2/18/15

REGRESSION

① Recall: Linear Regression

Model $\hat{y} = w_0 + w_1 x_1 + \cdots + w_d x_d$

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

Least Squares Fitting / Ordinary Least Squares

dataset: $D = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)\}$

$$\hat{w}_{OLS} = \underset{\vec{w}}{\operatorname{argmin}} \; L(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \frac{1}{n} \| Y - Xw \|_2^2$$

$$= \frac{1}{n}\left[Y^T Y - 2 Y^T X w + w^T X^T X w \right]$$

Detour: Matrix / Vector differentiation

Scalars: x, y

Vectors: $\vec{x}, \vec{y}$

Matrices: X, Y

	S	V	M
S	$\frac{\partial y}{\partial x}$	$\frac{\partial \vec{y}}{\partial x}$	$\frac{\partial Y}{\partial x}$
V	$\frac{\partial y}{\partial \vec{x}}$	$\frac{\partial \vec{y}}{\partial \vec{x}}$	
M	$\frac{\partial y}{\partial X}$		

Convention: $\frac{\partial \vec{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$ ↓ numerator = dim 1 / col-vector

→ denominator

$\frac{\partial y}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$ = dim 2 / row-vector

$\frac{\partial \vec{y}}{\partial \vec{x}} = \begin{bmatrix} & \vdots & \\ \cdots & \frac{\partial y_i}{\partial x_j} & \cdots \end{bmatrix}_{m \times n}$ (i ↓, j →)

Now,

$$\frac{\partial (\vec{w}^T \vec{x})}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial \vec{w}^T \vec{x}}{\partial w_0} & \cdots & \frac{\partial \vec{w}^T \vec{x}}{\partial w_d} \end{bmatrix}$$

$$\frac{\partial}{\partial w_0} \left[\sum_{i=0}^{d} w_i x_i \right] = x_0 \quad (=1 \text{ in this case})$$

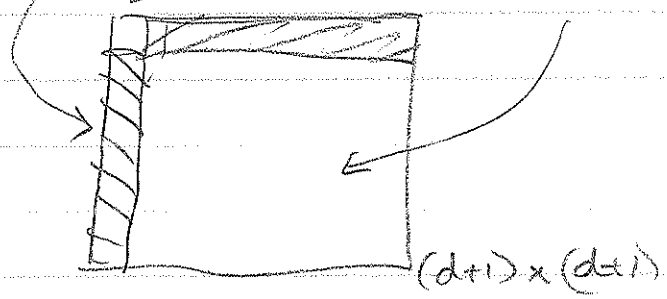
Thus $\frac{\partial (\vec{w}^T \vec{x})}{\partial \vec{w}} = \begin{bmatrix} x_0 & x_1 & \cdots & x_d \end{bmatrix} = \vec{x}^T$

(very intuitive)

Also, $\frac{\partial(\vec{w}^T A \vec{w})}{\partial \vec{w}} = \left[\frac{\partial(w^T A w)}{\partial w_0} \; - - - \; \frac{\partial(w^T A w)}{\partial w_d} \right]$

$$\frac{\partial}{\partial w_0}\left[w^T A w \right] = \frac{\partial}{\partial w_0}\left[\sum_{i=0}^{d} \sum_{j=0}^{d} w_i a_{ij} w_j \right]$$

$$= \frac{\partial}{\partial w_0}\left[a_{00} \cdot w_0^2 + \sum_{i \neq 0} a_{ij} \cdot w_j \cdot w_0 + \sum_{j \neq 0} a_{0j} w_0 w_j + \sum_{i \neq 0} \sum_{j \neq 0} w_i a_{ij} w_j \right]$$



$$= 2 a_{00} w_0 + \sum_{i \neq 0} a_{ij} w_i + \sum_{j \neq 0} a_{0j} w_j$$

$$= 2 a_{00} w_0 + 2 \sum_{i \neq 0} a_{ij} w_i$$

$$= 2 w^T A_0 \leftarrow \text{first column of } A$$

$$\Rightarrow \frac{\partial(w^T A w)}{\partial \vec{w}} = \left[2 w^T A_0 \;\; 2 w^T A_1 \; - - - \; 2 w^T A_d \right]$$

$$= 2 w^T A \quad \text{(also intuitive! Just drop } \vec{w}\text{)}$$

END DETOUR

Putting it all together

$$\frac{\partial \mathcal{L}(\vec{w})}{\partial \vec{w}} = \frac{1}{n}\left[-2 Y^T X + 2 w^T X^T X \right] = 0$$

$$\Rightarrow \quad w^T X^T X = Y^T X$$

$$\Rightarrow \quad (X^T X) w = X^T Y$$

$$\Rightarrow \quad \boxed{\hat{w}_{OLS} = (X^T X)^{-1} X^T Y}$$

Single clean equation

Note 1: $(X^T X)^{-1} X^T$ is called the Moore-Penrose Pseudo Inverse of X

notation $X^{\dagger} = (X^T X)^{-1} X^T$

Why? behaves like an inverse

$$X^{\dagger} X = (X^T X)^{-1} X^T X = I$$

Recall $Y = Xw \Rightarrow w = X^{-1} Y$

$Y \approx \hat{Y} = Xw \Rightarrow w = X^{\dagger} w$ ← See the resemblance?

Squared loss

Note 2: (X^TX) may not be invertible

Specifically,

$$(X^TX)_{(d+1)\times(d+1)} = \begin{bmatrix} \uparrow \\ \vec{x}_i \\ \downarrow \end{bmatrix}_{(d+1)\times n} \begin{bmatrix} \longleftarrow & \vec{x}_i & \longrightarrow \end{bmatrix}_{n\times(d+1)}$$

$$= \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T$$

$\vec{x}_i \vec{x}_i^T$ ⇝ rank 1 matrix

n-summations

⇒ if $n < (d+1)$

$\sum \vec{x}_i \vec{x}_i^T$ has rank $< (d+1)$

⇒ X^TX not invertible.

② Least Squares as MLE

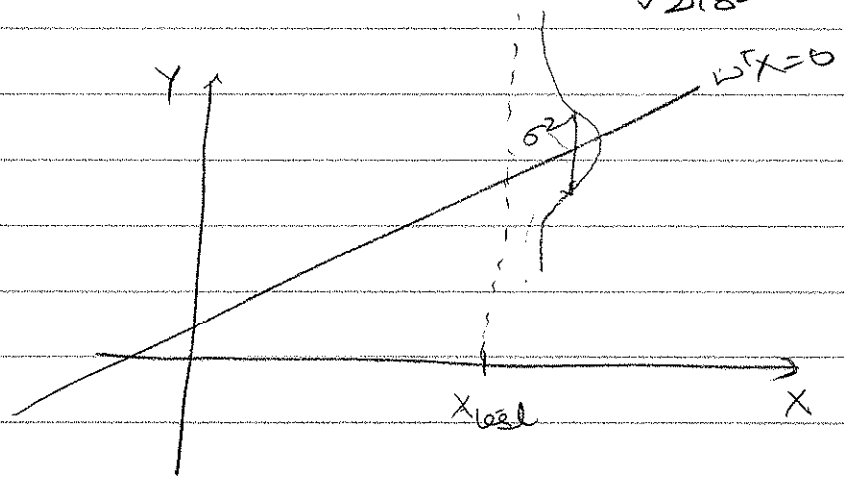
Consider a probabilistic model

$$X \sim P(X) \quad [\text{Don't Care}]$$

$$Y|X=x \sim N(w^T x, \sigma^2)$$

(σ^2: fixed / don't care)

$$\Rightarrow p(y|X=x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-w^T x)^2}{2\sigma^2}}$$



$$e = y - w^T x \sim N(0, \sigma^2) \Leftrightarrow p(e) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-w^T x)^2}{2\sigma^2}}$$

Now, given dataset $D = \{(\vec{x}_1, y_1) \cdots (\vec{x}_n, y_n)\}$

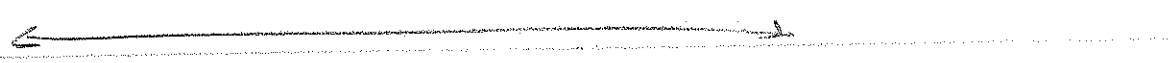
$$\hat{w}_{MLE} = \underset{\vec{w}}{\operatorname{argmax}} \log P(D|\vec{w})$$

$$= \underset{\vec{w}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(e_i|w)$$

$$= \underset{\vec{w}}{\operatorname{argmax}} \sum_{i=1}^{n} \left[-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right]$$

$$\Rightarrow \hat{w}_{MLE} = \underset{w}{\operatorname{argmax}} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - w^T x_i)^2 = \hat{w}_{OLS}$$

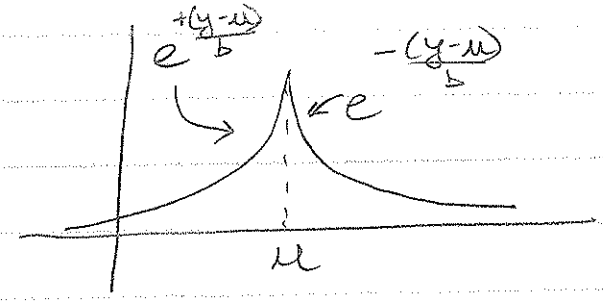


③ Robust Least Squares

$$Y \mid X = x \sim \text{Laplacian}(w^T x, b)$$ ← again fixed

Recall,

$$Y \sim \text{Lap}(\mu, b) \Leftrightarrow p(y \mid \mu, b) = \frac{1}{2b} e^{-\frac{|y-\mu|}{b}}$$



$$\hat{w}_{MLE} = \underset{\vec{w}}{\operatorname{argmax}} \log P(D \mid w)$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} \left[-\frac{|y_i - w^T x_i|}{b} - \log 2b \right]$$

$$= \underset{\vec{w}}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - w^T x_i|$$

So Laplace likelihood $\Leftrightarrow$ Least Absolute Error!

How do we solve this? HW2!