Microsoft Whiteboard

Mehul S Raval

Regularisation - Linear Regression

Inputs

$x_1^1 \quad x_1^2 \quad x_1^3 \quad \cdots \cdots \quad x_1^n$ output y_1

x_2

x_3

m example

x_m $\qquad y_m$

$n+$

Training pairs

$\mathbb{R}^{m \times n+1}$

$f: \mathbb{R}^{n+1} \longrightarrow \mathbb{R} = h_\theta(x)$

$h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1 + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$i = 1, \ldots m$

Loss function

Cost function

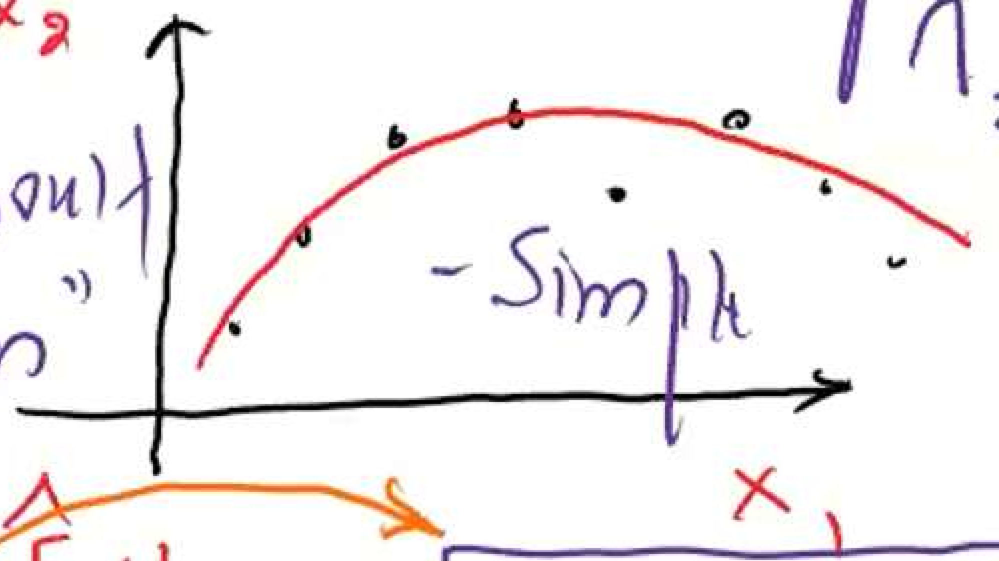
Microsoft Whiteboard

Training loss ≈ 0

$\mathbb{R}^{m \times n+1}$

Generalisation

prediction should fit "unknown" data well



$\boxed{\text{Training}} = \text{Test}$

$\lambda = 0$

$$x^0 = 1$$

$$h_\theta(x) = \boxed{\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4}$$

$$f(x) = h_\theta(x) = \boxed{\theta_0 + \theta_1 x + \theta_2 x^2}$$

"Better fit"

"fit" to data

$$f(x) = h_\theta(x) \approx y$$

Loss ≈ 0

$$\theta_3 \approx 0, \quad \theta_4 \approx 0$$

Training error = Testing error

Microsoft Whiteboard

Ahmedabad University
Mehul S Raval

$x^{(0)} = 1$

$$\boxed{h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2} + \cancel{\theta_3 x^3} + \cancel{\theta_4 x^4} \xrightarrow{\text{kill para.}} \theta_0 + \theta_1 x + \theta_2 x^2$$

Regularisation

~~Kill features~~

Information loss?

$\boxed{n = 100}$

Retain all features

$$\sum_{i=1}^{n+1} \theta_i x^i$$

Which features are important?

$m < n$

example # features

Cost / loss / Energy

Goal

$$\min_{\theta} J(\theta) = \left\{ \left[\frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \right] + 10000\,\theta_3^2 + 10000\,\theta_4^2 \right\}$$

$\boxed{\theta_3 \approx 0, \quad \theta_4 \approx 0}$

"Dimensionality Reduction"

Microsoft Whiteboard

$$\sum_{i=1}^{n+1} x^{(i)} \theta_i$$

Regularisation

Regularisation term

$$\min_{\theta} \frac{\partial J(\theta)}{\partial \theta_j} = \left[\frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right] \quad \|\cdot\|^2$$

Error

~~Data~~ term

$$\min \left[A + \lambda B \right]$$

R_{emp} "fit well the training data"

?

Regularisation parameter

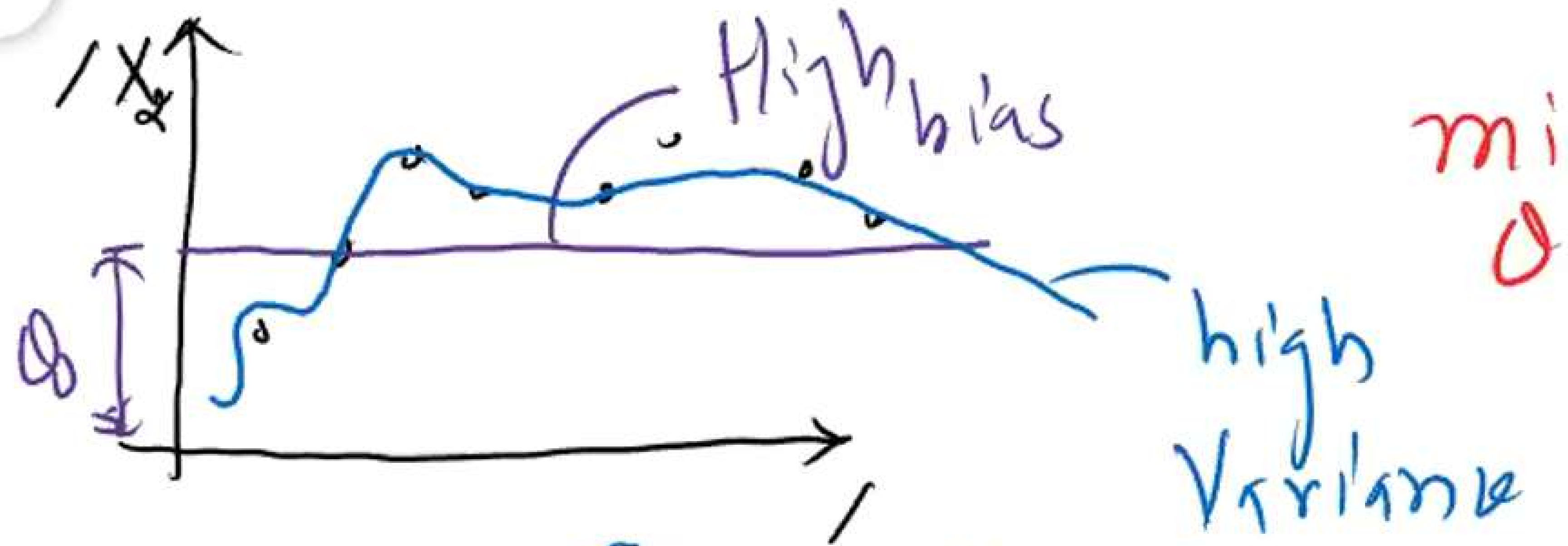
$$J(\theta) = A + \lambda B$$

Keep θ_j = small

Simple hypothesis

Microsoft Whiteboard

Mehul S Raval



$$h_\theta(x) = \theta_0 + \cancel{\theta_1 x} + \cancel{\theta_2 x^2} + \cancel{\theta_3 x^3}$$

$$\min_\theta J(\theta)$$

$$\boxed{\min_\theta \left[A + \boxed{\lambda B} \right]}$$

Keep θ small

λ: "Very large"

"Very small"

$$\sum_{i=1}^{n}$$

Hyper parameter [Regularization parameter]

$\boxed{\lambda: \text{Very Small}}$

Microsoft Whiteboard

Mehul S Raval

BATCH

Gradient Descent

$$J(\theta) = \min_{\theta} \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \boxed{\lambda \sum_{j=1}^{n} \theta_j^2}\right]$$

Learning rate

$$\boxed{\theta_{new} = \theta_{old} - \alpha \frac{\partial J(\theta)}{\partial \theta_j}}$$ Derivative

$x_0^{(i)} = 1$

$\theta_{old} - \theta_{new} < \epsilon$

iterations

$$\frac{\partial J(\theta)}{\partial \theta_j} = 0$$

Repeat till Convergence

{

$$\theta_0 := \theta_0 - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0^{(i)}$$

$$\theta_j := \theta_j - \frac{\alpha}{m}\boxed{\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}} - \boxed{\frac{\lambda}{m}\theta_j}$$

}

$$\boxed{1 - \frac{\alpha\lambda}{m}} < 1$$

0.9

$$\boxed{\theta_j := \theta_j\left(1 - \frac{\alpha\lambda}{m}\right)} - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}$$