Linear regression

Note Title 9/28/2017

$$y = w^T x + t \qquad x \in \mathbb{R}^D$$

$$\Rightarrow \epsilon = y - w^T x \quad \text{(residual)} \qquad w \in \mathbb{R}^D$$

$$= \sum_{i=1}^{D} w_i x_i + \epsilon$$

observing

$$\epsilon \sim N(0, \sigma^2)$$

$$P(y \mid X, \theta) = N(y \mid w^T x, \sigma^2)$$

A system L is linear if

$$L(ax + y) = aL(x) + L(y)$$

example

$$L = A_{n \times n}$$

$$A_{n \times n}[aX + Y] = aAX + AY$$

Note: linear system can be used to model non-linear system.

right now x_i are feature

$$[x_1, x_2 \quad \cdots x_D, \; x_1^2, \; x_D^2, \quad \cdots \quad x_{1D}^2] = Y$$

$$w_1 \; w_2 \quad w_D \qquad\qquad w_{2D} = W$$

In general

$$Y = W^T Y + \epsilon$$

$\wedge$ We can modify features $x \xrightarrow{\phi} \phi(x)$

$\phi \rightarrow$ polynomial

$\rightarrow$ kernel methods

$\rightarrow$ deep learning does feature engineering

$D \rightarrow 2D$

In general we can model non-linear system as

$$\boxed{N(y \mid \phi(x), \sigma^2)}$$

$\rightarrow x^T x_2 \in \mathbb{R}$

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

later we will see that we only need to know about interactions between feature x_i

MLE estimation (Least square estimation)

let say we observed IID sample $D = \{(x_i, y_i)\}_{i=1}^{N}$

$P(D|\theta)$

$$\hat{\theta} = (\hat{w}) = \arg\max_{\theta} \log P(D|\theta)$$

$P(\theta|D)$

$$= \arg\max_{\theta} \log \prod_{i=1}^{N} \mathcal{N}(y_i | w^T x_i, \sigma^2)$$

$$= \arg\min_{\theta} -\log \underbrace{\prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left(-\frac{1}{2}\frac{(y_i - w^T x_i)^2}{\sigma^2}\right)}_{LLL(\theta)}$$

(because lot of machine learning packages has minimization abstraction (API) built-in

$$= \arg\min_{\theta} \left[-\sum_{i=1}^{N} \log \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} + \sum_{i=1}^{N} \frac{1}{2\sigma^2} (y_i - w^T x_i)^2 \right]$$

$$= \arg\min_{\theta} \sum_{i=1}^{N} \frac{1}{2\sigma^2} (y_i - w^T x_i)^2$$

$$= \arg\min_{w=\theta} \sum_{i=1}^{N} (y_i - w^T x_i)^2 \qquad w \in \mathbb{R}^D$$

$$= \arg\min_{w=\theta} \sum_{i=1}^{N} \|\epsilon_i\|^2$$

RSS – Residual sum of Square

MSE (mean square error) $= \frac{RSS}{N}$

$$= \arg\min_{w=\theta} \|\epsilon\|^2 \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

$$= (y - Xw)^T (y - Xw)$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$$

$$X = \begin{bmatrix} - x_1^T - \\ - x_2^T - \\ \vdots \\ - x_i^T - \\ \vdots \end{bmatrix} \qquad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix} \in \mathbb{R}^D$$

$$y - Xw$$

$$= \begin{bmatrix} y_1 \\ y_2 \\ \\ \\ y_N \end{bmatrix} - \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \\ \vdots \\ x_N^T \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$$

$$= \begin{bmatrix} y_1 \\ y_2 \\ \\ y_N \end{bmatrix} - \begin{bmatrix} x_1^T w_1 = w_1^T x_1 \\ x_2^T w_2 \\ \vdots \\ x_N^T w_N \end{bmatrix}$$

$$= \begin{bmatrix} y_1 - x_1^T w_1 \\ y_2 - x_2^T w_2 \\ \vdots \\ y_N - x_N^T w_N \end{bmatrix}$$

Side

$$\begin{bmatrix} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{bmatrix} \quad \begin{bmatrix} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ \hline C_{21} & C_{22} \end{bmatrix}$$

$$C_{11} = A_{11}B_{11} + A_{12}B_{21}$$

For matrices block wise multiplication is also true.

$$\left(\underset{1 \times D}{(x_1^T)} \; \underset{D \times 1}{w} \right)_{1 \times 1}$$

$$x^T x = \|x\|_2^2$$

$$= x_1^2 + x_2^2 + x_3^2$$

$$(y - Xw)^T (y - Xw) = \ell(w)$$

$$= \begin{bmatrix} y_1 - x_1^T w_1 \\ y_2 - x_2^T w_2 \\ \vdots \\ y_N - x_N^T w_N \end{bmatrix}^T \begin{bmatrix} y_1 - x_1^T w_1 \\ y_2 - x_2^T w_2 \\ \vdots \\ y_N - x_N^T w_N \end{bmatrix}$$

$$\underset{w}{\operatorname{argmin}} = \sum_{i=1}^{N} (y_i - x_i^T w_i)^2 = \ell(w)$$

$$\ell(w) = (y - Xw)^T (y - Xw)$$

$$= (y^T - (Xw)^T)(y - Xw)$$

$$= (y^T - w^T X^T)(y - Xw)$$

$$= y^T y - y^T Xw - w^T X^T y + w^T X^T Xw$$

$$= -2 y^T Xw$$

$$\frac{\partial L(w)}{\partial w} = 0 - 2y^T X + w^T\left(x^T X + (X^T X)^T\right)$$

$$= -2y^T X + w^T\left(x^T X + x^T (x^T)^T\right)$$

$$= -2y^T X + 2w^T x^T X$$

Let put it equal to 0

$$2w^T x^T X = +2y^T X$$

$$x^T X w = x^T y$$

$$w = \underline{\underline{(x^T X)^{-1} x^T y}}$$

if $\leftarrow$ =

$X_{N \times D}$ as far as column rank of X is D

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

$$\frac{dl(w)}{dw^2} = +2X^T X_{D \times D}$$

So w is minimizer if

$X^T X$ is positive

definite matrix

=

positive definite matrix $A_{m \times m}$

= A is a positive definite matrix

if for any non-zero vector

$$\min_{x \in \mathbb{R}^n} \frac{1}{(Ax)_{mx1}} \quad A_{mxn} \; x_{nx1}$$

s.t.

$$x \geq 0$$

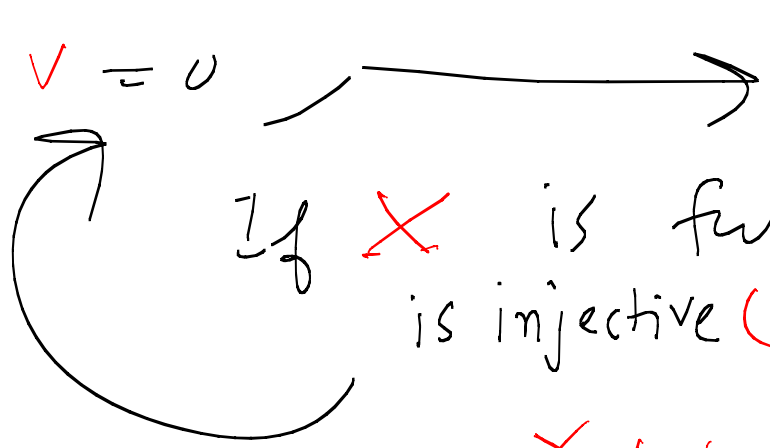
$$\|x\|_2 = (x^T x)^{1/2} = 1$$

$$v^T x^T x v = 1$$

$$x^T x \quad \text{Positive definite.}$$

$$v^T \underbrace{A^T A}_{nxn} v \geq 0$$

$$1x1$$



If X is full rank, then it is injective (1-1), Hence

$XV \neq 0$ for $V \neq 0$

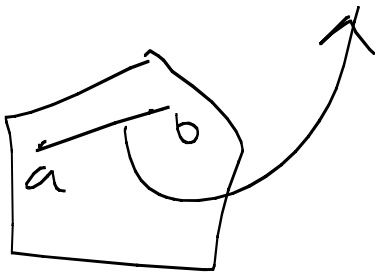
$\Rightarrow \|XV\|_2 \neq 0$

Convex Function

a function $f(\theta)$ is called Convex (*)

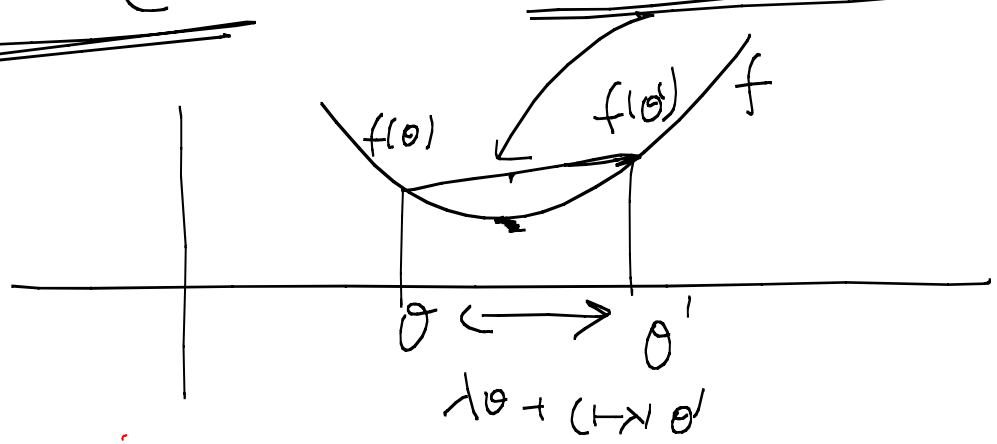
Convex Set A = A set is convex if for point $a \in A$, $b \in A$

$\lambda a + (1-\lambda) b \in A \quad \lambda \in [0\ 1]$



(*) on a convex set S if for any $\theta, \theta' \in S$

$$f(\lambda\theta + (1-\lambda)\theta') \leq \lambda f(\theta) + (1-\lambda) f(\theta')$$



i.e line joining $f(\theta)$ and $f(\theta')$ is always above function between θ and θ'.