1.1

$$f(\beta) = \|y - X\beta\|^2$$

$$= (y - X\beta)^T (y - X\beta)$$

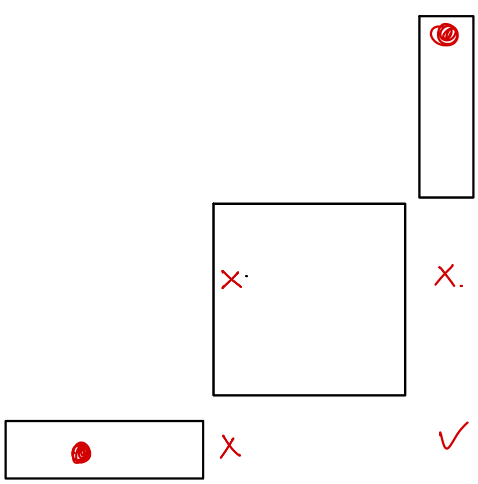
$$\nabla_\beta f(\beta) = \nabla_\beta (y^T - \beta^T X^T)(y - X\beta)$$

$$= \nabla_\beta \cancel{y^T y} - \underbrace{y^T X\beta}_{X^T y} - \underbrace{\beta^T X^T y}_{X^T y} + \underbrace{\beta^T X^T X\beta}_{2X^T X\beta}$$

$$= 2X^T X\beta - 2X^T y = 0 \quad (\text{극점})$$

then $X^T y = X^T X\beta$

직관.



$$\nabla_\beta \beta^T A\beta = \nabla_\beta \sum_{i,j} \beta_i A_{ij} \beta_j$$

$$= (A + A^T)\beta$$

1.2

- likelyhood $L(x_1, x_2, \dots, x_n | \theta)$

$$= p(x_1|\theta)\, p(x_2|\theta) \cdots p(x_n|\theta)$$

- MLE

Max likelyhood estimate

$p(X|\theta)$를 높이는 θ를 찾자!

Bernoulli

random variable

$$P\{x_1 = 1 | \theta\} = \theta$$

$$P\{x_1 = 0 | \theta\} = 1 - \theta.$$

$$p(x_i|\theta) = \theta^{x_i} \cdot (1-\theta)^{1-x_i} \xrightarrow{\log} x_i \log\theta + (1-x_i)\log(1-\theta)$$

$$p(x|\theta) = \theta^{\Sigma x_i} \cdot (1-\theta)^{n - \Sigma x_i} \longrightarrow \Sigma x_i \log\theta + (n - \Sigma x_i)\log(1-\theta)$$

θ에 대한 미분!

$$\Sigma x_i \cdot \frac{1}{\theta} + (n - \Sigma x_i)\frac{-1}{1-\theta} = 0$$

$$(1-\theta)\Sigma x_i = (n - \Sigma x_i)\theta$$

$$\theta = \frac{\Sigma x_i}{n}$$

$p(\theta)$ prior

$p(data|\theta)$ likelyhood

$p(data)$ evidence

$p(\theta|data)$ posterior

예시 θ : 감기

data : 기침.

$$p(\theta|data) = \frac{p(data|\theta)\, p(\theta)}{p(data)}$$

$$\propto p(data|\theta) \cdot p(\theta)$$

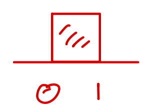
MAP : $p(\theta|data)$ 를

Max A Posterior Estimate

높이는 θ를 찾기.

Bernoulli

let $p(\theta) =$



$P\{x_1=1|\theta\} = \theta$

$p\{x_1=0|\theta\} = 1-\theta$.

$p(x_i|\theta) = \theta^{x_i} \cdot (1-\theta)^{1-x_i}$

$p(x|\theta) = \theta^{\sum x_i} \cdot (1-\theta)^{n-\sum x_i}$

$p(\theta|data) \propto p(\theta) \cdot p(data|\theta)$

$p(\theta|data) \propto \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \quad (0<\theta<1)$

$$p(\theta|data) = \left(\frac{\Gamma(n+2)}{\Gamma(\sum x_i+1)\,\Gamma(n-\sum x_i+1)} \cdot \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \right) \quad \text{in} \quad 0<\theta<1$$

$\Gamma(n) = (n-1)!$

$\Rightarrow$ $p(\theta)=1$ 라는 건 $(0<\theta<1)$ θ에 대한 정보가 없음을 의미한다.

$\Rightarrow$ MLE = MAP

$X \sim Beta(\alpha, \beta)$

pdf $\quad f(x) = (\text{constant}) \cdot x^{\alpha-1} \cdot (1-x)^{\beta-1} \qquad \boxed{0 \le x \le 1}$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} x^{\alpha-1}\cdot(1-x)^{\beta-1} \qquad \boxed{\Gamma(n) = (n-1)!}$$

(증명은 어려움.)

Bernoulli Distribution.

H	T
7	3

MLE $\quad \cdot \; p(data|\theta) = \theta^7(1-\theta)^3 \longrightarrow \theta_{MLE} = 0.7$

MAP $\quad \cdot$ let $\theta \sim Beta(2,2)$

$$p(\theta) = \frac{\Gamma(4)}{\Gamma(2)\cdot\Gamma(2)} \theta(1-\theta)$$

$$= \frac{3!}{1!\cdot 1!} \theta(1-\theta) \qquad 0 \le \theta \le 1$$

$\cdot \; p(\theta|data) \propto p(\theta) \cdot p(data|\theta)$

$$\propto \theta^8(1-\theta)^4 \longrightarrow MAP = 0.666\ldots$$

Amazon.

MAP

	data $\oplus$	$\ominus$	
Seller ①	90	10	$\longrightarrow \theta_1$
Seller ②	2	0	$\longrightarrow \theta_2$.

assume

$p(\theta_1) = 1 \qquad 0 < \theta_1 < 1$

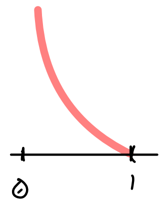
$p(\theta_2) = 1 \qquad 0 < \theta_2 < 1$

$$p(\theta_1|data) \propto \theta_1^{90}(1-\theta_1)^{10}$$

$$p(\theta_2|data) \propto \theta_2^{2}.$$

$$p(\theta_1 > \theta_2|data) = \int_0^1\int_0^1 I\{\theta_1 > \theta_2\}\, Beta(\theta_1|91,11)\, Beta(\theta_2|3,1)\, d\theta_1 d\theta_2 \approx 71\%$$

· $h(x) = -\log p(x)$



· indep x, y : $h(x, y) = h(x) + h(y)$

· entropy : $H(X) = E[-\log p(X)]$

① $H(X) \geq 0$

② $H(X) = 0$ if $\exists x \ p(x) = 1$

$$H(X, Y) = H(X) + H(Y|X)$$

$$-\log p(X, Y) = -\log p(X) - \log \frac{p(X, Y)}{p(X)}$$

- $D_{KL}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

example. Bernoulli $\begin{cases} p \Rightarrow r \\ q \Rightarrow s \end{cases}$

→ symmetric (X)

– triangular inequality (X)

$$D_{KL}(p\|q) = (1-r) \log \frac{1-r}{1-s} + (r) \log \frac{r}{s}$$

$D_{KL}(p\|q) \geq 0$ $\quad \left[D_{KL}(p\|q) = 0 \text{ iff } \forall x \; p(x) = q(x) \right]$

pf. $D_{KL}(p\|q)$

$$= E_{p(x)}\left(-\log \frac{q(x)}{p(x)}\right)$$

$$\geq -\log E_{p(x)}\left[\frac{q(x)}{p(x)}\right]$$

$$= -\log \sum_x \cancel{p(x)} \cdot \frac{q(x)}{\cancel{p(x)}} = 0$$

Jensen's Inequality : f: convex. $E[f(x)] \geq f(E(x))$

equality holds when $f(x)$ = constant.

등호 조건 $\quad \text{const} = \frac{q(x)}{p(x)} = 1$

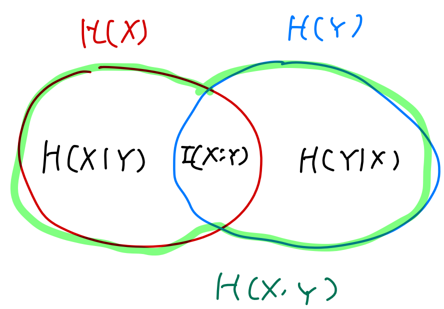
$\begin{cases} D_{KL}(p\|q) \Downarrow : \text{모든 영역} \\ D_{KL}(q\|p) \Downarrow : \text{봉우리} \end{cases}$

$$I(X;Y) = D_{KL}(p(x,y) \,\|\, p(x)p(y))$$

$$= \sum p(x,y) \cdot \log \frac{p(x,y)}{p(x)p(y)}$$

$$= H(X) + H(Y) - H(X,Y)$$

$$= H(X) - H(X|Y)$$



- Cross Entropy $H_p(g) = E_p[-\log g(X)]$

$$= -\sum_x p(x) \log g(x)$$

- $$D_{KL}(p\|g) = \sum_x p(x) \log \frac{p(x)}{g(x)}$$

$H(p)$ $H_p(g)$

$$= H_p(g) - \cancel{H(p)} \geq 0$$

- p가 실제, g가 예측이면

$D_{KL}(p\|g)$ 대신 $H_p(g)$ 사용 가능

- let $y = \beta_0 \cdot 1 + \beta_1 \cdot x + \beta_2 x + \cdots + \beta_m x_m + \varepsilon$

$\langle x_i, \beta \rangle$ $\sim N(0, \sigma^2)$

- $Y_i | X_i, \beta \sim N(\langle X_i, \beta \rangle, \sigma^2)$

- $$P(Y_i | X_i, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle x_i, \beta \rangle)^2}{2\sigma^2}\right) \xrightarrow{\log} ② - \frac{(y_i - \langle x_i, \beta \rangle)^2}{2\sigma^2}$$

- $P(Y | X, \beta)$ ← MLE

$$= P(y_1, y_2, \ldots, y_n \mid x_1, \ldots, x_n, \beta)$$

$$= P(y_1 | x_1, \beta) \cdot P(y_2 | x_2, \beta) \cdots P(y_n | x_n, \beta)$$

Max Likelyhood Estimate = Least Square Estimate.

$$\xrightarrow{\log} ② - \sum_i \frac{(y_i - \langle x_i, \beta \rangle)^2}{2\sigma^2}$$ ← Least Square Estimate