CS 307 Lecture 4

Linear Regression Recap

Training set $= \{(\underbrace{x^{(i)}}_{\in \mathbb{R}^{d+1}}, \overbrace{y^{(i)}}^{\in \mathbb{R}})\}_{i=1}^{n}$

Goal is identify $h_\theta : \mathbb{R}^{d+1} \to \mathbb{R}$ where

$$h_\theta(x) = \theta_0 \underbrace{x_0}_{x_0 = 1} + \theta_1 x_1 + \dots + \theta_d x_d.$$

$$= x^T \theta$$

such that h_θ minimizes

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left((x^{(i)})^T \theta - y^{(i)} \right)^2$$

$$= \frac{1}{2n} \| X\theta - y \|_2^2 = \frac{1}{2n} (X\theta - y)^T (X\theta - y)$$

$$1 = x_0^{(i)}, \qquad X = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & \cdots & x_d^{(1)} \\ \vdots & & \vdots & \\ x_0^{(i)} & \cdots & \cdots & x_d^{(i)} \\ \vdots & & & \end{bmatrix}, \qquad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$$\nabla_\theta J(\theta) = \frac{1}{n} X^T (X\theta - y).$$

Optimal θ is obtained when $\nabla_\theta J(\theta) = 0$

i.e. $X^T X \theta = X^T y$ (normal equations)

<u>Proposition</u>: If θ^* satisfies $X^T X \theta^* = X^T y$ then

$$J(\theta^*) = \min_\theta J(\theta).$$

One solution to the normal equations is

$$\theta_{OLS} = \underbrace{(X^T X)^{+}}_{\text{Moore Penrose pseudo inverse}} X^T y$$

← ordinary least squares sol.

$$J(\theta) = \frac{1}{2n}(X\theta - y)^T (X\theta - y)$$

$$\nabla_\theta J(\theta) = \frac{1}{n} X^T (X\theta - y)$$

$$[\nabla_\theta J(\theta)]_j = \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)}\left(\sum_{k=0}^{d} x_k^{(i)}\theta_k - y^{(i)}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

$$\begin{bmatrix} x_0^{(0)} & & x_0^{(i)} & \\ x_1^{(0)} & & \cdot & \\ \cdot & \cdots & \cdot & \cdots \\ \cdot & & \cdot & \\ x_d^{(0)} & & x_d^{(i)} & \end{bmatrix} = X^T$$

Gradient descent:

Batch Gradient descent.

θ = initial assignment ($\theta = 0$)

Repeat until convergence — θ doesn't change by much.

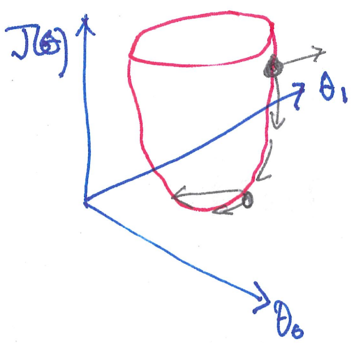
learning rate (hyperparameters)

$$\theta = \theta - \alpha \nabla_\theta J(\theta)$$

$$= \theta - \alpha \frac{1}{n} X^T (X\theta - y)$$

Each j

$$\theta_j = \theta_j - \underbrace{\alpha \frac{1}{n}}_{\alpha'} \sum_{i=1}^{n} x_j^{(i)}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$



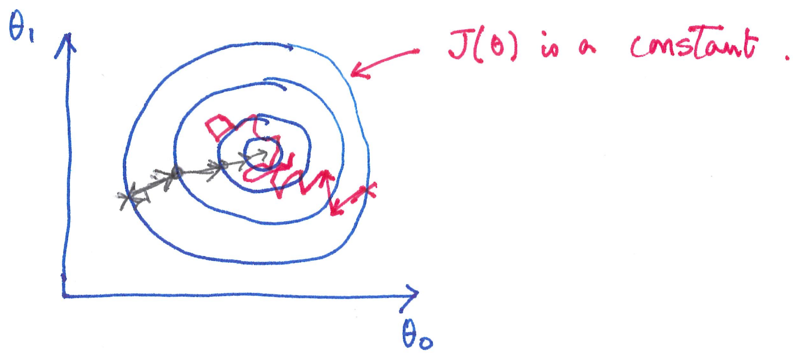
Stochastic Gradient Descent.

θ = initial assignments

Repeat

for $i = 1$ to n

$$\theta_j = \theta_j - \underbrace{\alpha \frac{1}{n}}_{\alpha'} \left(x_j^{(i)} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \right)$$



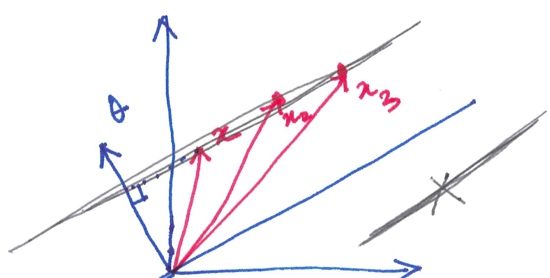
Linear Classification:

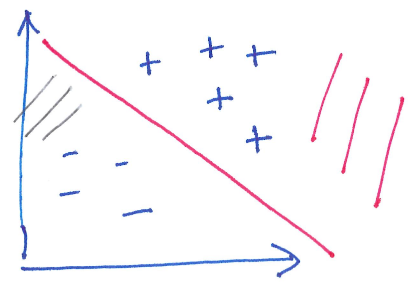
Predicting one out of 2 outputs (Binary classification).

Training set $= \{(\underbrace{x^{(i)}}_{\in \mathbb{R}^d}, \underbrace{y^{(i)}}_{\in \{-1,+1\}})\}_{i=1}^{n}$.

Goal: Compute a hypothesis $h_\theta : \mathbb{R}^d \to \{-1,+1\}$

$$\cancel{h_\theta(x) = 1[(x^T\theta) \geq 0]}$$





$x^T\theta = \underbrace{(\text{length of the projection along } \theta)}\,\|\theta\|_2$

$x^T\theta = c$

$x^T\theta = 0$

Goal: Compute a hypothesis $h_\theta : \mathbb{R}^d \to \{-1,+1\}$

$$h_\theta(x) = \begin{cases} +1 & x^T\theta > 0 \\ -1 & \underbrace{x^T\theta}_{\theta_0 1 + \theta_1 x_1 + \cdots \theta_d x_d} \leq 0 \end{cases}$$

0-1 Loss function

$$L(h_\theta, x, y) = \begin{cases} 1 & \text{if } y \neq h_\theta(x) \\ 0 & \text{o.w.} \end{cases}$$

$$= 1[y \neq h_\theta(x)]$$

$1(b) = \begin{cases} 1 & \text{if } b \text{ true} \\ 0 & \text{o.w} \end{cases}$

Assuming $y \in \{+1,-1\}$ and $h_\theta(x)$'s sign is prediction.

$$L(h_\theta, x, y) = 1[y \cdot h_\theta(x) > 0]$$