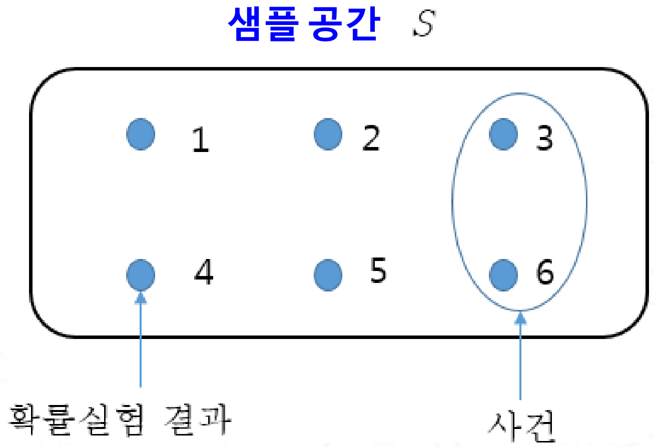
Probabilistic Model

Probability

- ❖ **Experiment**(실험)
  - **Procedure**(절차) + **Observation**(관측)
    - **Procedure** : Tossing a coin and causing it to fall to the ground
    - **Observation** : Confirmation of the coin front
  - **Outcome**(관측 결과) = **Sample**(표본)
- ❖ **Sample space**
  - A set of all possible observations
- ❖ **Event**
  - A set of any samples you may find interesting
  - A subset of the sample space



Sample space S = {1, 2, 3, 4, 5, 6}

Event E_1 = {4 이상의 결과} = {4, 5, 6}

Event E_2 = {짝수인 결과}= {2, 4,6}

❖ Definition

- The likelihood of any event
- **Relative frequency probability**(상대빈도확률)
  - Frequentist probability(빈도주의자확률)
  - Relative frequency of events of interest compared to total experiment recalls

색상	자주	초록	파랑	노랑	분홍
빈도	24	32	14	22	8

$$P(\text{자주}) = \frac{24}{24+32+14+22+8} = \frac{24}{100}$$

- **Subjective probability**(주관적 확률)
  - **확신** 또는 **믿음의 정도**(degree of belief)
  - There is a 40% chance of rain tomorrow morning

Probability

❖ Joint probability(결합 확률)

- $P(A, B)$, $P(A \cap B)$, $P(AB)$
- Probability of events A and B occurring at the same time
  - A : First dice is an even number, B : Second dice is an odd number
    - $P(A, B) = \frac{9}{36}$ = 0.25

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Probability

❖ Conditional probability(조건부확률)

- $P(A|B)$
- The probability that A will happen when B is given

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)} = \frac{P(A,B)}{P(B)} \quad \text{where } P(B) > 0.$$

A : 두 주사위의 합이 8이다
B : 첫 번째 주사위는 3이다

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

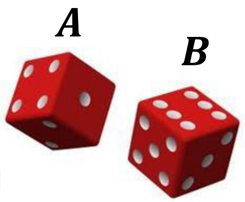
$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{1/36}{6/36} = \frac{1}{6}$$

Probability

❖ Marginal probability(주변 확률)

- When two or more events can occur at the same time, ignore the others
- The probability of noting one specific event

$$P(A) = \sum_B P(A, B)$$ **(합 규칙, sum rule)**



(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

$$\begin{aligned} P(A=3) &= \Sigma_b\, P(A=3, B=b) \\ &= P(A=3, B=1) + P(A=3, B=2) + P(A=3, B=3) \\ &\quad + P(A=3, B=4) + P(A=3, B=5) + P(A=3, B=6) \end{aligned}$$

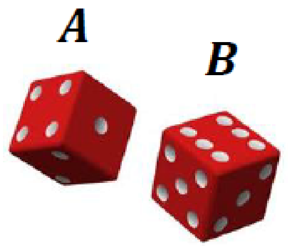
$$\frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36}$$

Probability

❖ Independent event(독립 사건)

- The probability of one event occurring does not affect the probability of the other occurring

$$P(A = 1, B = 2) = P(A = 1)P(B = 2)$$



$$\frac{1}{36} = \frac{1}{6} \times \frac{1}{6}$$

❖ Relationship between Independent event and conditional probability

- Independent event A, B

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

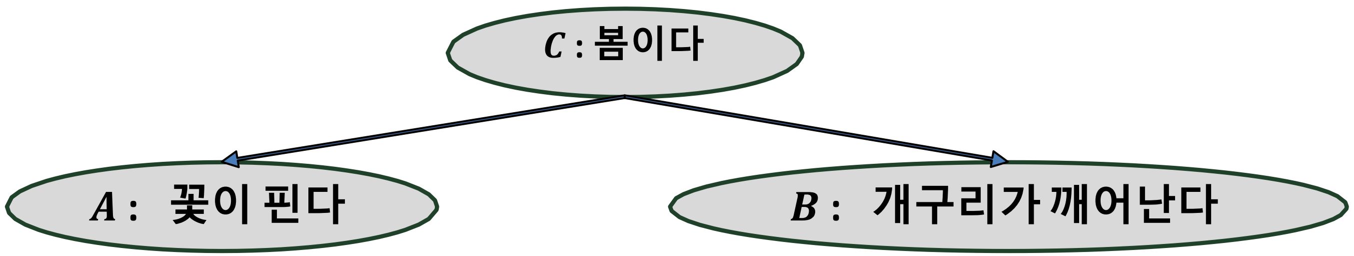
예) A : 꽃이 피었다

B : 자동차가 창밖을 지나간다

Probability

❖ Conditional independence(조건부독립)

- When $P(A|B,C) = P(A|C)$ holds, then Given C, A is conditionally independent of B



$$\mathbf{P(A,B|C) = P(A|C)P(B|C)}$$

$$P(A|B,C) = \frac{P(A,B,C)}{P(B,C)} = P(A|C)$$

$$\Rightarrow P(A,B,C) = P(B,C)P(A|C)$$

$$\Rightarrow P(A,B,C) = P(B|C)P(C)P(A|C)$$

$$\Rightarrow P(A,B|C) = \frac{P(A,B,C)}{P(C)}$$

$$\Rightarrow P(A,B|C) = P(A|C)P(B|C)$$

Probability

❖ **Mutually independent event**(상호독립사건)

$$P(A_1, A_2, \dots, A_n) = P(A_1)P(A_2)\ \dots P(A_n)$$

❖ **Conditional independence of mutually independent event**

$$P(A_1, A_2, \dots, A_n \mid B) = P(A_1 \mid B)P(A_2 \mid B)\ \dots P(A_n \mid B)$$

❖ **Product rule**

$$P(A, B) = P(B|A)P(A) = \frac{P(A, B)}{P(A)}P(A)$$

$$P(A_1, A_2, \dots, A_n) = P(A_1|A_2, \dots, A_n)P(A_2|A_3, \dots, A_n)\ \dots\ P(A_n)$$

Random Variable

❖ Random variable(확률변수)

- Numerical function of experimental results
- **Discrete random variable**(이산확률변수)
  - When you can count the value of a random variable X, you can turn that random variable X into a discrete random variable
  - e.g., coin toss, dice toss

이산확률분포

이산확률변수 X의 값 $x_1, x_2, x_3, \cdots, x_n$ 각각에 대한 확률 $p_1, p_2, p_3, \cdots, p_n$의 대응관계를 이산확률변수 X에 대한 확률분포라고 하고, 이 대응관계를

$$\mathrm{P}(X=x_i)=p_i \quad (i=1,\ 2,\ 3,\ \cdots,\ n)$$

과 같이 나타낸다. 이때, $\mathrm{P}(X=x_i)$를 이산확률변수 X의 확률질량함수라고 한다. 또한, 이 확률분포를 표로 나타낸 것을 확률분포표라고 한다.

X	x_1	x_2	$\cdots$	x_n
$\mathrm{P}(X=x_i)$	p_1	p_2	$\cdots$	p_n

Random Variable

❖ Random variable(확률변수)

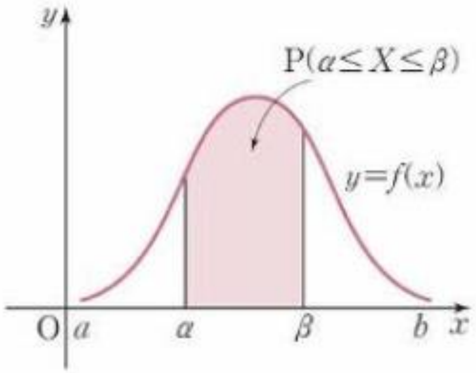
- Numerical function of experimental results
- **Continuous random variable**(연속확률변수)
  - When a random variable X has all the real values in a certain interval, such as a person's weight, commute to school, boiling point of each substance, etc., that X is a continuous random variable

연속확률변수 X의 확률밀도함수 $f(x)$ $(a \le x \le b)$에 대하여

1) $f(x) \ge 0$

2) $f(x)$의 그래프와 x축 및 두 직선 $x=a$, $x=b$로 둘러싸인 부분의 넓이는 1이다.

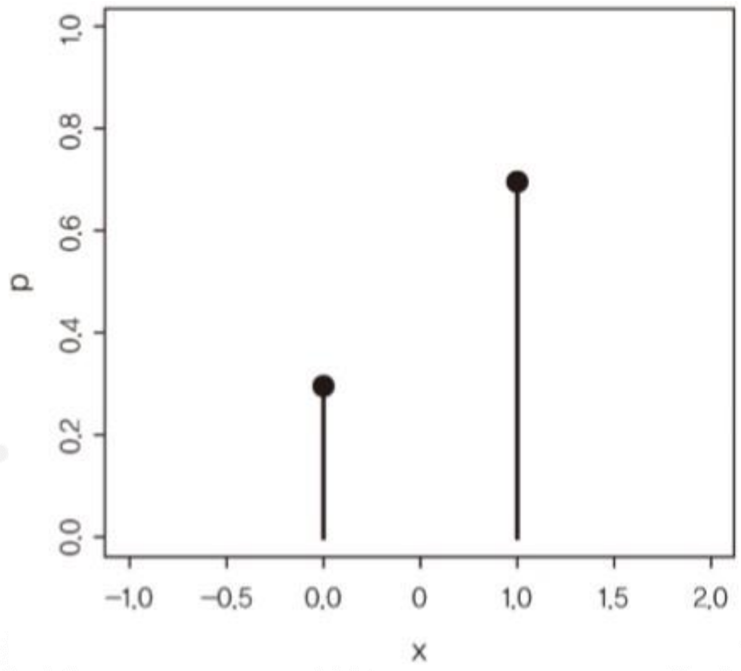
3) 두 상수 α, β $(a \le \alpha \le \beta \le b)$에 대하여 $\mathrm{P}(\alpha \le X \le \beta)$는 $f(x)$의 그래프와 x축 및 두 직선 $x=\alpha$, $x=\beta$로 둘러싸인 부분의 넓이이다.



Probability Distribution

❖ Probability mass function(확률질량함수)

- Probability distribution of discrete random variable values



$$P(x = 0) = 0.3$$

$$P(x = 1) = 0.7$$

❖ Probability mass function(확률질량함수)

주사위를 던져서 나오는 눈의 수를 4로 나눈 나머지를 확률변수 X라고 할 때, 다음 물음에 답하시오.

(1) X의 확률분포를 구하시오.

(2) $P(1 \le X \le 2)$를 구하시오.

(1) 주사위를 던져서 나오는 눈의 수는 1, 2, 3, 4, 5, 6의 6가지가 있고, 이들을 4로 나눈 나머지는 각각 1, 2, 3, 0, 1, 2이다. 따라서 확률변수가 가질 수 있는 값들은 0, 1, 2, 3이 되고, 이들에 대응되는 확률분포는 다음 표와 같다.

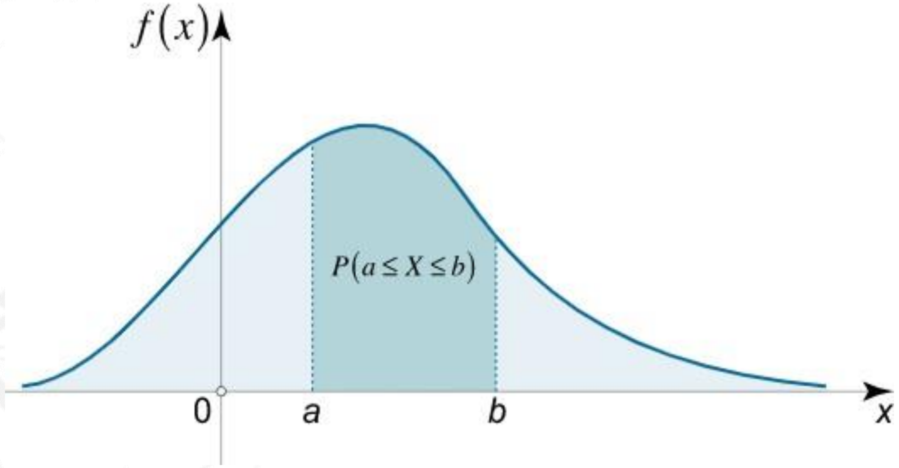
X	0	1	2	3
$P(X=x_i)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

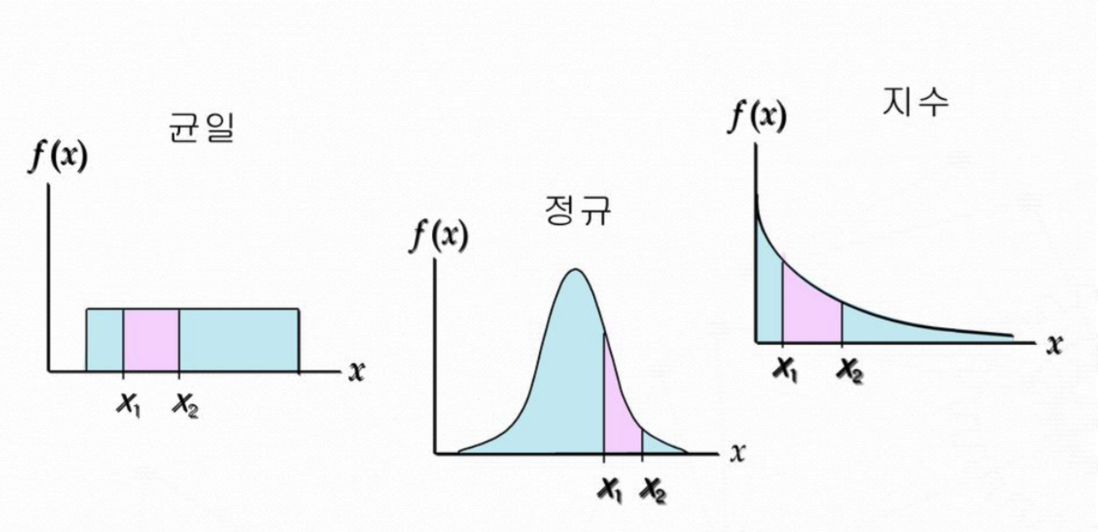
(2) $P(1 \le X \le 2) = P(X=1) + P(X=2) = \frac{2}{6} + \frac{2}{6} = \frac{2}{3}$

Probability Distribution

❖ Probability density function(확률밀도함수)

- Distribution of continuous random variables





$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \qquad P(a \le x \le b) = \int_{a}^{b} f(x)\,dx$$

❖ Probability density function(확률밀도함수)

- Distribution of continuous random variables

연속확률변수 X의 확률밀도함수가 $f(x)=kx\ (0 \le x \le 3)$일 때, 다음 물음에 답하시오. (단, k는 상수)

(1) k의 값을 구하시오.

(2) $\mathrm{P}(1 \le X \le 2)$를 구하시오.

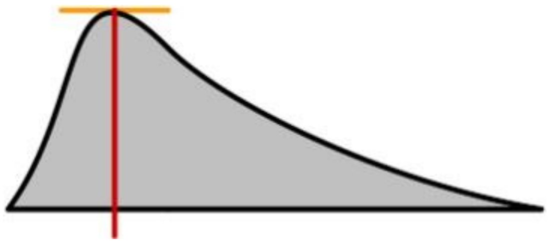
(1) $\int_0^3 kx\,dx = \left[\frac{k}{2}x^2\right]_0^3 = \frac{9}{2}k = 1$에서 $k=\frac{2}{9}$

(2) $\mathrm{P}(1 \le X \le 2)=\int_1^2 \frac{2}{9}x\,dx=\left[\frac{1}{9}x^2\right]_1^2=\frac{4-1}{9}=\frac{3}{9}=\frac{1}{3}$

Probability Distribution

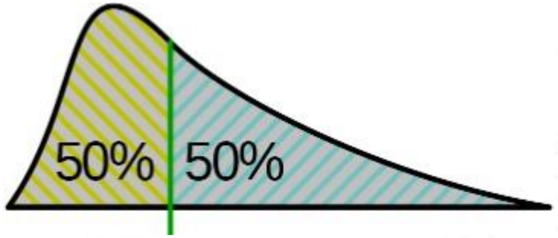
❖ **Mode**(최빈값)

- In a discrete probability distribution,
  the number with the largest probability value

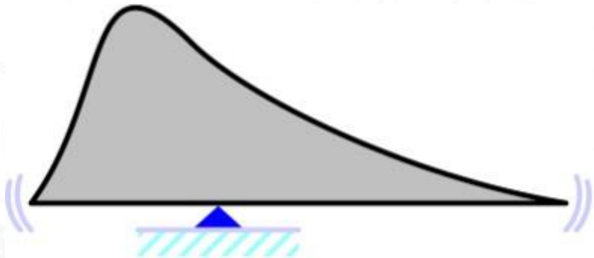


❖ **Median**(중위수)

- The probability of a greater than the median value and the probability of a smaller value are equal to 0.5



❖ **Mean**(평균)



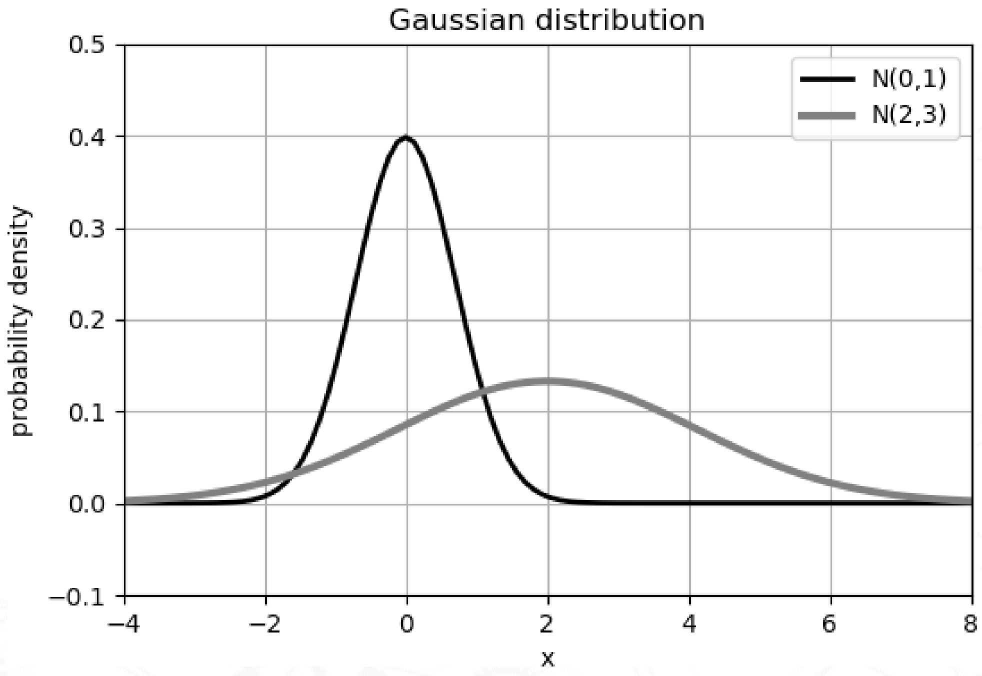
https://en.wikipedia.org/wiki/Probability_density_function

Probability Distribution

❖ Gaussian distribution

- Representative probability density function for continuous random variables in the shape of a bell
- **평균**(mean, μ)과 **표준편차**(standard deviation, σ)에 의해 형태 결정
- **Normal distribution**(정규분포)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)$$



Probability Distribution

❖ 2D Gaussian distribution

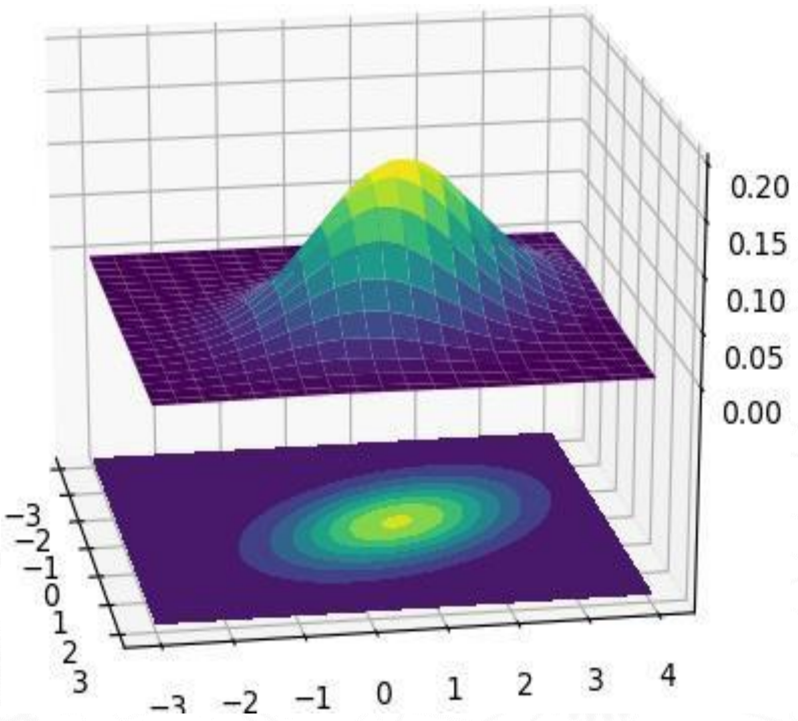
- Representative PDF for 2D bell-shaped continuous random variables
- Mean vector, μ

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

- Covariance matrix, Σ

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

$$f(x) = \frac{1}{2\pi\sqrt{|\Sigma|}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



- **Meaning of covariance matrix**
  - How similar are the spread of each (structural) feature?
  - How to make a linear transformation of (geometric) data?

Probability Distribution

❖ Bernouli Distribution(베르누이 분포)

- **Bernouli Trial**
  - The result is either Success or Fail
  - e.g., coin toss : heads (H), tails (T)
- **베르누이 분포**
  - A discrete random variable in which a random variable can have only one of the values of 0 or 1
  - Probability of 1 : θ
  - Probability of 0 : $1 - \theta$

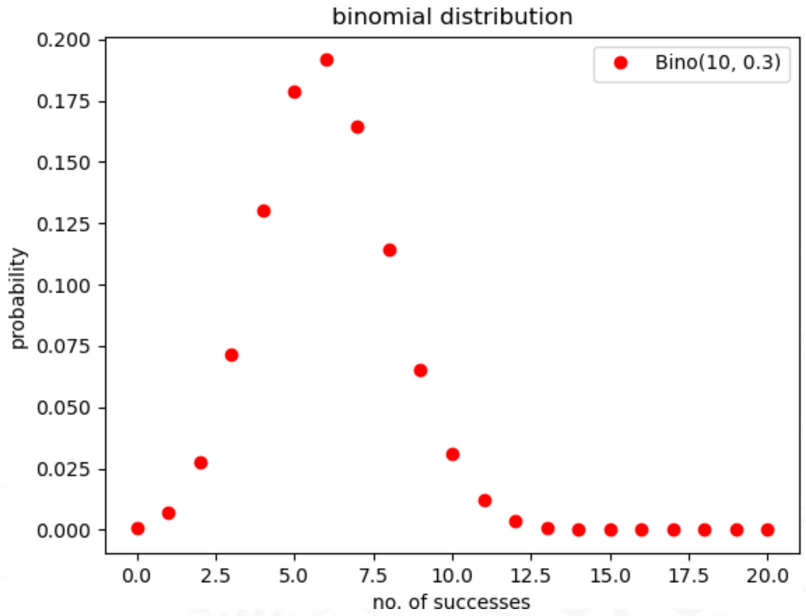
$$\mathrm{Bern}(x;\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

Probability Distribution

❖ Binomial Distribution(이항분포)

- Distribution of the random variable X, which indicates the number of successes when the Bernoulli trial is repeated n times with a probability of success p

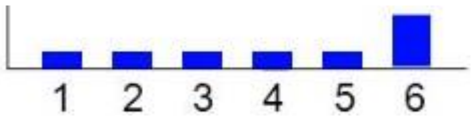
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$



Probability Distribution

❖ Multinormial Distribution(다항 분포)

- The probability distribution that arises from repeated trials of events with three or more outcomes
- 사건 $E_1, E_2, \ldots, Ek$ 중 어느 사건이 일어날 확률
  - $P_1, P_2, \ldots, P_k$
  - $P_1 + P_2 + \ldots + P_k = 1$



- n번 독립적 시행, E_1이 n_1번, E_2이 n_2번, …, E_k가 n_k번 발생할 확률

$$\frac{n!}{n_1!\, n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

$$n_1 + n_2 + \cdots + n_k = n$$

$$P_1 = 0.2, P_2 = 0.3, P_3 = 0.5$$
$$n_1 = 1, n_2 = 2, n_3 = 3$$

$$\frac{6!}{1!2!3!}(0.2^1)(0.3^2)(0.5^3) = 0.135$$

Discrete Random Variable

❖ **Expectation** (mean)

$$E(X) = \sum_{i=1}^{n} x_i p_i \; = m.$$

❖ **Variance**

$$V(X) = \sum_{i=1}^{n} (x_i - m)^2 p_i$$

❖ **Standard deviation**

$$\sigma(X) = \sqrt{V(X)}$$

X	2	3	4
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$E(X) = 2 \times \frac{1}{4} + 3 \times \frac{1}{2} + 4 \times \frac{1}{4} = 3$$

$$V(X) = 2^2 \times \frac{1}{4} + 3^2 \times \frac{1}{2} + 4^2 \times \frac{1}{4} - 3^2 = \frac{1}{2}$$

Continuous Random Variable

❖ Expectation (mean)

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

$f(x)$: 확률 밀도함수

❖ Variance

$$V(X) = \int_{-\infty}^{\infty} (x - m)^2 f(x) dx$$

❖ Standard deviation

$$\sigma(X) = \sqrt{V(X)}$$

Bayesian Theorem

❖ Bayesian theorem

사후 확률 가능도 사전 확률

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

증거

- prior probability
- likelihood
- posterior probability
- evidence

Bayes's theorem

likelihood

Prior Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Posterior Probability

Marginal Probability

Bayesian Theorem

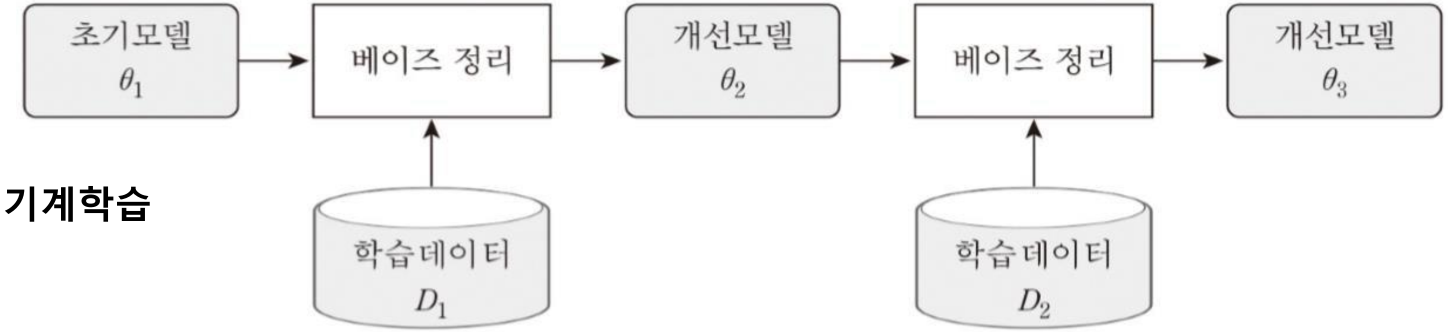
❖ Using conditional probability

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

- Accuracy of hypotheses based on evidence
  - ***P***(가설 | 증거)
- Diagnosis for symptoms and estimates the cause
  - ***P***(원인 | 증상), ***P***(부류 | 특징)
- Application of machine learning
  - ***P***(학습모델 | 학습데이터)

$$P(\theta_{i+1}|D_i) = \frac{P(D_i|\theta_i)P(\theta_i)}{P(D_i)}$$

점진적 기계학습



Bayesian Theorem

❖ Classification

- Class : C_1, C_2
- Attribute : X_1, X_2, X_3
- $\boldsymbol{C1}$ Probability: $P(C_1 \mid X_1, X_2, X_3)$
- $\boldsymbol{C2}$Probability: $P(C_2 \mid X_1, X_2, X_3)$

속성			부류
Pattern	*Outline*	*Dot*	*Shape*
수직	점선	무	삼각형
수직	점선	유	삼각형
대각선	점선	무	사각형

$P($삼각형 $\mid$ 수직, 점선, 무$)$

$P($사각형 $\mid$ 수직, 점선, 무$)$

If $P(C_1 \mid X_1, X_2, X_3) > P(C_2 \mid X_1, X_2, X_3)$, class $= C_1$
If $P(C_1 \mid X_1, X_2, X_3) < P(C_2 \mid X_1, X_2, X_3)$, class $= C_2$

❖ Multiple classification

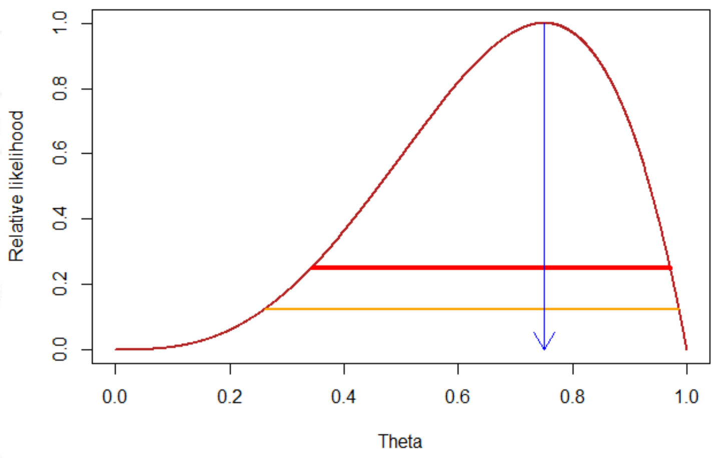
$$C^* = \arg max_{C_i} P(C_i \mid X_1, X_2, \ldots, X_n)$$

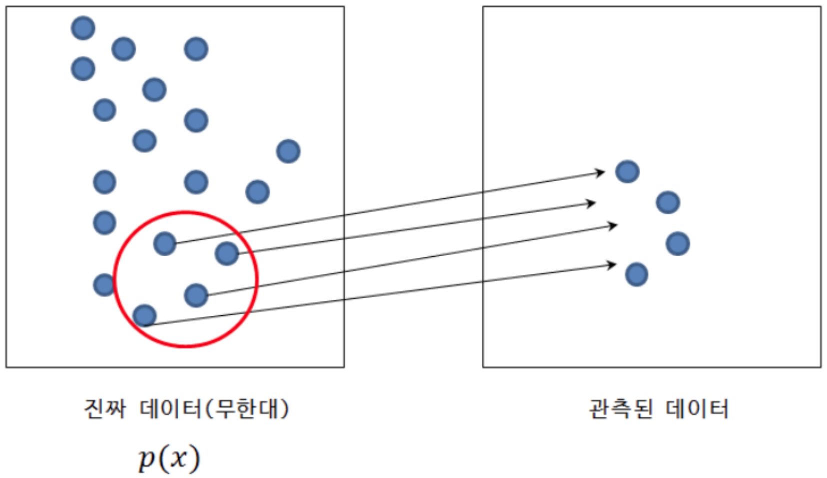
$$= \arg max_{C_i} \frac{P(X_1, X_2, \ldots, X_n \mid C) P(C_i)}{P(X_1, X_2, \ldots, X_n)}$$

MLE and MLP

❖ MLE(Maximum Likelihood Estimation, 최대 우도 추정)

- Finding the likelihood that maximizes the probability of the current dataset coming out
- Finding the parameter theta of the probabilistic model that best represents the given training data
  - If the data used for training is not representative of the actual data group, a major weakness occurs in the probability distribution

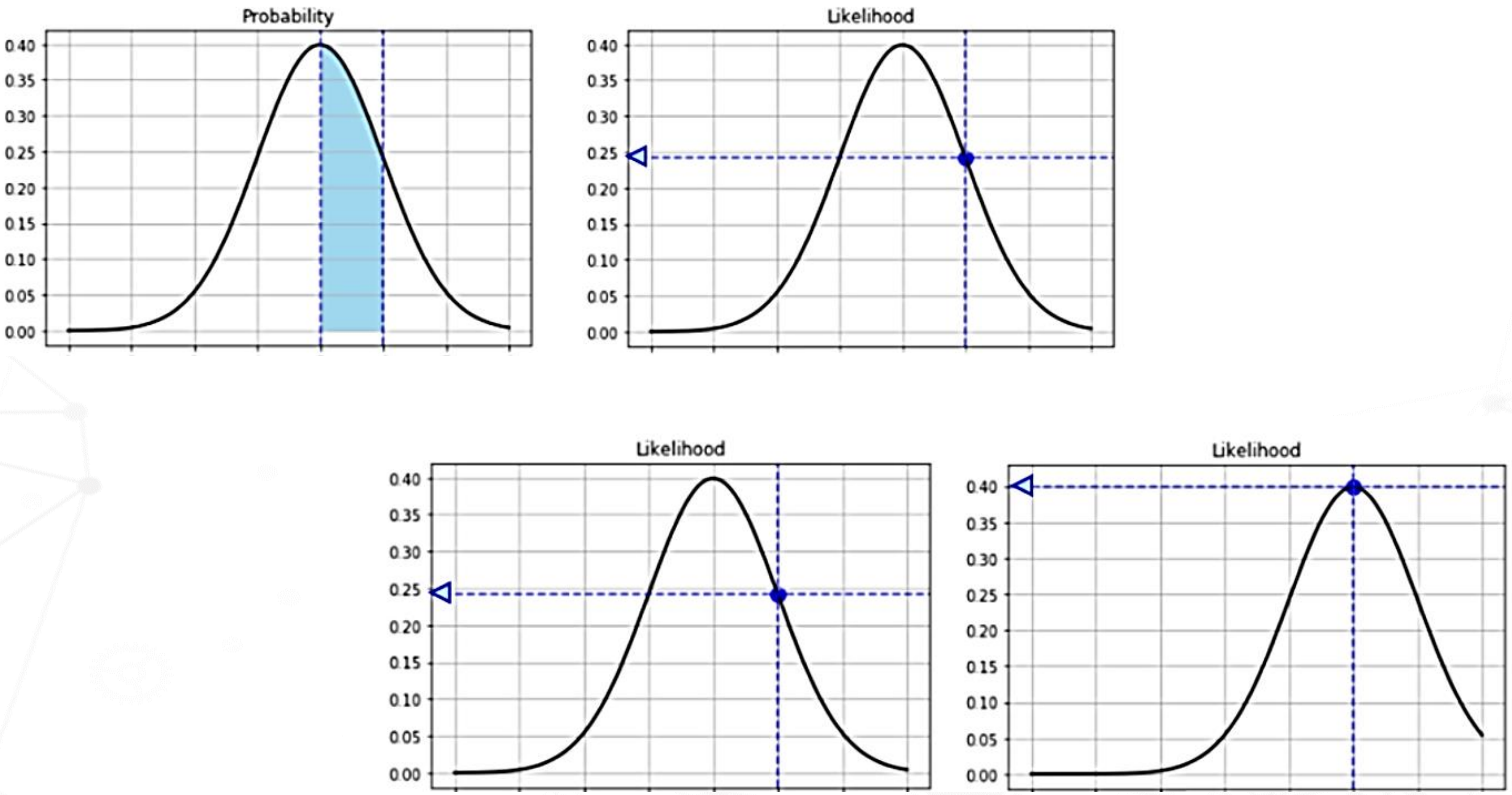




이미지 출처 : Stackoverflow.com

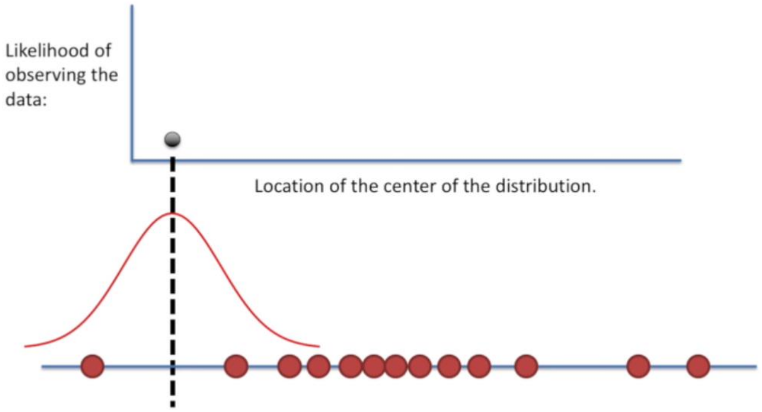
MLE and MLP

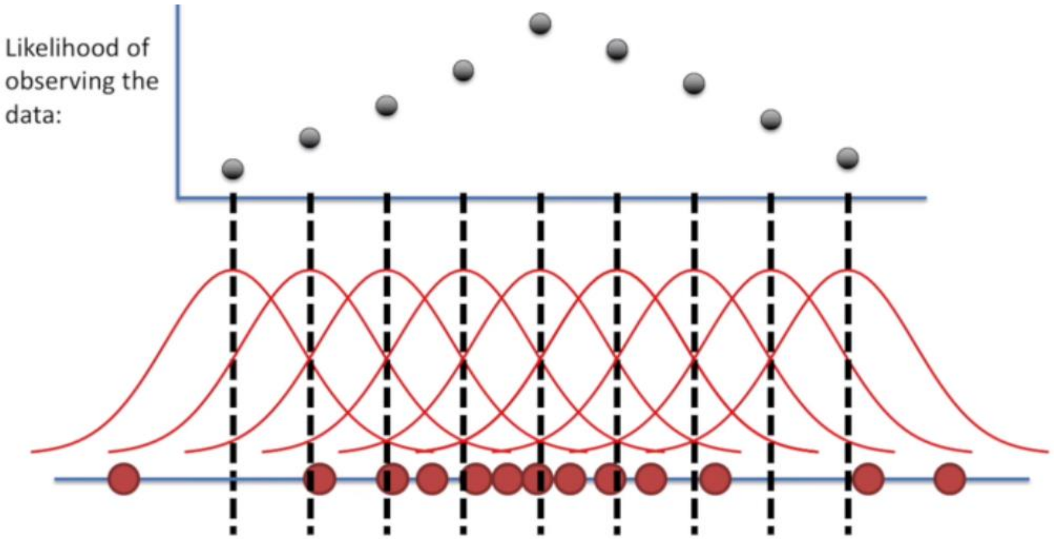
❖ **MLE(Maximum Likelihood Estimation, 최대 우도 추정)**



MLE and MLP

❖ **MLE(Maximum Likelihood Estimation,** 최대 우도 추정**)**





MLE and MLP

❖ MAP(Maximum A Posterior, 최대사후확률) Estimation

- A method of combining a given observation result with 'prior knowledge (prior probability)' to find the optimal parameter

- Advantages
  - That it reflects prior knowledge
  - Reflecting existing information rather than simply relying on observation results

- Disadvantages
  - Modeling of prior knowledge is required, and the accuracy of the posterior probability, which is the inference result, is greatly depended by modeling

MLE and MLP

❖ MAP과 MLE 적용을 위한 필요사항

$$C^* = \arg max_{C_i} \ P(C_i | X_1, X_2, \dots, X_n)$$

MAP : $C^* = \arg max_{C_i} \dfrac{P(X_1, X_2, \dots, X_n | C_i) P(C_i)}{P(X_1, X_2, \dots, X_n)}$

ML : $C^* = \arg max_{C_i} P(X_1, X_2, \dots, X_n | C_i)$

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

- **사전 확률** $P(C_i)$
  - 사전 지식 사용한 결정
  - 학습 데이터에서 부류별 비율
- **가능도** $P(X_1, X_2, \dots, X_n | C_i)$
  - 부류별 확률 분포 결정
  - 계산 부담
- **증거** $P(X_1, X_2, \dots, X_n)$
  - MAP에서 분모에 사용, 모든 부류에 대해 동일한 값
  - 계산 불필요

$$\boxed{\underbrace{P(c|x_1, x_2, \cdots . x_n)}_{\text{사후 확률}} = \frac{\overbrace{P(x_1, x_2, \cdots, x_n | c)}^{\text{가능도}} \overbrace{P(c)}^{\text{사전 확률}}}{\underbrace{P(x_1, x_2, \cdots, x_n)}_{\text{증거}}}}$$

Naïve Bayesian Classifier

❖ **Naïve Bayesian Classifier**(단순 베이즈 분류기)

- Able to handle predictors of any size, whether continuous or discrete
- Given the data, we use the posterior probabilities that would have come from this group of data to proceed with the classification

$$p(C_j|x) = \frac{p(C_j)p(x|C_j)}{p(x)}, \quad j = 1, 2, \ldots, K$$

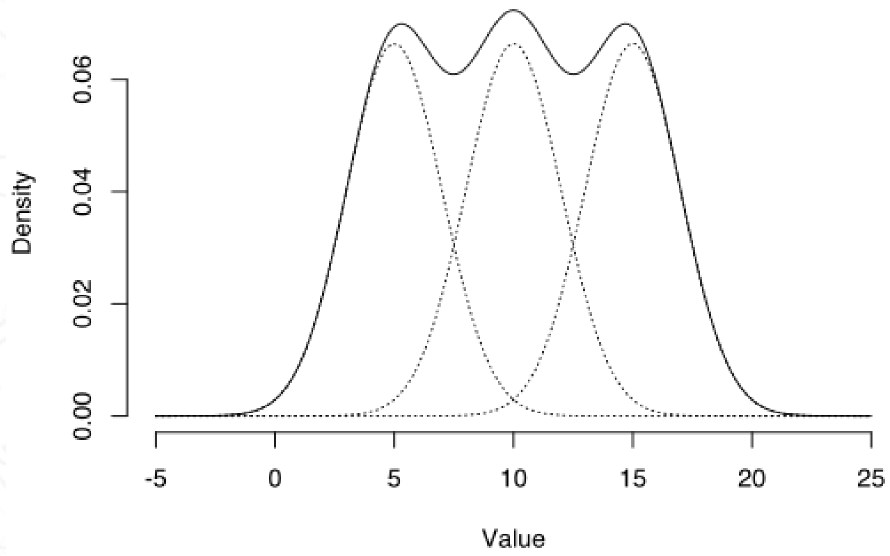
- In general Bayesian classification, the group with the greatest posterior probability is classified
- Therefore, it is assumed that each predictor is independent of each other

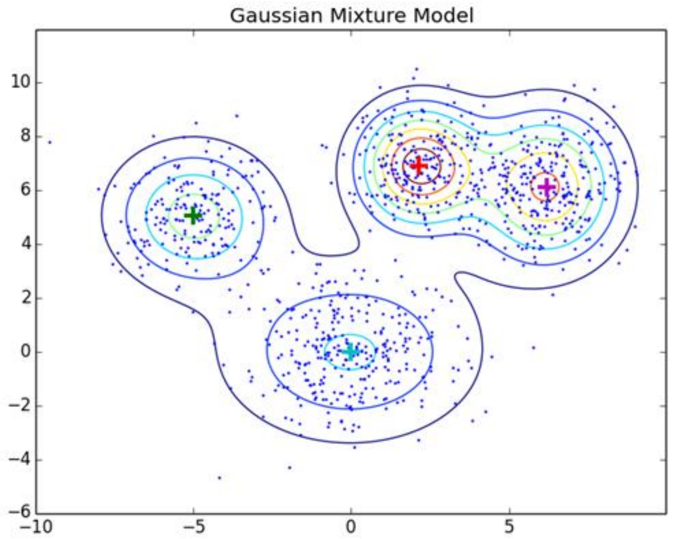
$$p(x|C_j) = p(x_1|C_j)p(x_2|C_j) \ldots p(x_d|C_j)$$

Gaussian Mixture Model

❖ Gaussian Mixture Model (가우시안 혼합 모델)

- Distribution created by linearly combining Gaussian distributions
  - Assuming a model, estimate the density by finding the parameters that go into the model





Sequential Data

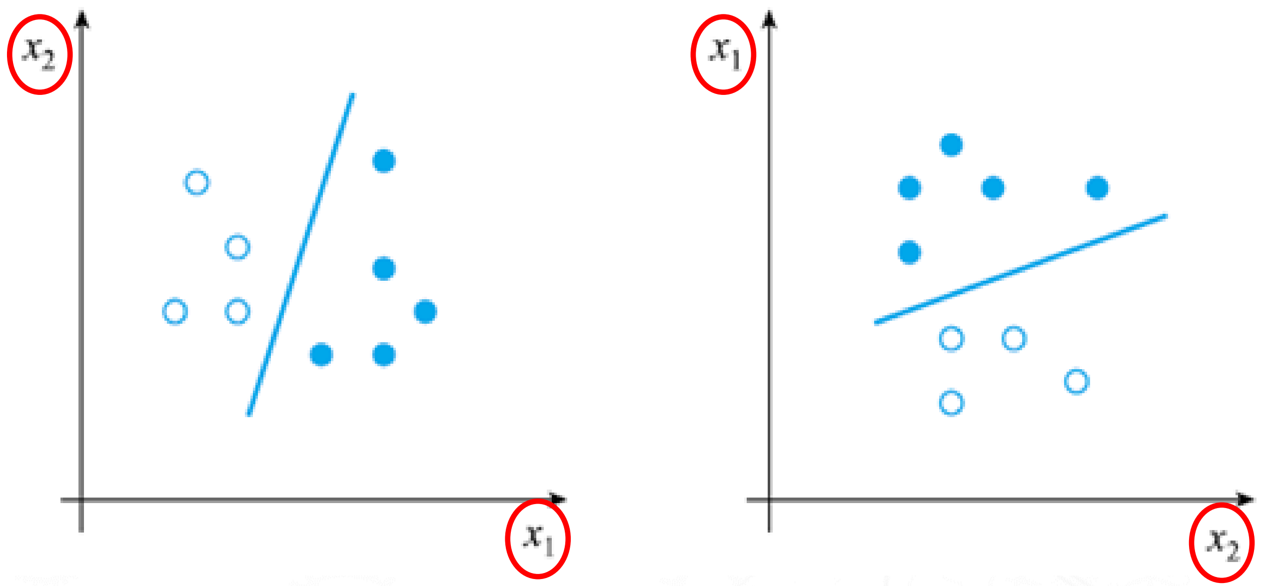
❖ Probability of Sequential Data

- Data with temporality between patterns
- Context-dependent data
- It is necessary to have a model that can properly express the temporality implied in the data and infer the desired information from it
  - Hidden Markov Model
    - It is based on probability theory, so it has a solid mathematical foundation
- Suppose a feature vector has a fixed length, and that length is denoted as d,
- However, sequential data has a variable length
  - Expressed as observation vectors instead of feature vectors
- Symbols have dependencies over time

Sequential Data

❖ Data without Sequence

- In feature vectors that are not temporal, it is not a problem for features to reposition each other



Reposition features in non-temporal data

Sequential Data

❖ Data with Sequence

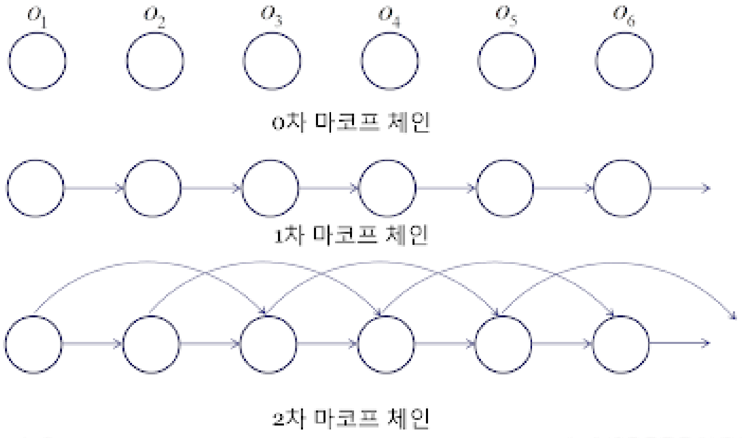
- Changing the order in which features appear distorts the physical properties of the pattern

Physical distortion when the order of features in temporal patterns is reversed

Markov Chain

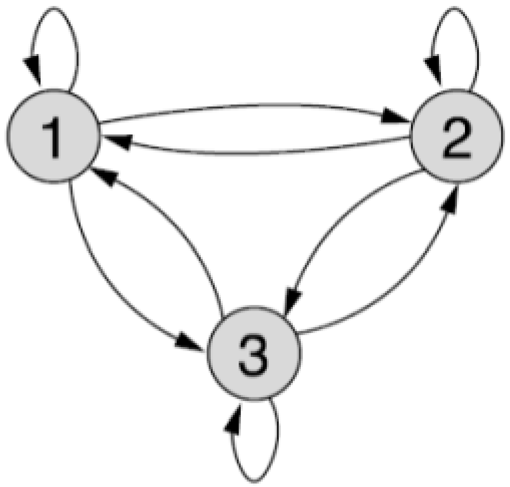
❖ Probability Matrix and Markov Chain

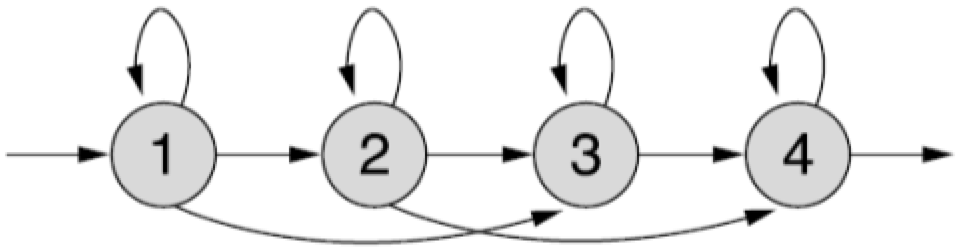
- Probability matrix : A matrix whose components are probabilities.
- Markov chain: A model in which the probability of changing from one state to another depends only on the present state rather than the traces of the past
  - The probability of a particular state depends only on the state in the past



❖ **Markov Model**

- The model generated depends on the shape of the state transition
- Ergodic model: Weather prediction
- Left-right model: Voice recognition



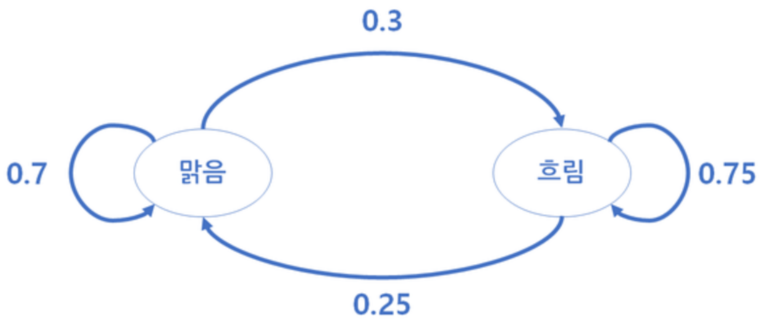


❖ Markov Model

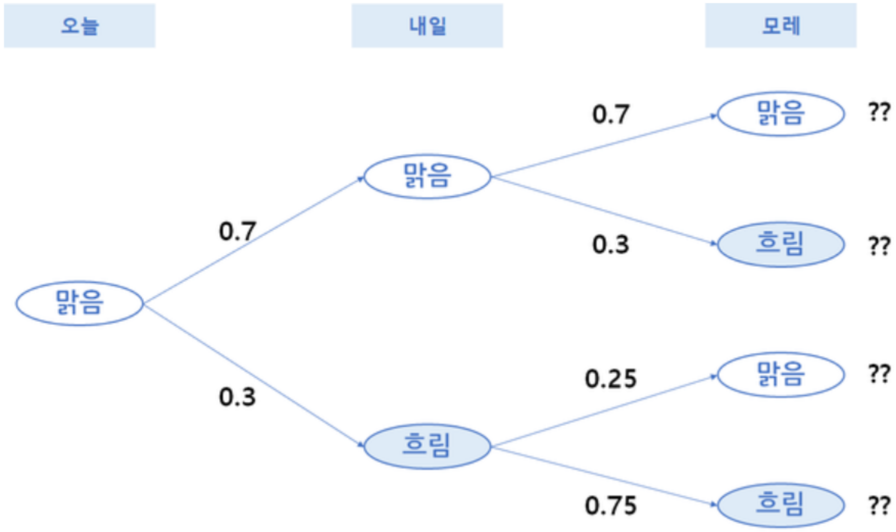
- The model generated depends on the shape of the state transition

오늘/내일	맑음	흐림
맑음	0.7	0.3
흐림	0.25	0.75

오늘 날씨에 기준하여 내일 특정 날씨가 발생할 확률



상태 전이도 예시



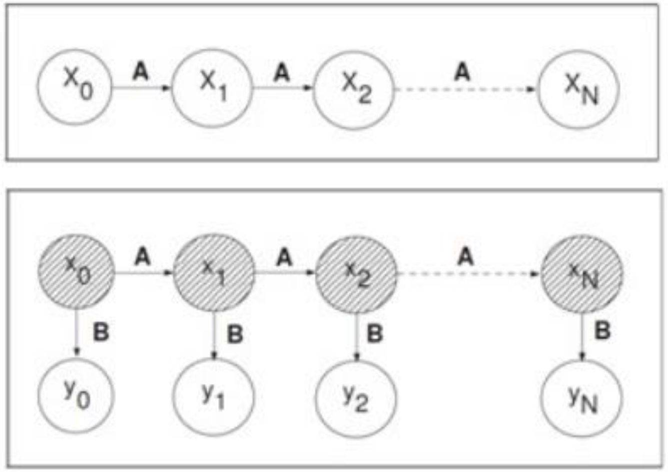
$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.25 & 0.75 \end{bmatrix}$$

$$P2 = \begin{bmatrix} 0.7 & 0.3 \\ 0.25 & 0.75 \end{bmatrix} \cdot \begin{bmatrix} 0.7 & 0.3 \\ 0.25 & 0.75 \end{bmatrix} = \begin{bmatrix} 0.565 & 0.435 \\ 0.362 & 0.637 \end{bmatrix}$$

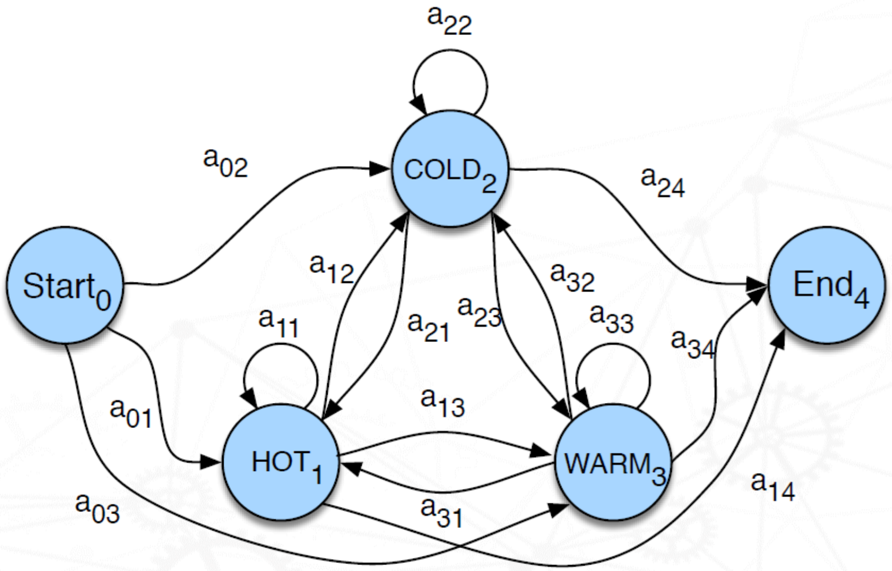
❖ Hidden Markov Model (은닉 마르코프 모델)

- One of the statistical Markov models, which considers a system to consist of two components: a hidden state and an observable outcome
- Hidden state : Direct causes that cause observable results
  - Only those states are observed as the results derived by the Markov process

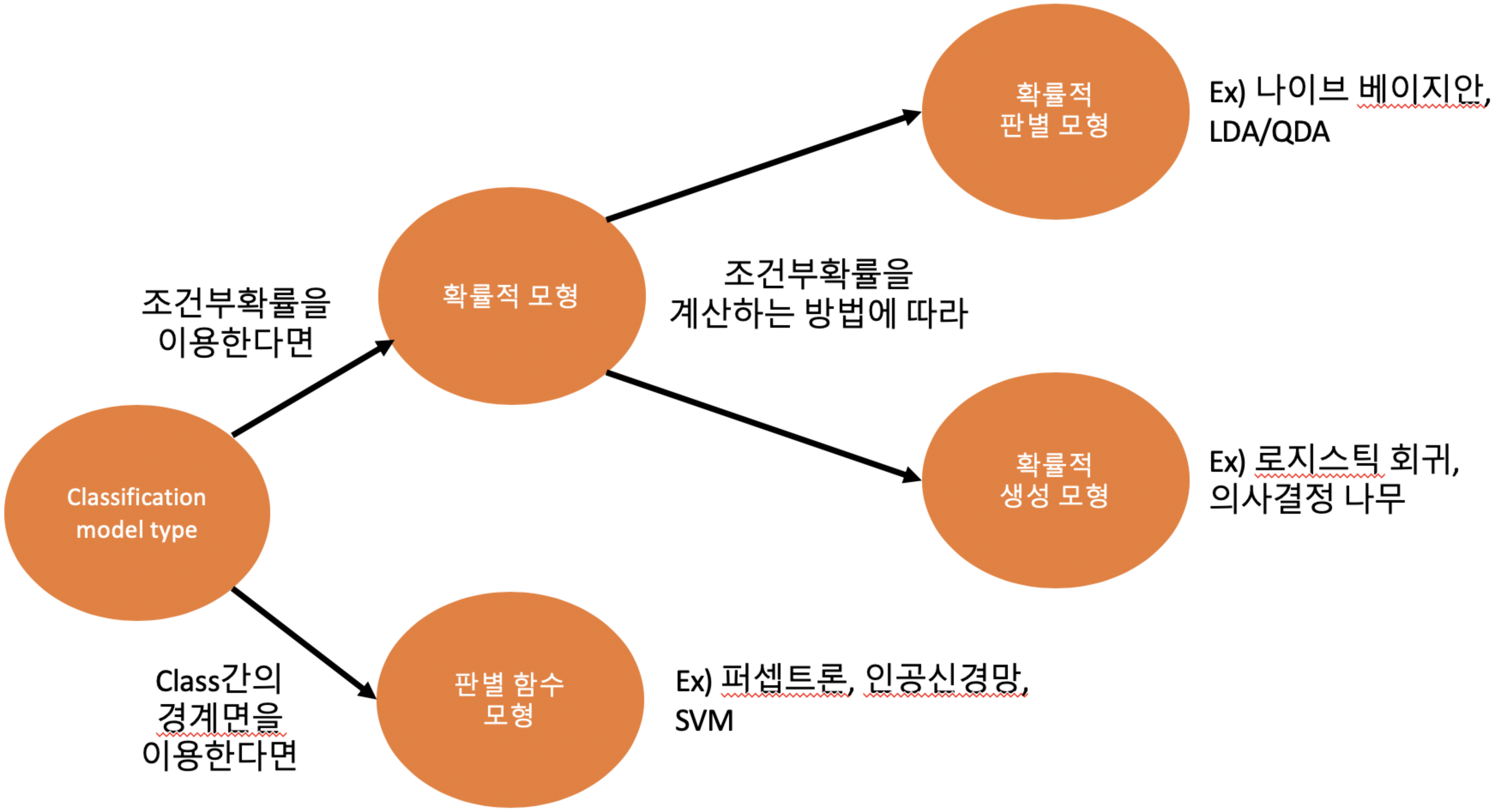
A Markov Model



A Hidden Markov Model



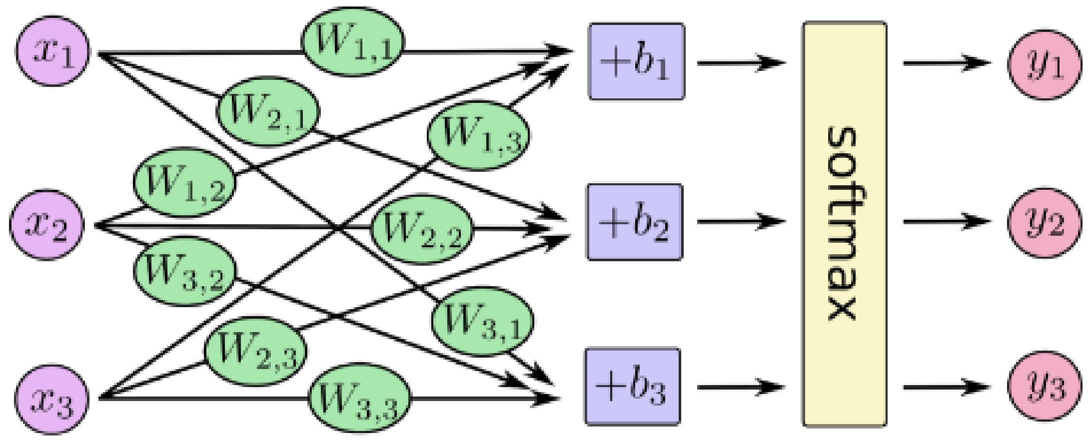
Applications



Applications

https://m.dongascience.com/news.php?idx=46883

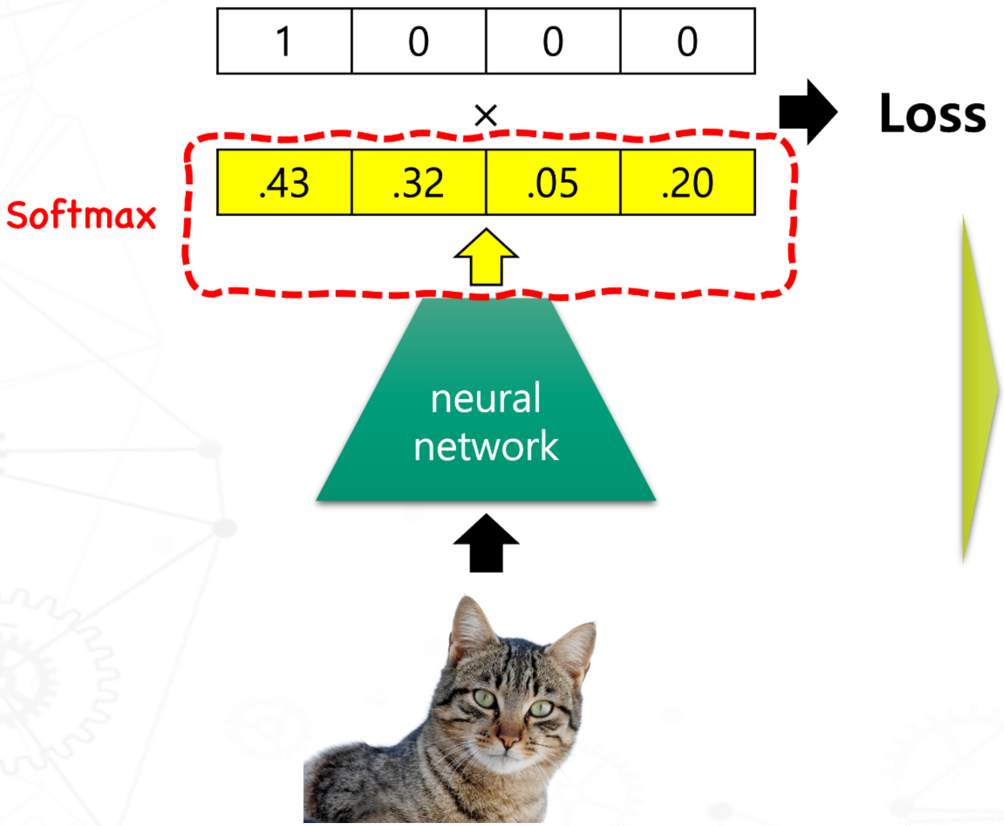
Applications

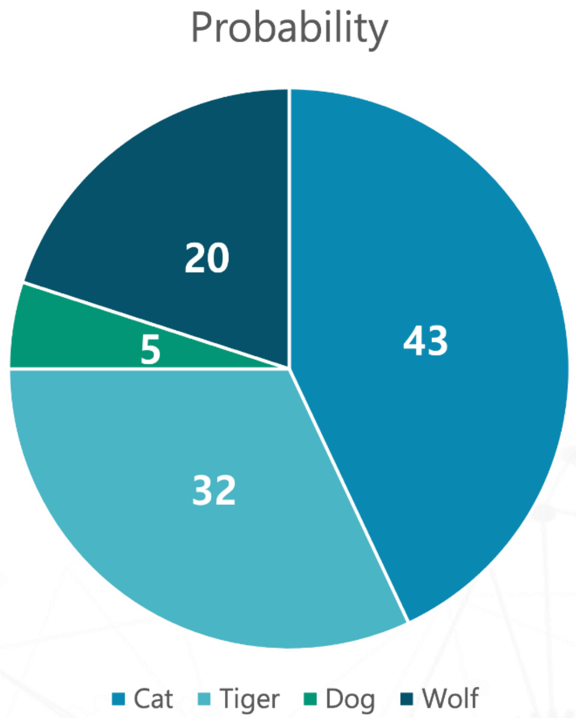


If we write it as an equation, we can get:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{softmax}\left(\begin{matrix} W_{1,1}x_1 + W_{1,2}x_1 + W_{1,3}x_1 + b_1 \\ W_{2,1}x_2 + W_{2,2}x_2 + W_{2,3}x_2 + b_2 \\ W_{3,1}x_3 + W_{3,2}x_3 + W_{3,3}x_3 + b_3 \end{matrix} \right)$$

Applications





https://kh-kim.github.io/nlp_with_deep_learning_blog/docs/1-13-deep_neural_networks_ii/05-softmax_and_cross_entropy/