, the more inaccurate the prediction result is. so we must choose a $\theta = [\theta_0, \theta_1, ..., \theta n]$ to minimise $J(\theta)$

Gradient Descent

we update θ to minimise $J(\theta)$ by

$$\theta_j = \theta_j - \alpha \cdot \frac{1}{m} \frac{\partial J(\theta)}{\partial \theta_j}$$

and when $m = 1$

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_i} &= 2 \cdot \frac{1}{2} (h_\theta - y) \cdot \frac{\partial}{\partial \theta j} (h_\theta x - y) \\
&= 2 \cdot \frac{1}{2} (h_\theta - y) \frac{\partial}{\partial \theta j} (\sum_{i=0}^{m} \theta_i x_i - y) \\
&= (h_\theta - y) x_j
\end{aligned}$$

so now

$$\theta_j = \theta_j - \alpha \cdot \frac{1}{m}(h_\theta - y)x_j$$

it is called gradient descent.

Normal equation

it is other way to get appropriate θ.

$$\theta = (X^T X)^{-1} \cdot (X^T y)$$

if $X^T X$ is noninvertible, it may caused by:

- Redundant features, where two features are very closely related (i.e. they are linearly dependent)
- Too many features (e.g. m ≤ n). In this case, delete some features or use "regularization" (to be explained in a later lesson).

Reference

Mechine Learning