# LINEAR

## REGRESSION

→ In linear regression, we find a set of numbers called parameters from the training data and use it to predict values for testing data.

→ $h_\theta(x) = \theta^T x = \sum_i^n \theta_i x_i$ ; $x_0 = 1$

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \end{bmatrix}$ $x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \end{bmatrix}$

$m \rightarrow$ no. of training examples.

$n \rightarrow$ no. of features.

→ Cost function;

$$\underbrace{J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left[ h_\theta(x^{(i)}) - y^{(i)} \right]^2}_{\text{Has to be minized.}}$$

→ So, which can be done by

Gradient - descent:

Start with some $\theta$ ;

then

$$\theta_j := \theta_j - \overbrace{\alpha}^{\text{learning rate}} \frac{\partial}{\partial \theta_j} J(\theta)$$

$$\Leftrightarrow \theta_j := \theta_j - \alpha \left[ h_\theta(x) - y \right] x_j$$

For single training ex

→ For multiple training examples m;

Repeat {
$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} \left[ h_\theta(x^{(i)}) - y^{(i)} \right] x_j^{(i)}$$
for $j = 0, 1, 2 \ldots n$.

}

→ Stochastic Gradient descent:

Repeat {

For $j=1$ to $m$ {

$$\theta_j := \theta_j - \alpha \left( h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

}

}

→ Other Algorithms:

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

→ Why least squares?

let us assume

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$$

Taken to be gaussian

"from central limit theorem"

$$\varepsilon^{(i)} \sim N(0, \sigma^2)$$

$$P(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( - \frac{(\varepsilon^{(i)})^2}{2\sigma^2} \right)$$

$\Rightarrow$ On writing $\varepsilon^{(i)}$ as $y^{(i)} - \theta^T x^{(i)}$

$$p(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

i.e. $\boxed{y^{(i)} \mid x^{(i)}; \theta \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)}$

$$\mathcal{L}(\theta) = p(\vec{y} \mid x; \theta)$$

$$= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$\ell(\theta) = \log \mathcal{L}(\theta)$$

$$= \sum_{i=1}^{m} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log(\exp \cdots)\right]$$

$$= m \log \frac{1}{\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{m} -\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}$$

So, $\mathcal{L}(\theta) \longrightarrow$ Maximised

$\Rightarrow \ell(\theta) \longrightarrow$ Maximised

$$\Rightarrow \underbrace{\frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2}_{\text{cost function.}} \longrightarrow \text{Minimised}$$