Lecture 1 : Linear Regression I

Example : Predicts level of PSA from various measurements on prostate

$(x_1, y_1)\ (x_2, y_2)\ \ldots\ (x_N, y_N) \rightarrow$ Training Data

$y_1, y_2, \ldots y_N$ are known, $y_i \in \mathbb{R}$

$x_1, x_2, \ldots x_N$ are measurements on prostate

$x_i \in \mathbb{R}^d$

Example : Netflix - (user, movie, rating) triplets

Jeff — Sorry to bother you — $5 ☆

$1MM Netflix

Goal : Predict y for a <u>new</u> x

Example : Predict whether email is spam or not

X = set of emails = $\{x_1, x_2 \ldots x_N\}, x_i \in \mathbb{R}^d$

$Y = \{\text{spam, normal}\}$

When Y is categorical - Classification

When Y is real-valued (continuous) - Regression

Regression Problem

Given (x_i, y_i), $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $i = 1, 2, 3, \ldots N$

$$x = \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(d) \end{bmatrix}$$

Prediction (linear): $w_i \in \mathbb{R}$

$$y(x) = w_0 + w_1 x(1) + w_2 x(2) + \ldots + w_d x(d).$$

$$= w_0 + \bar{w}^T x, \quad \bar{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$= w^T \bar{x}$$

where $w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$, $\bar{x} = \begin{bmatrix} 1 \\ x(1) \\ x(2) \\ \vdots \\ x(d) \end{bmatrix}$. $w, \bar{x} \in \mathbb{R}^{d+1}$

$\bar{x}_1^T w \approx y_1 \rightarrow (\bar{x}_1^T w - y_1)^2$ is small

$\bar{x}_2^T w \approx y_2 \rightarrow (\bar{x}_2^T w - y_2)^2$ " "

$\bar{x}_N^T w \approx y_N \rightarrow (\bar{x}_N^T w - y_N)^2$ is small

$$F(w) = \frac{1}{2} \sum_{i}^{N} (\bar{x}_i^T w - y_i)^2$$. Minimize $F(w)$

$$X = \begin{matrix} \\ d+1 \end{matrix} \overset{N}{\begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_N \end{bmatrix}} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1(1) & x_2(1) & & x_N(1) \\ x_1(2) & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ x_1(d) & x_2(d) & \ldots & x_N(d) \end{bmatrix}$$

$X \in \mathbb{R}^{(d+1) \times N}$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad y \in \mathbb{R}^N$$

$$F(w) = \frac{1}{2}\|X^T w - y\|_2^2$$ - Least Squares Objective

$$= \frac{1}{2}(X^T w - y)^T (X^T w - y)$$

$$= \frac{1}{2}(w^T X - y^T)(X^T w - y)$$

$$= \frac{1}{2}(w^T X X^T w - 2 y^T X^T w + y^T y)$$

$$\nabla_w F(w) = X X^T w - X y$$

$$w^* = \arg\min F(w)$$

$$\nabla F(w^*) = 0$$

$$X X^T w^* - X y = 0$$

$$\boxed{(X X^T) w^* = X y}$$

$X X^T \in \mathbb{R}^{(d+1)\times(d+1)}$

$w^* = (X X^T)^{-1} X y$ - ok if $X X^T$ is non-singular

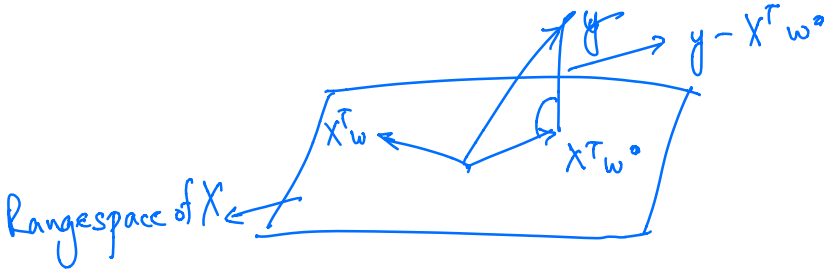
Normal Equations

Geometric View

$$X^T = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \\ 1 & x_N^T \end{bmatrix} = \begin{bmatrix} \bar{x}_1^T \\ \bar{x}_2^T \\ \vdots \\ \bar{x}_N^T \end{bmatrix}$$

$X^T w \approx y$

$X^T w$ ← Linear Combination of columns of X^T



$$X^T w \perp y - X^T w^* \quad \forall w$$

$$(X^T w)^T (y - X^T w^*) = 0$$

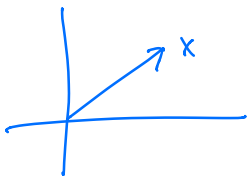
$$\Rightarrow \quad w^T (Xy - X X^T w^*) = 0$$

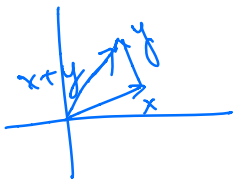
$$\Rightarrow \quad \boxed{X X^T w^* = X y}$$

Normal Equations (again! but derived from a geometric viewpoint)

Linear Algebra Background.

$$x \in \mathbb{R}^d, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}, \quad x^T = [x_1, x_2 \dots x_d]$$





Vector Norms $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_+$

1. $\|x\| \geq 0$ & $\|x\| = 0 \Leftrightarrow x = 0$ — positivity
2. $\|\alpha x\| = |\alpha| \, \|x\|$ — Homogeneity
3. $\|x+y\| \leq \|x\| + \|y\|$ — Triangle inequality

$$\|x\|_p = \left(|x_1|^p + |x_2|^p + \dots |x_d|^p \right)^{1/p} \quad \text{— } L_p \text{ norm}$$

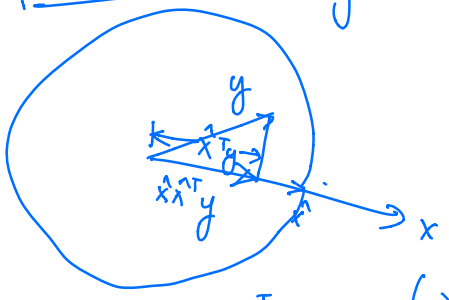
$$\|x\|_1 = |x_1| + |x_2| + \dots + |x_d| \quad \text{— } L_1 \text{ norm}$$

$$\|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \dots |x_d|^2} \quad \text{— } L_2 \text{ norm}$$

$$\|x\|_\infty = \max_{1 \leq i \leq d} |x_i| \quad \text{— } L_\infty \text{ norm}$$

$\hat{x} = \frac{x}{\|x\|_2}$ — unit vector in direction of x

$\hat{x}\hat{x}^T$ → $d \times d$ matrix → orthogonal projector onto the vector x (or $\hat{x}$)



$$(\hat{x}\hat{x}^T)y = \hat{x}(\hat{x}^T y)$$

($\hat{x}$: unit vector; $\hat{x}^T y$: scalar)

$$\hat{x}^T y = \left(\frac{x}{\|x\|}\right)^T y = \frac{x^T y}{\|x\|}$$

$$-1 \leq \cos\theta = \frac{x^T y}{\|x\| \cdot \|y\|} \leq 1 \Rightarrow |x^T y| \leq \|x\| \cdot \|y\|$$

(Cauchy Schwarz Inequality)

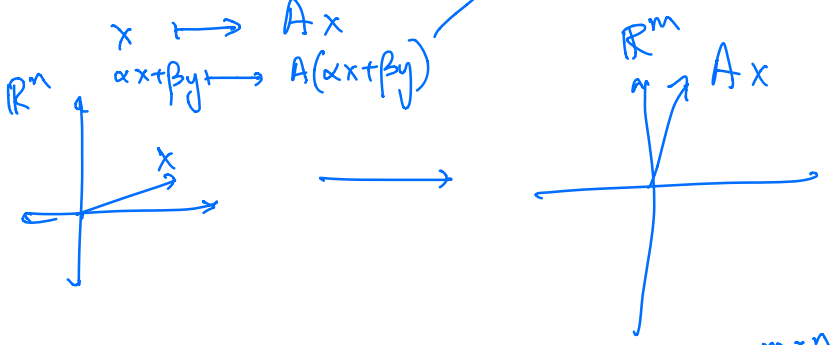
$x \perp y \iff x^T y = 0$

$A \in \mathbb{R}^{m \times n}$

$A : \mathbb{R}^n \to \mathbb{R}^m$

$x \mapsto Ax$

$\alpha x + \beta y \mapsto A(\alpha x + \beta y) = \alpha Ax + \beta Ay$



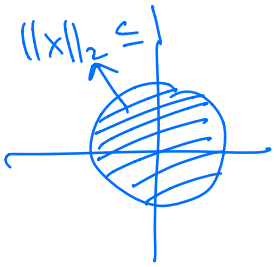
Matrix Norms, $A \in \mathbb{R}^{m \times n}$, $\|\cdot\| : \mathbb{R}^{m \times n} \to \mathbb{R}_+$

1. $\|A\| \geq 0$ & $\|A\| = 0 \iff A = 0$
2. $\|\alpha A\| = |\alpha| \cdot \|A\|$
3. $\|A + B\| \leq \|A\| + \|B\|$

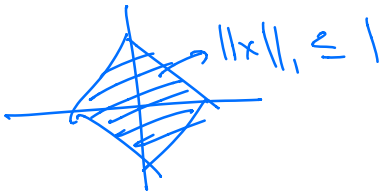
$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2}$ → Frobenius Norm

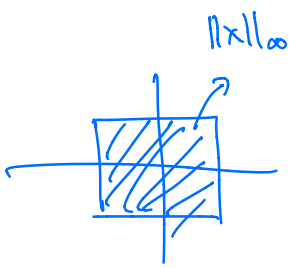
$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2 \leq 1} \|Ax\|_2 = \sigma_1$

σ_1 ↓ maximum singular value of A



$\|A\|_1 = \max_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{\|x\|_1 \leq 1} \|Ax\|_1$





$$\|A\|_\infty = \max_{\|x\|_\infty \le 1} \|Ax\|_\infty$$

$$\|A\|_{p,q} = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_q}$$

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_1 \text{ (max sing. value)}$$

$$\min_{x \neq 0} \frac{\|Ax\|}{\|x\|_2} = \sigma_n \text{ minimum singular value of } A$$

$\kappa(A)$ – condition number of A

$$\kappa(A) = \|A\| \|A^{-1}\|$$

$$\kappa_2(A) = \frac{\sigma_1}{\sigma_n} = \left(\|A\|_2 \|A^{-1}\|_2\right)$$

Eigenvalues / Eigenvectors

$$Ax = \lambda x, \quad x \neq 0, \quad A \in \mathbb{R}^{n \times n}$$

λ – eigenvalue

x – eigenvector $(\|x\|_2 = 1)$

$$Ax - \lambda x = 0$$

$$(A - \lambda I)x = 0 \Rightarrow A - \lambda I \text{ is singular}$$

$\Rightarrow \det(A - \lambda I) = 0$

$$\det\left(\begin{bmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & \cdots & & a_{nn}-\lambda \end{bmatrix}\right)$$ is polynomial of degree n in λ

If A is $n \times n$, it has n eigenvalues

If $A = A^T$ (symmetric) $\Rightarrow$ all eigenvalues are real and it has a full set (n) of mutually orthogonal eigenvectors

$A = A^T$, $Av_1 = v_1\lambda_1$ $\quad \lambda_i \in \mathbb{R}$

$Av_2 = v_2\lambda_2$

$\vdots$

$Av_n = v_n\lambda_n$

$v_i^T v_j = 0, \; i \neq j$

$v_i^T v_i = 1$

$\Downarrow$ $V = [v_1 v_2 \cdots v_n]$

$V^T V = I = V V^T$

$$A[v_1 \cdots v_n] = [v_1 \cdots v_n] \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

$AV = V\Lambda$, $\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$

$\boxed{A = V\Lambda V^T}$ – Eigenvalue Decomposition

A is positive semi-definite if $x^T A x \geq 0 \; \forall x$ $(A \succeq 0)$

A is positive definite if $x^T A x > 0 \; \forall x \neq 0$ $(A \succ 0)$

$A = A^T$, $x^T A x = (x^T V) \Lambda (V^T x)$

Let $z = V^T x$, then $x^T A x = z^T \Lambda z = \sum_{i=1}^{n} \lambda_i z_i^2$

$x^T A x \geq 0 \Leftrightarrow \lambda_i \geq 0$

$A = GG^T$, $x^T A x = x^T G G^T x = z^T z \geq 0 \quad (z = G^T x)$