Lecture 1: Linear Regression

Example 1: Predict levels of PSA from various measurements on the prostate

Training Data: $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4) \ldots, (x_N, y_N)$

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

$x_1, x_2, \ldots x_N$ are measurements on prostate, $x_i \in \mathbb{R}^d$

Example 2: Netflix - (user, movie, rating)

$x \leftarrow$ (Inderjit, Breaking Bad Ep 1, Season 1, 4 ★★★★) — y_i = 4 ★★★★

\$1M Netflix Prize

Goal: Predict y for a new x, $y_i \in \mathbb{R}$ (regression)

Example 3: Predict whether an email is spam or not

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

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

When Y is categorical $\rightarrow$ Classification

When Y is real-valued $\rightarrow$ Regression

Regression Problem

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

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

Prediction (linear):

$$y(x) = w_0 + w_1 x(1) + w_2 x(2) + \ldots + w_d x(d)$$
$$= w_0 + \tilde{w}^T x, \quad \tilde{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$
$$= w_0 \cdot 1 + \tilde{w}^T x$$

$$= w^T \bar{x} \quad , \quad x = \begin{bmatrix} 1 \\ x(1) \\ x(2) \\ \vdots \\ x(d) \end{bmatrix}, \quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$w, \bar{x} \in \mathbb{R}^{d+1}$$

$$y(x_1) = w^T \bar{x}_1 \simeq y_1 \longrightarrow (w^T \bar{x}_1 - y_1)^2 \text{ is small}$$

$$y(x_2) = w^T \bar{x}_2 \simeq y_2 \longrightarrow (w^T \bar{x}_2 - y_2)^2 \text{ is small}$$

$$y(x_3) = w^T \bar{x}_3 \simeq y_3 \longrightarrow (w^T \bar{x}_3 - y_3)^2 \text{ is small}$$

$$\vdots$$

$$y(x_N) = w^T \bar{x}_N \simeq y_N \longrightarrow (w^T \bar{x}_N - y_N)^2 \text{ is small}$$

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

Goal: $\underset{w}{\text{Minimize}} \; F(w)$ — Least Squares Regression

$$X = \begin{array}{c} \\ \uparrow \\ d+1 \\ \downarrow \end{array} \overbrace{\begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_N \end{bmatrix}}^{N} \quad , \quad X \in \mathbb{R}^{(d+1) \times N}$$

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

$$X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1(1) & x_2(1) & \cdots & x_N(1) \\ x_1(2) & x_2(2) & \cdots & x_N(2) \\ \vdots & \vdots & & \vdots \\ x_1(d) & x_2(d) & \cdots & x_N(d) \end{bmatrix} \quad , \quad X \in \mathbb{R}^{(d+1) \times N}$$

$${}_{d+1}\overset{N}{X} \quad , \quad y \in \mathbb{R}^N \quad , \quad w \in \mathbb{R}^{d+1}$$

$$\boxed{F(w) = \frac{1}{2} \| X^T w - y \|_2^2} \qquad {}_{N}\overset{d+1}{X^T} \cdot w_{\,d+1} \simeq y(X)$$

Least Squares Objective

Goal: Find w such that $F(w)$ is minimized

$$\|z\|_2^2 = z^T z \qquad a^T b = b^T a$$

$$
\begin{aligned}
F(w) &= \tfrac{1}{2}\left(X^T w - y\right)^T \left(X^T w - y\right) \\
&= \tfrac{1}{2}\left(w^T X - y^T\right)\left(X^T w - y\right) \\
&= \tfrac{1}{2}\left(w^T X X^T w - \underline{y^T X^T w} - \underline{w^T X y} + y^T y\right) \\
&= \tfrac{1}{2}\left(w^T X X^T w - 2\, y^T X^T w + y^T y\right)
\end{aligned}
$$

$$\nabla_w F(w) = \tfrac{1}{2}\left(2 X X^T w - 2 X y\right) = X X^T w - X y$$

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

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

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

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

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

$w \in \mathbb{R}^{d+1}$

$d+1$ linear equations

$d+1$ unknowns

$$\boxed{w^* = \left(X X^T\right)^{-1} X y}$$

- ok if $X X^T$ is non-singular

Normal Equations

Geometric View

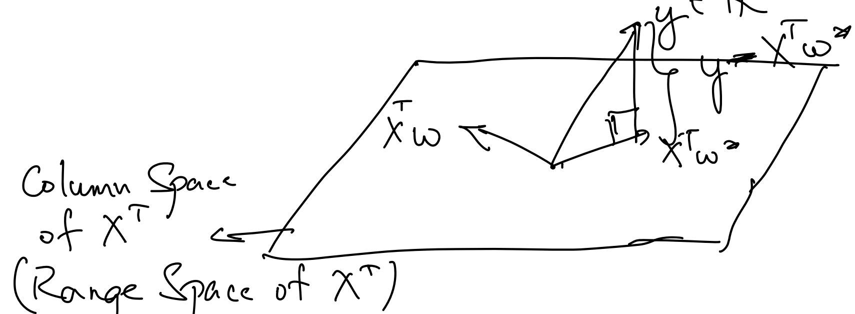
$$X^T = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_N^T \end{bmatrix} = \begin{bmatrix} \bar{x}_1^T \\ \bar{x}_2^T \\ \vdots \\ \bar{x}_N^T \end{bmatrix} \qquad X^T \in \mathbb{R}^{N \times (d+1)}$$

$$X^T w \approx y$$

$X^T w \leftarrow$ Linear Combination of columns of X^T

$= w_0 \times \text{col}1 + w_1 \times \text{col}2 + \dots\ w_d \times \text{col}(d+1)$

$\in \mathbb{R}^N$



$$X^T w \perp y - X^T w^* \quad \forall w \quad \left(\text{to minimize } F(w) \quad w^* = \arg\min F(w)\right)$$

$a \perp b \quad a^T b = 0$

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

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

$$w^T \underbrace{(Xy - XX^T w^*)}_{0} = 0 \quad \forall w$$

$$\Rightarrow \boxed{XX^T w^* = Xy}$$

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

$d+1$ equations in $d+1$ unknowns $(w_0, w_1, \ldots w_d)$

Solve $XX^T w^* = Xy$ - linear system of equations

How?

(1) Do Gaussian Elimination + solve $\underline{XX^T w^*} = \underline{Xy}$

- Form the matrix $\underline{XX^T = A}$ - $O(d^2 N)$ operations
- Form the right hand side $Xy = b$ - $O(dN)$

$$A w^* = b$$

$A = LU$ decomposition (Gaussian Elimination)

$A = XX^T$ - positive semi-definite matrix

$A = LL^T$ (Cholesky Decomposition) - L is lower triangular, L^T is upper triangular

$O(d^3)$ operations

$$LL^T w^* = b$$

Solve $Lz = b$ $\quad \triangle \, \cdot = \cdot$ $\quad$ Forward Substitution $O(d^2)$

$L^T w^* = z$ $\quad \nabla \, \cdot = \cdot$ $\quad$ Backward Substitution $O(d^2)$

Solve normal equations by Gaussian Elimination in $O(d^2 N + d^3)$ operations

Inverse of A (XX^T) exists only if A is non-singular

Condition Number of A is large $\equiv$ A is close to singularity

Solving normal equations can yield large error when A is poorly conditioned

Other methods which are better when A is close to singular:

(1) Use QR decomposition of X

☆☆☆☆☆ (2) Use SVD of X.

Singular value decomposition

Most accurate but more expensive to compute.