

model + cost $\rightarrow$ opt.
linear SSE normal
GD 

max. likelihood estimate (MLE) 

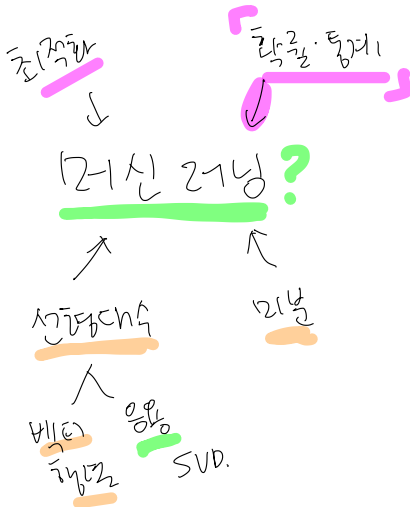
Inclass 15: MLE and MAP ; statistical inference.

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.ac.kr)

School of AI Convergence & Department of Artificial Intelligence, Dongguk University

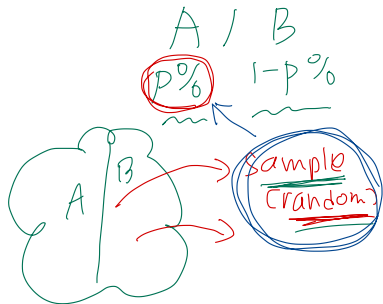


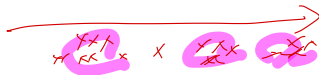
prob
 $\bar{z}_1, \bar{z}_2$
 $\bar{z}_1, \bar{z}_2$
+

$\frac{1}{n} \sum$
 $\frac{1}{n} \sum$

$\uparrow$

Stat.
 $E_1 = 1$





기초



보조



심화

linear
regression

gradient
descent

MLE

MAP

Bayesian

perceptron → NN

logistic

SVM → connex

k-nn

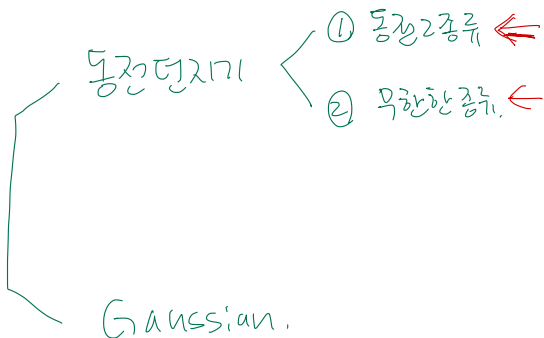
decision tree
Random forest.

NN, backpropagation

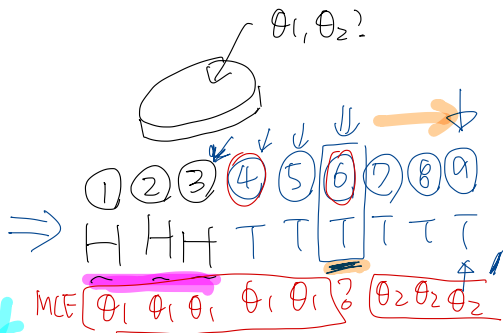
CNN : ~

① Likelihood (probability) ?

② max likelihood estimate. ?



	H	T
θ_1	0.9	0.1
θ_2	0.1	0.9



likelihood (prob)
parameter

$$L(\theta_1) = \Pr(HHH; \theta_1) = 0.9 \times 0.9 \times 0.9 = 0.729$$

$$L(\theta_2) = \Pr(HHH; \theta_2) = 0.1 \times 0.1 \times 0.1 = 0.001$$

$$\underline{\theta} \rightarrow \underline{\mathcal{L}(\theta)} = \mathcal{L}(\theta; x_1, x_2, \dots, x_N)$$

$\uparrow$
 $x_1, x_2, \dots, x_N$

① Likelihood : $\text{pr}(x; \theta)$

② θ is random variable 이든 아니든 상관없음.

Maximum likelihood estimate

Maximum likelihood estimation, or MLE, is on flavor of parameter estimation in machine learning. In order to perform parameter estimation, we need:

- some data $\mathbf{x}$
- some hypothesized generating function of the data $f(\mathbf{x}, \theta)$
- a set of parameters from that function θ
- some evaluation of the goodness of our parameters

Maximum likelihood estimate

In MLE, the objective function (evaluation) we chose is the likelihood of the data given our model. To find the best θ then, we need to find the θ which maximizes our evaluation function (the likelihood).

Therefore, in its general form the MLE is:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \mathcal{L}(\theta) \quad (1)$$

$$= \arg \max_{\theta} \log \mathcal{L}(\theta) \quad (2)$$

$$\mathcal{L}(\theta) = \prod \text{pr}(x_n; \theta)$$

$$\log \mathcal{L}(\theta) = \sum \log \text{Pr}(x_n; \theta)$$

	H	T
θ	θ	$1-\theta$

$$\begin{pmatrix} x_1 & x_2 & x_3 \\ H & H & T \\ \hline \uparrow & \uparrow & \end{pmatrix} \frac{2}{3}$$

$$\mathcal{L}(\theta) = \Pr(x_1, x_2, x_3; \theta)$$

$$= \Pr(H; \theta) \Pr(H; \theta) \Pr(T; \theta)$$

$$= \theta^2 (1-\theta)$$

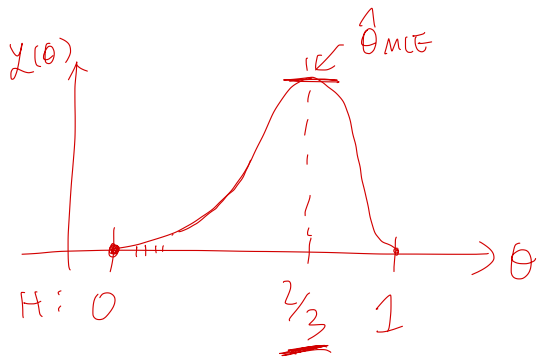
$$\log \mathcal{L}(\theta) = 2 \log \theta + \log(1-\theta)$$

$$\frac{d}{d\theta} \log \mathcal{L}(\theta) = \frac{2}{\theta} - \frac{1}{1-\theta} = 0$$

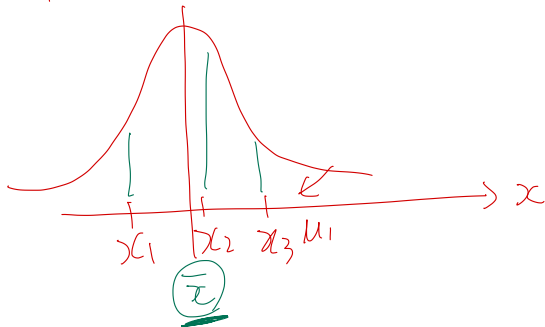
↓

$$\hat{\theta} = \frac{2}{3}$$

$$\frac{d}{d\theta} \log \mathcal{L}(\theta) = 0$$



Gaussian distribution



$$\theta = \mu$$

$$\sigma^2 = 1$$

$$x_1, x_2, x_3 \in$$

$$\mathcal{L}(\theta) = \Pr(x_1; \theta) \Pr(x_2; \theta) \Pr(x_3; \theta)$$

$$= \prod_{n=1}^3 \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x_n - \theta)^2}{2} \right\}$$

$$\log \mathcal{L}(\theta) = \sum_{n=1}^3 \cancel{\log \frac{1}{\sqrt{2\pi}}} - \frac{(x_n - \theta)^2}{2}$$

$$\frac{d}{d\theta} \log \mathcal{L}(\theta) = \sum_{n=1}^3 (x_n - \theta) = 0$$

$$\frac{1}{3} \sum_{n=1}^3 x_n = \hat{\theta}_{MLE}$$

$$\begin{aligned}
 \textcircled{\theta} &\rightarrow \frac{\mathcal{L}(\theta)}{=} \Pr(x_1 \cdots x_N; \theta) \\
 &\quad \uparrow \\
 &\quad \underline{x_1 \cdots x_N}
 \end{aligned}$$

$$\textcircled{v} \quad \underline{\underline{\Pr(x; \theta)}}$$

$\frac{x}{n}$ 이 관찰된 $\frac{\bar{x}}{n}$ 이
 θ 의 영향을 받는다.
 $\frac{\bar{x}}{n}$ 을 분포의 알아야 함.

Maximum a posteriori estimate

- Maximum likelihood estimate

$$\underline{p(x; \theta)}$$

$$\hat{\theta}_{\text{MLE}} = \arg \max \mathcal{L}(\theta) \quad (3)$$

$$= \arg \max p(x_1, x_2, \dots, x_N | \theta) \quad (4)$$

- Maximum a posteriori estimate

: MLE $p(x; \theta)$ + prior $p(\theta)$

$$\hat{\theta}_{\text{MAP}} = \arg \max p(\theta | x_1, x_2, \dots, x_N) \quad (5)$$

$$= \arg \max p(x_1, x_2, \dots, x_N | \theta) p(\theta) \quad (6)$$

prior = $\theta \sim \mathcal{Z}$ random variable.

$p(\theta), p(\theta | x)$
↖ ↗ ↖ ↗

MLE: likelihood $\leftarrow \frac{p(x; \theta)^n}{\Pr(x_1, \dots, x_n; \theta) = \mathcal{L}(\theta)}$
 $\{x_1, x_2, \dots, x_n\}$
sample

$$\hat{\theta}_{MLE} = \arg \max \mathcal{L}(\theta)$$

MAP: maximum a posteriori estimate

$$\hat{\theta}_{MAP} = \arg \max \text{posterior}$$

posterior

$$P(\theta | X) \propto$$

고관찰량이 주어졌을 때의
 θ 의 확률 분포.

likelihood prior

$$\frac{P(X|\theta)}{P(\theta)}$$

Bayes' theorem

prior: 사전 정보, 앞부분.
 이 부분도 주어지지 않았음.

	<u>H</u>	<u>T</u>	
θ_1	<u>0.9</u>	<u>0.1</u>	$\leftarrow$
θ_2	<u>0.1</u>	<u>0.9</u>	$\leftarrow$

	<u>θ_1</u>	<u>θ_2</u>	prior
<u>θ_1</u>	<u>$\frac{1}{11}$</u>	<u>$\frac{10}{11}$</u>	

$$x_1 = H$$

$$L(\theta_1) = p(H; \theta_1) = 0.9$$

$$L(\theta_2) = p(H; \theta_2) = 0.1$$

$$\hat{\theta}_{MLE} = \theta_1$$

$$\hat{\theta}_{MAP} = \theta_2$$

$$p(\theta_1 | H) \propto p(H | \theta_1) p(\theta_1)$$

$$= \frac{9}{110}$$

$$p(\theta_2 | H) \propto p(H | \theta_2) p(\theta_2)$$

$$= \frac{10}{110}$$

$$\underline{P(\theta_1 | H)} \propto \underline{P(H | \theta_1) P(\theta_1)} = \underline{\frac{9}{110}}$$

$$P(\theta_1 | H)$$

$$= \frac{P(H | \theta_1) P(\theta_1)}{P(H | \theta_1) P(\theta_1) + P(H | \theta_2) P(\theta_2)} = \frac{9}{19}$$

$$\underline{P(\theta_2 | H)} \propto \underline{P(H | \theta_2) P(\theta_2)} = \underline{\frac{10}{110}}$$

$$= \frac{P(H | \theta_2) P(\theta_2)}{P(H | \theta_1) P(\theta_1) + P(H | \theta_2) P(\theta_2)} = \frac{10}{19}$$

$$x_1 = H$$

$$\underline{x_2 = H.} \leftarrow$$

$$\rightarrow \mathcal{L}(\theta_1) = p(HH; \theta_1) = \underline{0.9} \times \underline{0.9} \quad \checkmark$$

$$\rightarrow \mathcal{L}(\theta_2) = p(HH; \theta_2) = \underline{0.1} \times \underline{0.1} \quad \checkmark \quad \text{(\underline{0.8})}$$

$$p(\underline{\theta_1 | HH}) \propto p(HH; \theta_1) \quad \text{(\underline{\theta_1})} = \frac{0.1}{100} \times \frac{1}{11} \neq \frac{0.1}{91}$$

$$p(\underline{\theta_2 | HH}) \propto p(HH; \theta_2) \quad \text{(\underline{\theta_2})} = \frac{1}{100} \times \frac{10}{11} \neq \frac{10}{91}$$

$$\hat{\theta}_{MAP} = \theta_1$$

