머신 러닝·딥러닝에 필요한 수학 기초 with 파이썬

확률, 조건부확률

베이즈정리, 엔트로피

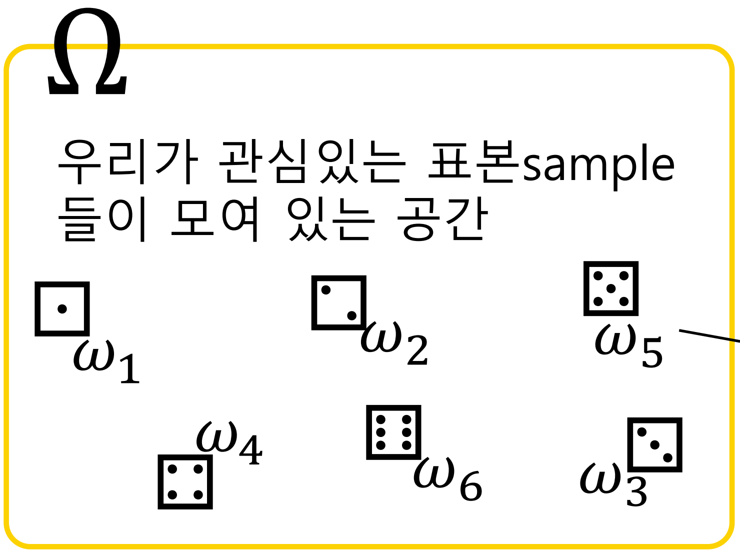
확률의 애매한 정의

- 수학적 확률
  - 어떤 시행에서 사건 A 가 일어날 가능성을 수로 나타낸 것: $P(A)$
  - 표본공간 S에서 사건 A 가 일어날 수학적 확률
    - 표본공간 S인 어떤 시행에서 각 결과가 일어날 가능성이 모두 같은 정도로 기대될 때
    - $P(A) = \frac{n(A)}{n(S)}$

- 통계적 확률
  - 일어날 가능성이 같은 정도로 기대될 수 없을 때
  - 같은 시행을 n 번 반복할 때 사건 A 가 일어난 횟수를 r_n
  - 시행 횟수 n 이 한없이 커짐에 따라 $\frac{r_n}{n}$이 일정한 값 P 에 가까워 지면
  - P 는 사건 A 의 통계적 확률

확률의 정의:표본공간

- 표본공간 Ω: 관심있는 표본들의 모임

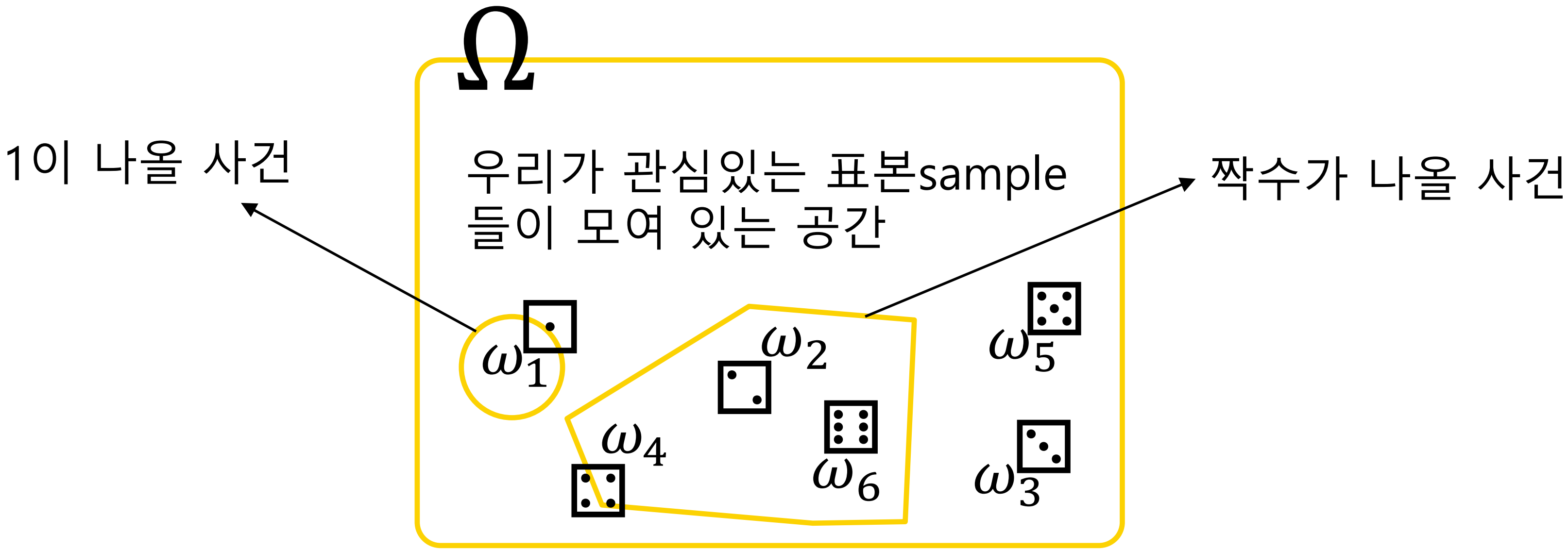


온 우주의 세상 만사
모든 것들: 하지만 관심 없음

어떤 시행에서 일어
날 수 있는 관심있는
관측치

확률의 정의:사건event

- 특정 조건을 만족시키는 표본들의 모임

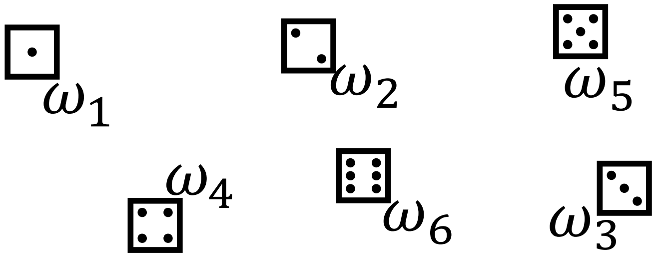


확률의 정의:사건공간

- 특정 조건을 만족시키는 표본들의 모임
- 사건들의 모임: 사건 공간event space $\mathcal{F}$

Ω

우리가 관심있는 표본sample
들이 모여 있는 공간



$\mathcal{F}$

$\{\omega_1, \omega_3, \omega_5\}$

$\{\omega_2, \omega_4, \omega_6\}$

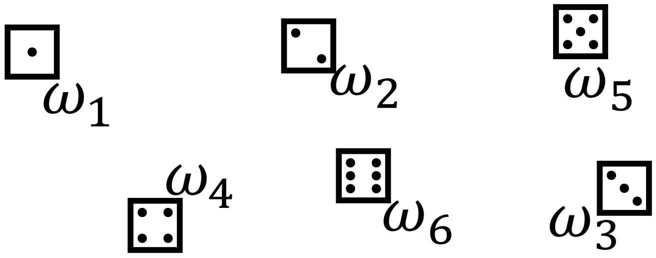
$\{\ \ \}$ $\qquad \Omega$

확률의 정의:확률

- 확률함수: 사건을 입력 받고 숫자를 출력하는 함수
- $\mathcal{F}$에 있는 모든 사건에 숫자를 할당하는 함수, 표본 ω에 숫자를 할당하는 것이 아님을 주의
- 아무렇 게나 할당?

Ω

우리가 관심있는 표본sample 들이 모여 있는 공간



$\mathcal{F}$

$$P(\{\omega_1, \omega_3, \omega_5\}) =?$$

$$P(\{\omega_2, \omega_4, \omega_6\}) =?$$

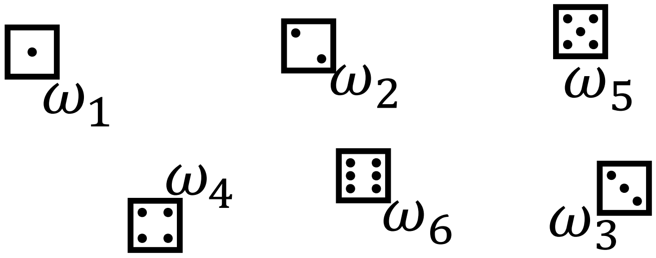
$$P(\{\ \}) = 0 \quad P(\Omega) = 1$$

확률의 정의:공리

- 각 확률은 0보다 크거나 같다.
- $P(\Omega) = 1$
- $i \neq j$일 때 임의의 두 사건 E_i, E_j 가 $E_i \cap E_j = \emptyset$ 이면 $P(E_i \cup E_j) = P(E_i) + P(E_j)$

Ω

우리가 관심있는 표본sample 들이 모여 있는 공간



$\mathcal{F}$

$$P(\{\omega_1, \omega_3, \omega_5\}) = 0.5$$

$$P(\{\omega_2, \omega_4, \omega_6\}) = 0.5$$

$$P(\{\ \ \}) = 0 \quad P(\Omega) = 1$$

확률의 정의:예

- 동전을 던지는 표본공간에서

Ω

H T

$\mathcal{F}$

$\{\}$ Ω

$\{H\}$ $\{T\}$

$\mathcal{F}$

$P(\{\}) = 0$ $P(\Omega) = 1$

$P(\{H\}) = 0.3$ $P(\{T\}) = 0.7$

확률은 면적

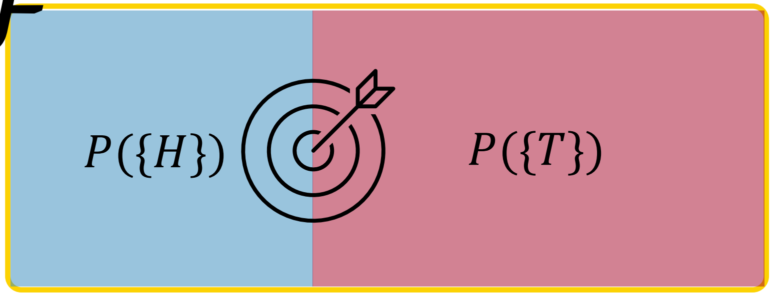
- 동전을 던지는 표본공간에서

$\mathcal{F}$

$P(\{\}) = 0$ $\quad$ $P(\Omega) = 1$

$P(\{H\}) = 0.3$ $\quad$ $P(\{T\}) = 0.7$

$\mathcal{F}$

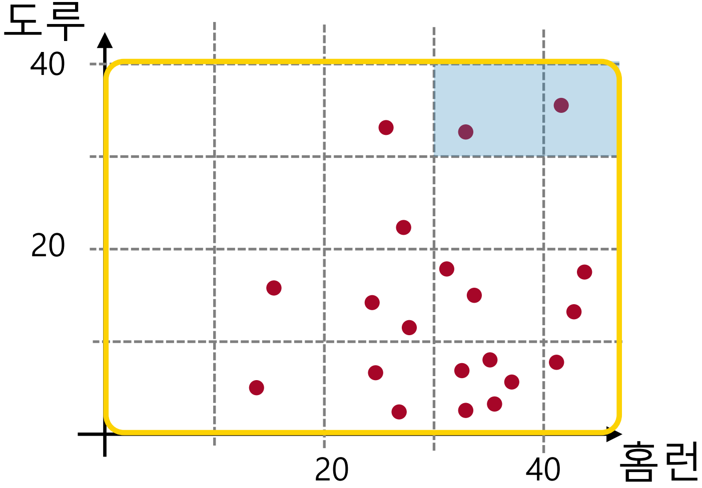


조건부 확률

- 두 사건 A, B에 대해
- 결합확률Joint Probability: 두 사건이 동시에 일어날 확률 $P(A,B)$
- 조건부확률Conditional Probability: 사건 A가 일어났을 때 사건B가 일어날 확률 $P(B|A)$

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

A: 홈런 30개 이상 12번 $P(A) = 12/20$
B: 도루 30개 이상 3번
$A \cap B$: 30,30 이상 2번 $P(A,B) = 2/20$

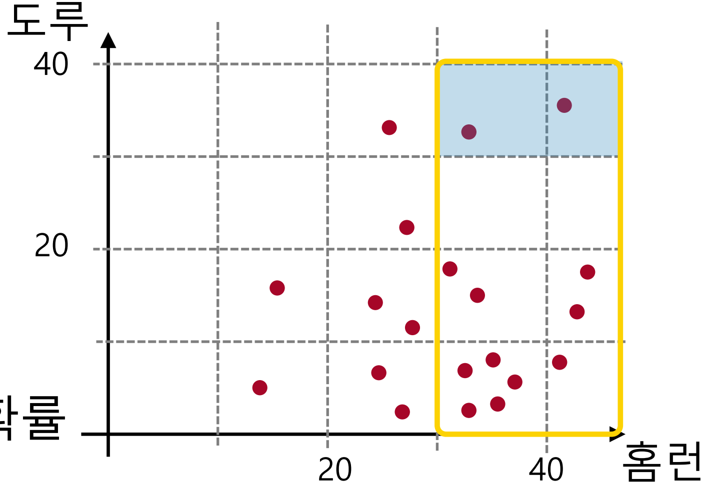


조건부 확률

- 두 사건 A, B에 대해
- 결합확률Joint Probability: 두 사건이 동시에 일어날 확률 $P(A,B)$
- 조건부확률Conditional Probability: 사건 A가 일어났을 때 사건B가 일어날 확률 $P(B|A)$

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

A: 홈런 30개 이상 12번 $P(A) = 12/20$
B: 도루 30개 이상 3번
$A \cap B$: 30,30 이상 2번 $P(A,B) = 2/20$



홈런 30개 이상 쳤을 때 도루 30개 이상할 확률
$P(B|A) = 2/12$, 분자 분모를 20으로 나누면

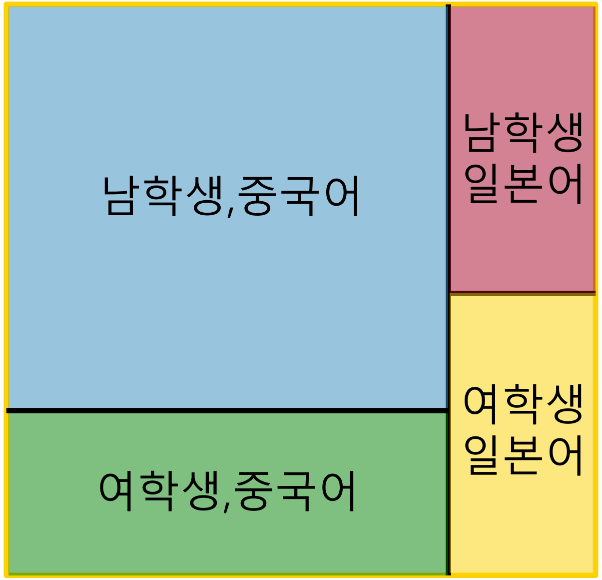
조건부 확률

- P(남학생,중국어)=45/100
- P(남학생|중국어)=45/70

	중국어	일본어
남학생 수	45	15
여학생 수	25	15

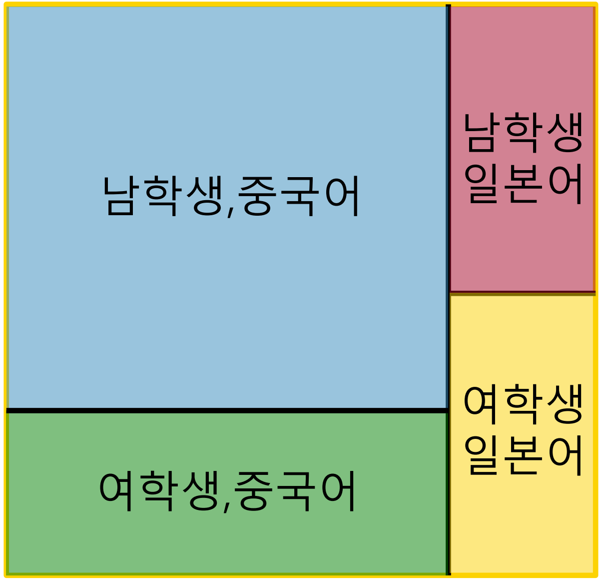
조건부 확률

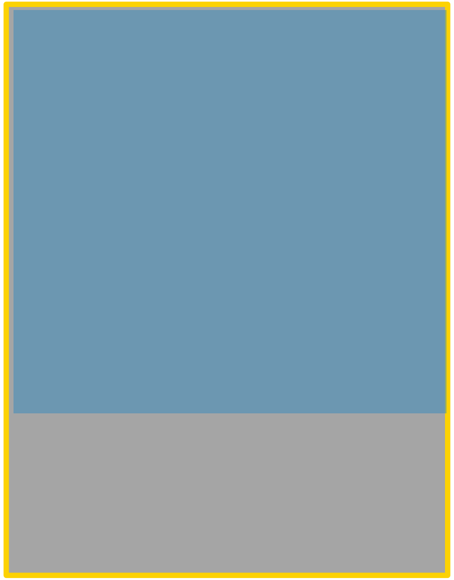
- P(남학생, 중국어)=45/100



조건부 확률

- P(남학생|중국어)=45/70





확률의 곱법칙

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

양변에 $P(A)$를 곱함

$$P(B|A)P(A) = P(A,B)$$

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

$$P(A|B)P(B) = P(A,B)$$

양변에 $P(B)$를 곱함

$$P(A,B) = P(B|A)P(A) = P(A|B)P(B)$$

공 10개중에 빨간공이 5개 있는데 두번 연속 빨간공을 꺼낼 확률?

$$\underbrace{\frac{5}{10}}_{P(A)} \times \underbrace{\frac{4}{9}}_{P(B|A)}$$

9 → 표본공간이 줄어듦

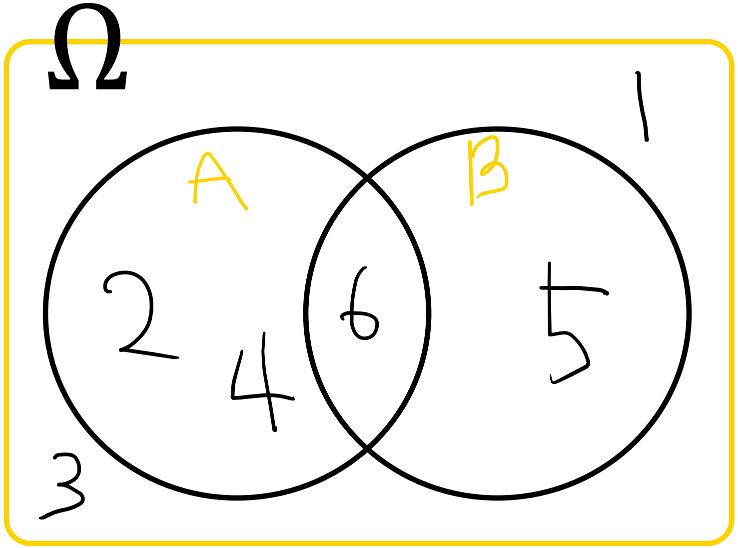
사건의 독립

- 두 사건 A, B에 대하여
- 사건 A가 일어나거나 일어나지 않는 것이 사건 B가 일어날 확률에 영향을 주지 않을 때

$$P(A,B) = P(B|A)P(A) = P(A|B)P(B)$$

$$P(B) \qquad\qquad P(A)$$

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



결합확률? 조건부확률?

- 10개중 3개 당첨 한 개씩 두 번 뽑을 때 당첨 제비가 한 개만 뽑힐 확률?
- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Ω

(1,2), (1,3)
(2,1), (2,3)
(3,1), (3,2)

6

21

(1,4), (1,5), (1,6), (1,7), (1,8), (1,9), (1,10)
(2,4), (2,5), (2,6), (2,7), (2,8), (2,9), (2,10)
(3,4), (3,5), (3,6), (3,7), (3,8), (3,9), (3,10)

(4,1), (4,2), (4,3)
(5,1), (5,2), (5,3)
(6,1), (6,2), (6,3)
(7,1), (7,2), (7,3)
(8,1), (8,2), (8,3)
(9,1), (9,2), (9,3)
(10,1), (10,2), (10,3)

21

(4,5), (4,6), (4,7), (4,8), (4,9), (4,10)
(5,4), (5,6), (5,7), (5,8), (5,9), (5,10)
(6,4), (6,5), (6,7), (6,8), (6,9), (6,10)
(7,4), (7,5), (7,6), (7,8), (7,9), (7,10)
(8,4), (8,5), (8,6), (8,7), (8,9), (8,10)
(9,4), (9,5), (9,6), (9,7), (9,8), (9,10)
(10,4), (10,5), (10,6), (10,7), (10,8), (10,9)

42

결합확률? 조건부확률?

- 20개 제비 가운데 4개 당첨
- 윤경이 뽑고 나서 건우가 뽑는데
- 윤경이의 당첨 확률과 건우의 당첨 확률은?

- 구하고 싶은 것: $P(\text{윤경당첨})$, $P(\text{건우당첨})$

$$P(\text{윤경당첨}) = \frac{4}{20} = \frac{1}{5}$$

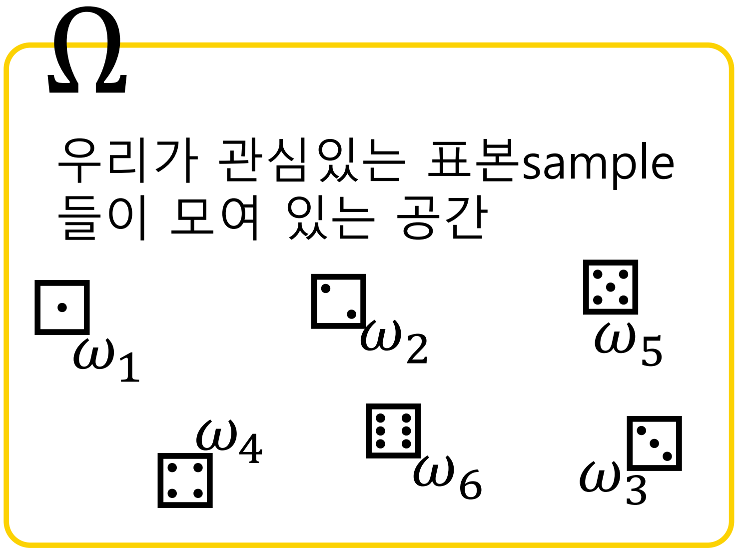
$$P(\text{건우당첨}) = \frac{4}{19} \times \frac{4}{5} \quad + \quad \frac{3}{19} \times \frac{1}{5}$$

$$\frac{4}{19} \qquad\qquad \frac{3}{19}$$

이 두 수가 무엇을 의미?

확률변수: 이산 확률변수

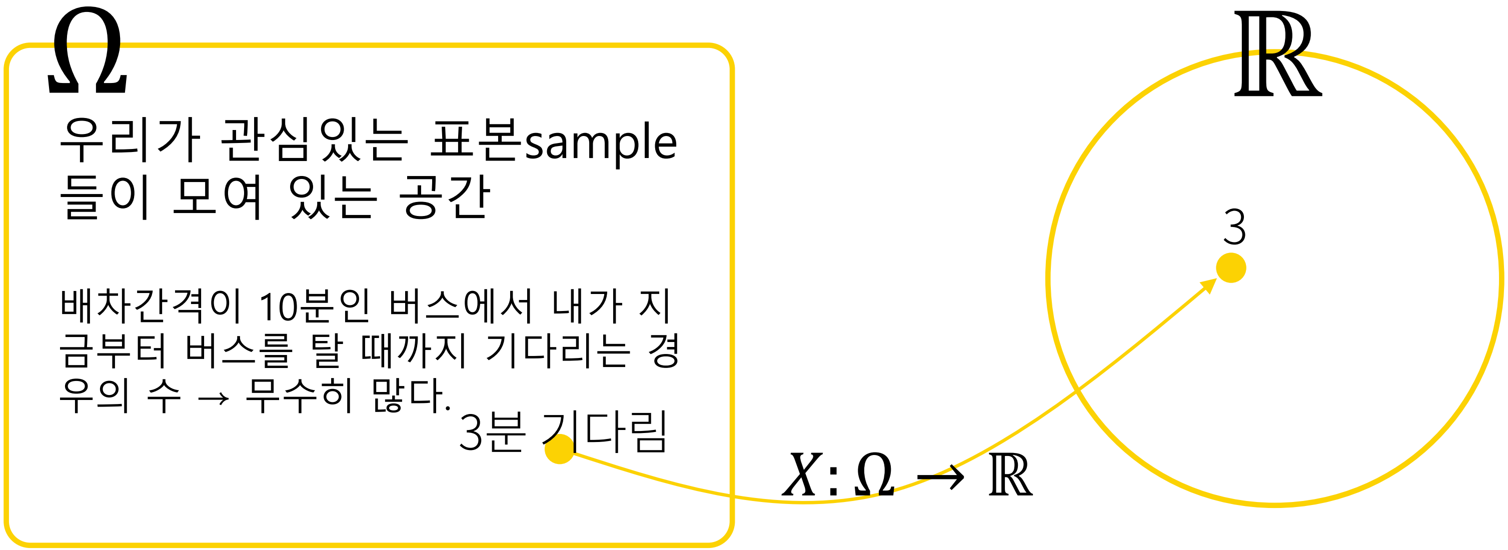
- 표본공간의 샘플에 숫자를 할당하는 함수
- 이산 확률변수 $X: \Omega = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\} \rightarrow \mathbb{R}$



$X(⚀) = 1$	$Y(⚀) = 1$	$Z(⚀) = 0$
$X(⚁) = 2$	$Y(⚁) = 0$	$Z(⚁) = 0$
$X(⚂) = 3$	$Y(⚂) = 1$	$Z(⚂) = 1$
$X(⚃) = 4$	$Y(⚃) = 0$	$Z(⚃) = 0$
$X(⚄) = 5$	$Y(⚄) = 1$	$Z(⚄) = 0$
$X(⚅) = 6$	$Y(⚅) = 0$	$Z(⚅) = 1$

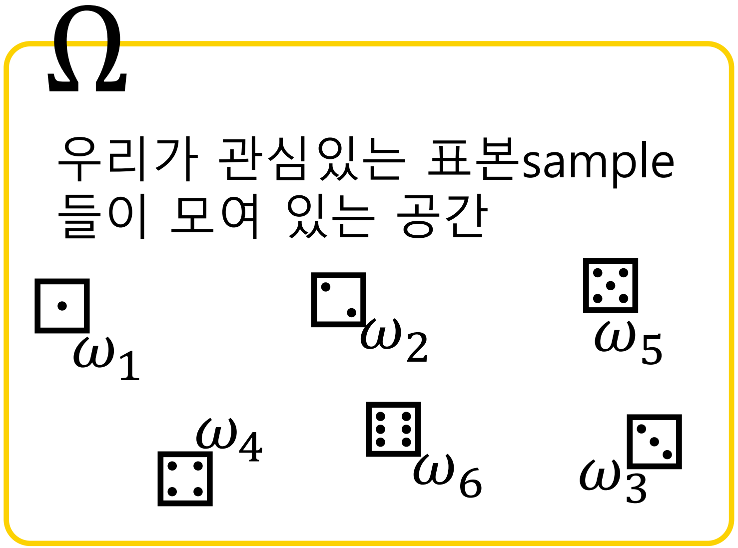
확률변수: 연속 확률변수

- 표본공간의 샘플에 숫자를 할당하는 함수
- 연속 확률변수 $X: \Omega \rightarrow \mathbb{R}$



확률분포

- 확률변수 X가 가질 수 있는 값과 확률의 대응 관계
  - 다시 말해 X가 가질 수 있는 값에 확률이 얼만 할당되었는가를 나타낸 확률의 펼쳐짐 정도
  - $P(X = x)$: 확률변수 X가 x값을 가질 확률
- 이산 확률변수



$X(⚀) = 1$

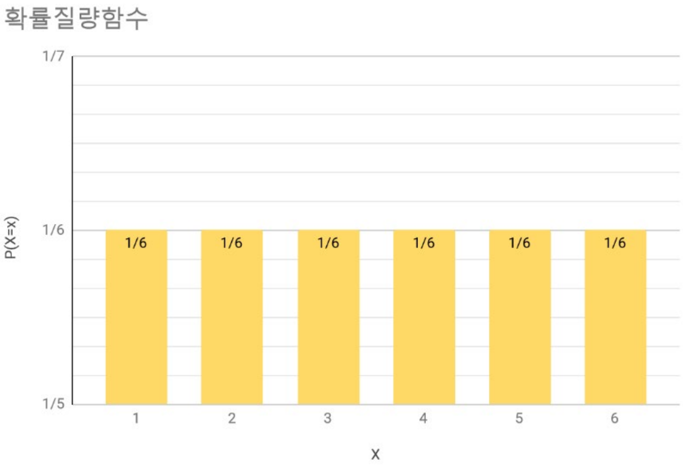
$X(⚁) = 2$

$X(⚂) = 3$

$X(⚃) = 4$

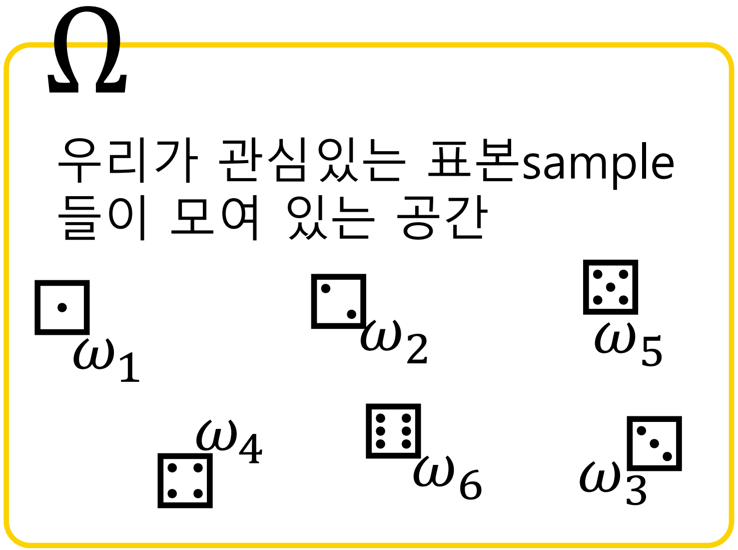
$X(⚄) = 5$

$X(⚅) = 6$



확률 질량함수

- 이산 확률변수의 다른 예



Z(⚀) = 0

Z(⚁) = 0

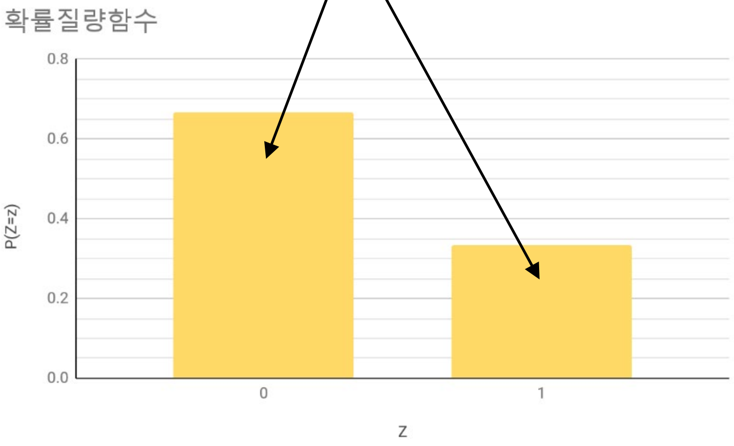
Z(⚂) = 1

Z(⚃) = 0

Z(⚄) = 0

Z(⚅) = 1

확률이 특정값에 덩어리처럼 몰려 있다. (질량)



연속 확률변수의 확률분포

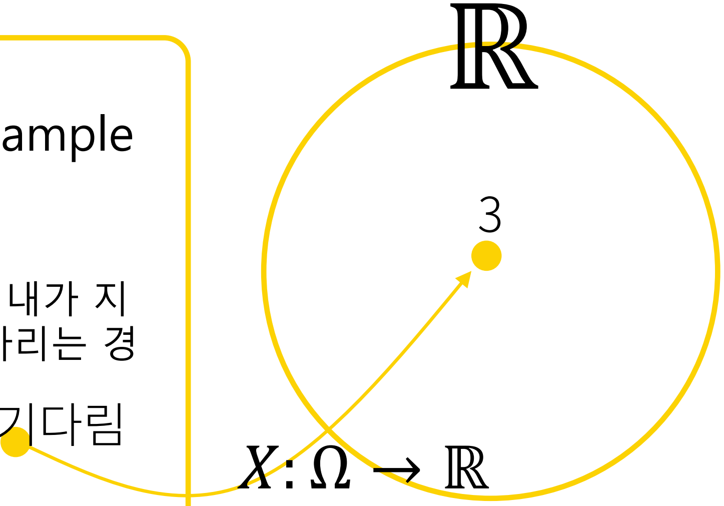
- 표본공간의 샘플에 숫자를 할당하는 함수
- 연속 확률변수 $X: \Omega \to \mathbb{R}$

Ω

우리가 관심있는 표본sample 들이 모여 있는 공간

배차간격이 10분인 버스에서 내가 지금부터 버스를 탈 때까지 기다리는 경우의 수 → 무수히 많다.

3분 기다림



$P(X = 3) = 0.3$

$P(X = 3.1) = ?$

$P(X = 3.01) = ?$

$P(X = 3.001) = ?$

$\sum_{x \in \Omega} P(X = x) \overset{?}{=} 1$

연속 확률변수의 누적 분포함수

- $F_X(x)$ 누적분포함수Cumulative distribution function
  - 확률변수 "구간"에 확률을 할당

$$F_X(x) = P(X < x)$$

Ω

우리가 관심있는 표본sample 들이 모여 있는 공간

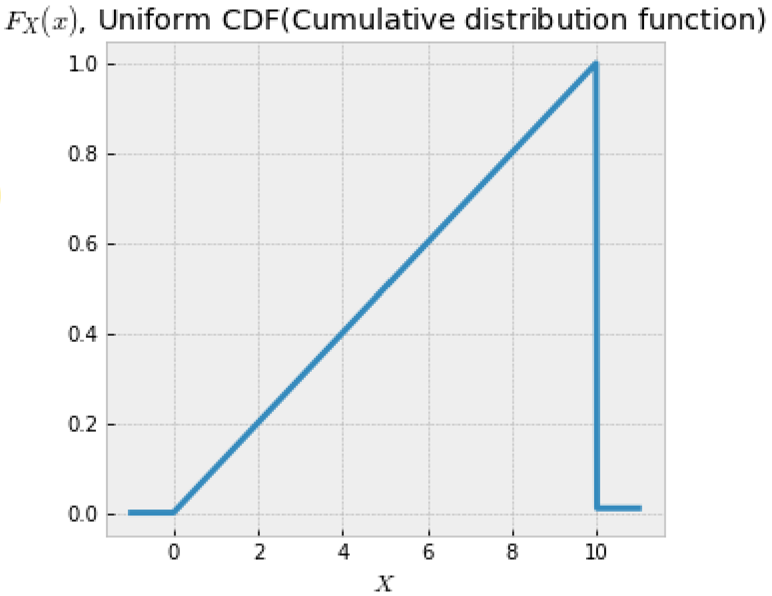
배차간격이 10분인 버스에서 내가 지금부터 버스를 탈 때까지 기다리는 경우의 수 → 무수히 많다.

3분 기다림

$\mathbb{R}$

3

$X: \Omega \to \mathbb{R}$



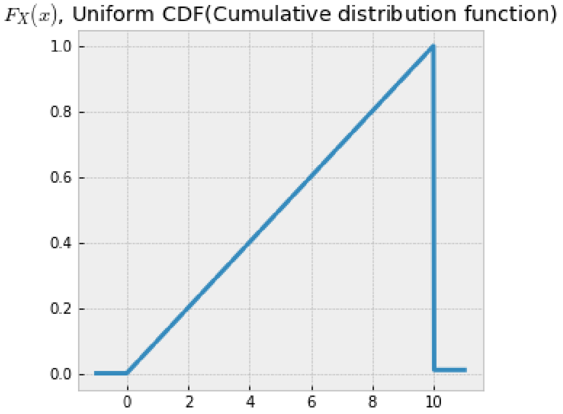
연속 확률변수의 확률 밀도함수

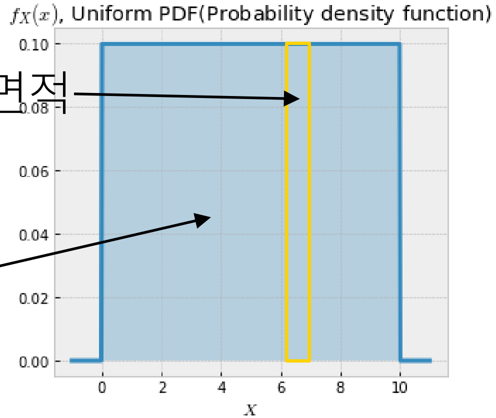
- $F_X(x)$에서 구간에 확률이 얼마나 할당되었는지 알고 싶은데?
- 얼마만큼 큰 구간에 할당된 확률을 볼까?→0.01!

$$dF_X(x) = \frac{F_X(x+0.01) - F_X(x)}{0.01}$$

- 0.01로 만족? 어디서 많이 본 것 같은...

$$f_X(x) = \lim_{h\to 0} \frac{F_X(x+h) - F_X(x)}{h}$$





확률: 구간에 대한 면적

확률이 전구간에 골고루
퍼져 있다. (밀도)

평균

- 평균
  - N개를 다 더해서 N으로 나눔

상금(원)	0	10,000	30,000	50,000	합계
행운권 수	872	100	20	8	1,000

  - 행운권 한 장 당 평균 상금: 총 상금/행운권 수

$$\frac{0 \times 872 + 10000 \times 100 + 30000 \times 20 + 50000 \times 8}{1000} = 2000$$

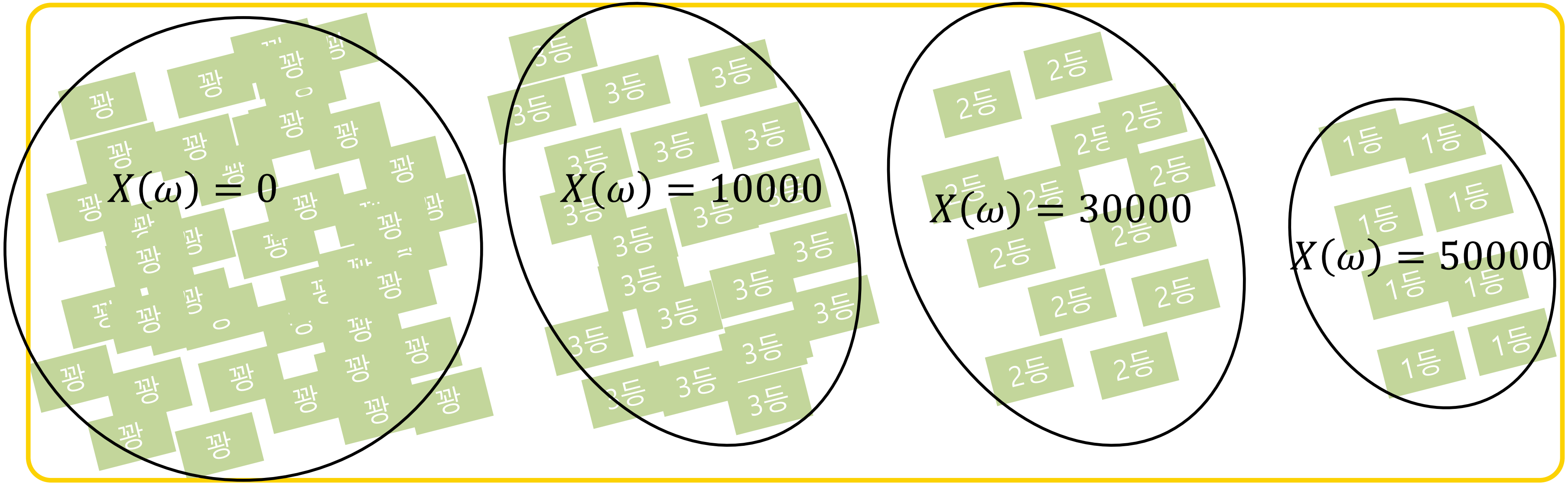
$$0 \times \frac{872}{1000} + 10000 \times \frac{100}{1000} + 30000 \times \frac{20}{1000} + 50000 \times \frac{8}{1000} = 2000$$

확률변수의 평균

- 평균
  - 행운권 한 장 당 평균 상금: 총 상금/행운권 수

$$\mathbb{E}[X] = \sum_{i=1}^{N} x_i p_i$$

$$0 \times \frac{872}{1000} + 10000 \times \frac{100}{1000} + 30000 \times \frac{20}{1000} + 50000 \times \frac{8}{1000} = 2000$$



확률변수의 평균, 분산, 표준편차

- 기댓값(평균)

$$\mathbb{E}[X] = \sum_{i=1}^{N} x_i p_i$$

- 분산: 편차 제곱의 평균

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{i=1}^{N} (x_i - \mathbb{E}[X])^2 p_i$$

- 표준편차

$$\sigma(X) = \sqrt{\mathbb{V}[X]}$$

$$\begin{aligned}
\mathbb{V}[X] &= \sum_{i=1}^{N} (x_i - \mathbb{E}[X])^2 p_i \\
&= \sum_{i=1}^{N} (x_i^2 - 2\mathbb{E}[X] x_i + \mathbb{E}[X]^2) p_i \\
&= \sum_{i=1}^{N} x_i^2 p_i - 2\mathbb{E}[X] \sum_{i=1}^{N} x_i p_i + \mathbb{E}[X]^2 \sum_{i=1}^{N} p_i \\
&= E[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\
&= E[X^2] - \mathbb{E}[X]^2
\end{aligned}$$

확률의 합법칙

$$P(X) = \sum_Y p(X,Y)$$
$$= \sum_Y p(X|Y)p(Y)$$

$$p(X = \text{남}) = p(X = \text{남}, Y = \text{중}) + p(X = \text{남}, Y = \text{일}) = \frac{45}{100} + \frac{15}{100} = \frac{60}{100}$$

$$p(X = \text{남}) = p(X = \text{남}|Y = \text{중})p(Y = \text{중}) + p(X = \text{남}|Y = \text{일})p(Y = \text{일})$$
$$= \frac{45}{70}\frac{70}{100} + \frac{15}{30}\frac{30}{100} = \frac{60}{100}$$

언어 사라짐

	중국어	일본어	
남학생 수	45 p(X=남,Y=중)	15 p(X=남,Y=일)	60
여학생 수	25 p(X=여,Y=중)	15 p(X=여,Y=일)	40
성별 사라짐 →	70	30	

$$p(X = \text{남}) = \frac{60}{100}$$
$$p(X = \text{여}) = \frac{40}{100}$$

$$p(Y = \text{중국어}) = \frac{70}{100} \quad p(Y = \text{일어}) = \frac{30}{100}$$

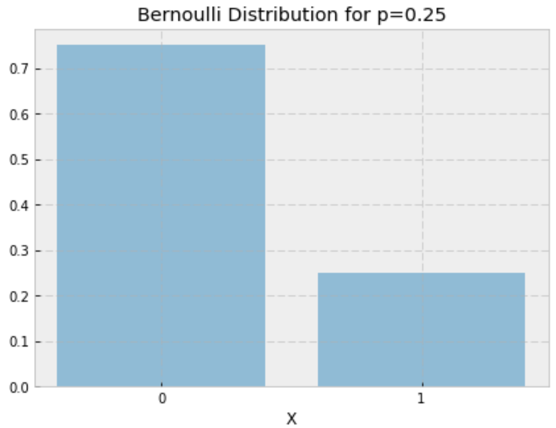
확률분포: 베르누이 분포

- 베르누이 분포Bernoulli distribution
  - 0 또는 1을 값으로 가지는 바이너리binary 확률변수에 대한 분포. 즉 $x \in \{0, 1\}$
  - 예: 동전 던지기, 동전의 앞면(Head)이 나오면 $x = 1$, 뒷면(Tail)이 나오면 $x = 0$

- 확률질량함수
  - $x \in \{0, 1\}$이므로 $x = 1$일 확률은 $p(x = 1|\mu) = \mu$, $x = 0$일 확률은 $p(x = 0|\mu) = 1 - \mu$

$$\mathrm{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

▷바이너리 확률변수에 값을 할당하게 하는 행위: 베르누이 시행bernoulli trial,bernoulli experiment



확률분포: 이항분포

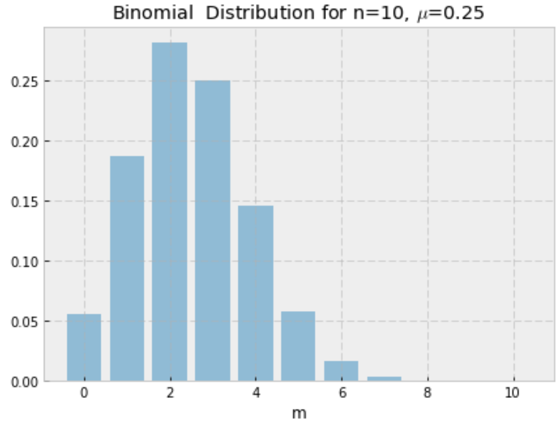
- 이항분포Binomial distribution
  - 베르누이 시행을 N번 독립적으로 해서 $x = 1$이 일어난 횟수 m이라 할 때 m에 대한 분포
  - 예: 동전을 열 번 던졌더니 $(N = 10)$ 앞면이 3번 나왔더라.$(m = 3)$
  - m은 $0 \sim N$이 될 수 있는데 각 숫자에 얼마나 확률이 할당되어 있는가에 대한 이야기

- 확률질량함수

경우의 수

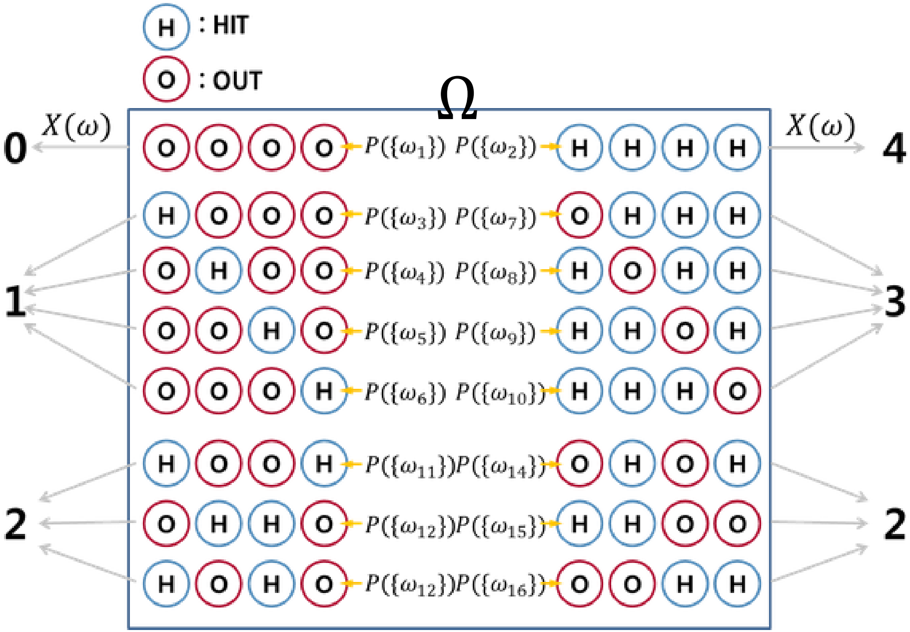
$$\mathrm{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

여러 번 독립 시행에서 발생한 곱



이항분포 예제

- 타율 0.250인 타자가 네 번 타석에 들어서서 안타를 친 횟수 X



$$P(\{\omega_1\}) = \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{81}{256}$$

$$P(\{\omega_2\}) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{256}$$

$$P(\{\omega_j\}) = \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{27}{256} \qquad j = 3, 4, 5, 6$$

$$P(\{\omega_j\}) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{256} \qquad j = 7, 8, 9, 10$$

$$P(\{\omega_j\}) = \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{9}{256} \qquad j \in [11, 16]$$

$$\sum_{j=1}^{16} P(\{\omega_j\}) = 1$$

$$p(X = x) = \binom{4}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{4-x}$$

확률분포: 정규분포

- 정규분포

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

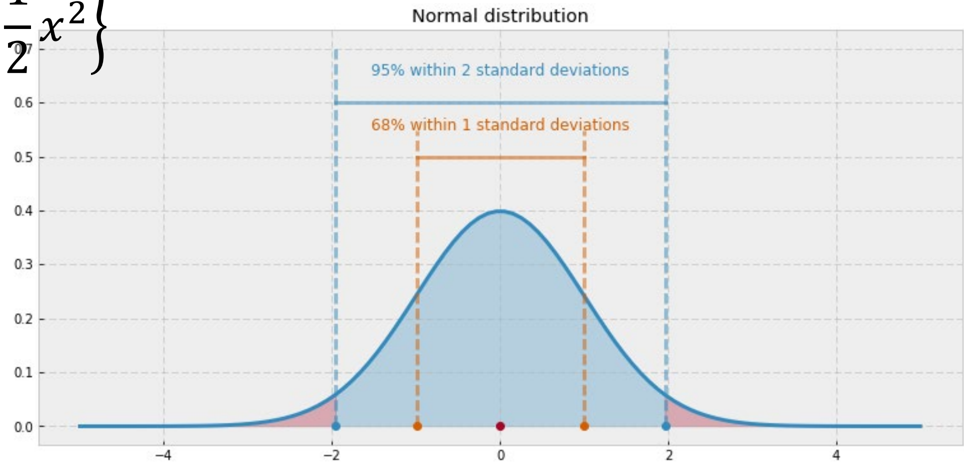
- 표준 정규분포: 평균 0, 분산 1인 정규분포

$$\mathcal{N}(x|\mu=0,\sigma^2=1) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}$$

- 표준화Standardization

$$X \backsim \mathcal{N}(\mu,\sigma^2)$$

$$Z = \frac{X-\mu}{\sigma} \qquad Z \backsim \mathcal{N}(0,1)$$



베이즈정리Bayes theorem

- 조건부 확률 정의에 의해

$$p(\mu \mid \mathcal{D}) = \frac{\overbrace{p(\mathcal{D}, \mu)}^{\text{결정해야할 파라미터}}}{\underbrace{p(\mathcal{D})}_{\text{데이터}}}$$

- 곱셈 정리에 의해

$$p(\mathcal{D}, \mu) = p(\mu \mid \mathcal{D})p(\mathcal{D})$$

$$p(\mathcal{D}, \mu) = p(\mathcal{D} \mid \mu)p(\mu)$$

$$\underbrace{p(\mu \mid \mathcal{D})}_{\substack{\text{사후확률분포} \\ \textit{Posterior}}} = \frac{\overbrace{p(\mathcal{D} \mid \mu)}^{\text{가능도}\textit{Likelihood}}\underbrace{p(\mu)}_{\substack{\text{사전확률분포} \\ \textit{Prior}}}}{p(\mathcal{D})}$$

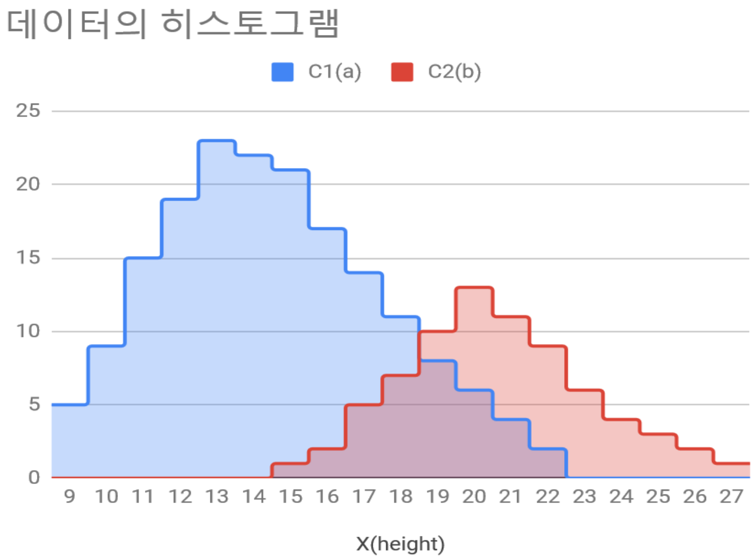
베이즈 정리를 이용한 분류

- a, b를 높이 값으로 분류*

a, b

a, b

a, b



X(height)	C1(a)	C2(b)
9	5	0
10	9	0
11	15	0
12	19	0
13	23	0
14	22	0
15	21	1
16	17	2
17	14	5
18	11	7
19	8	10
20	6	13
21	4	11
22	2	9
23	0	6
24	0	4
25	0	3
26	0	2
27	0	1

*Machine Learning An Algorithm Perspective 2nd Ed., Stephen Marsland, CRC Press

가능도Likelihood

- $p(\mu\,|\mathcal{D})$, $p(\mathcal{D}\,|\,\mu)$ 둘 다 조건부 '확률'인데 왜 후자는 확률이라 하지 않고 가능도라고 하는가?

가능도*Likelihood*

$$p(\mu \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mu)p(\mu)}{p(\mathcal{D})}$$

사전확률분포
Prior

사후확률분포
Posterior

가능도함수Likelihood func.

- μ 를 바꾸면서 $p(\mathcal{D} \mid \mu)$ 값을 계산하면 각각 다른 μ 를 조건으로 하는 조건부 확률분포함수의 함숫값이 구해짐
- 즉, 그렇게 구한 함숫값들의 모임은 μ에 대한 함수의 함숫값이고 이 함수는 확률함수나 확률분포함수가 아님
- 다시 말해 특정 $p(\mathcal{D} \mid \mu)$ 값은 확률분포함수의 함숫값과 일치하지만 μ를 변수로 보기 때문에 전혀 다른 함수→가능도함수

고정된 데이터, 확률을 알고 싶은 대상

$p(\mathcal{D} \mid \mu)$

변경되는 확률분포함수의 파라미터
$\mathcal{D}$ 의 확률을 결정하는 함수가 $p()$ 인데 $p()$의 모양을 조정하는 숫자

최대 가능도 추정MLE

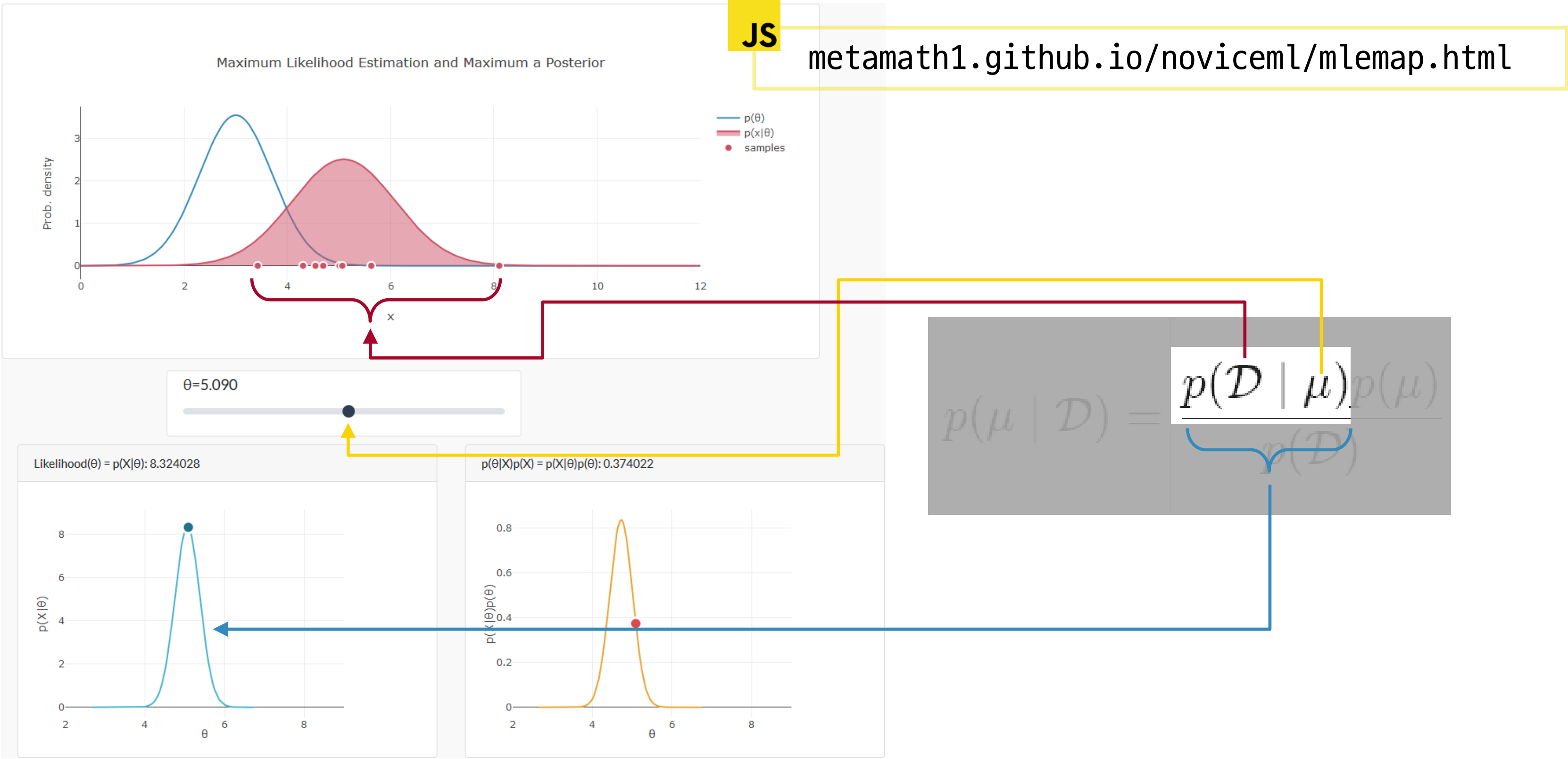
- $p(\mathcal{D}\mid\mu)$ 에서 μ 를 바꿔가면서 $p(\mathcal{D}\mid\mu)$의 값이 최대가 되게 조정하는 과정
- $\mathcal{D} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\mathrm{N})$

$$\frac{\partial}{\partial\mu} p(\mathcal{D}\mid\mu) \triangleq 0$$

$$\frac{\partial}{\partial\mu} p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\mathrm{N}\mid\mu) \triangleq 0 \xrightarrow{\text{독립 가정}} \frac{\partial}{\partial\mu} \overbrace{p(\mathbf{x}_1\mid\mu)p(\mathbf{x}_2\mid\mu)\cdots p(\mathbf{x}_\mathrm{N}\mid\mu)}^{\text{곱의 미분?}} \triangleq 0$$

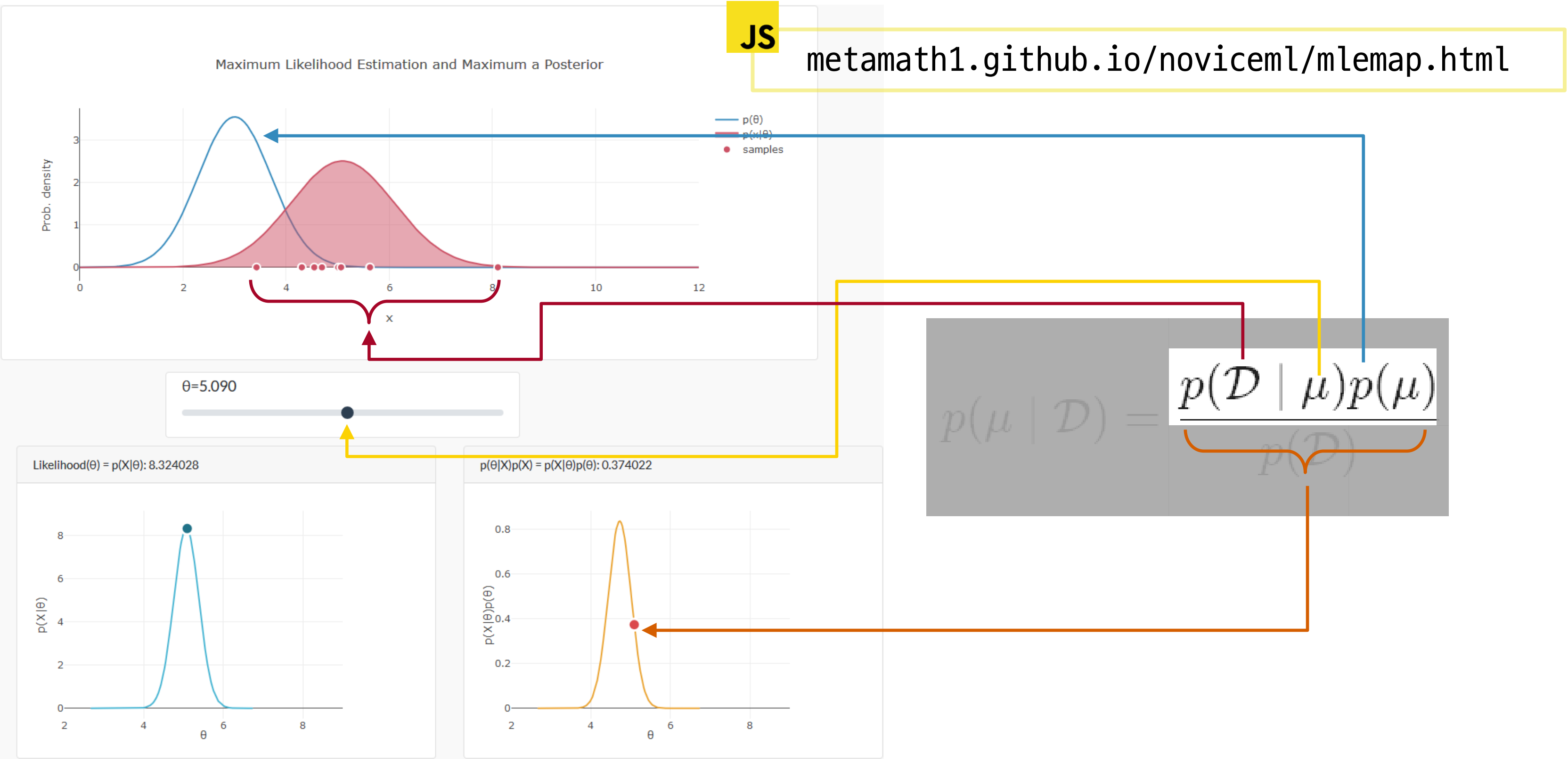
$$\xrightarrow{\log} \frac{\partial}{\partial\mu}\{\log p(\mathbf{x}_1\mid\mu) + \log p(\mathbf{x}_2\mid\mu) + \cdots + \log p(\mathbf{x}_\mathrm{N}\mid\mu)\} \triangleq 0$$

MLE 감잡기(Maximum Likelihood Estimation)



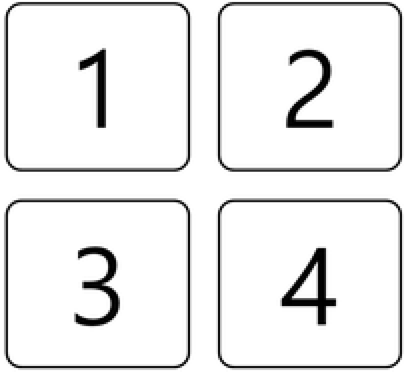
$$p(\mu \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mu)p(\mu)}{p(\mathcal{D})}$$

MAP 감잡기(Maximum A Posterior)



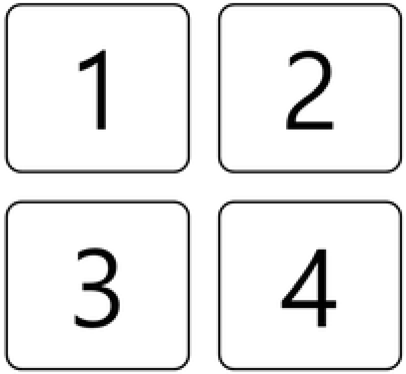
숫자출력 기계 for entropy

- 무작위로 숫자에 불이 들어오는 기계: 확률변수 X
- 1000번 관찰한 결과
  - 1: 250번, 2: 250번, 3: 250번, 4: 250번
- 1001번 째 어떤 번호에 불이 들어왔는지 알아내기 위해 필요한 질문?



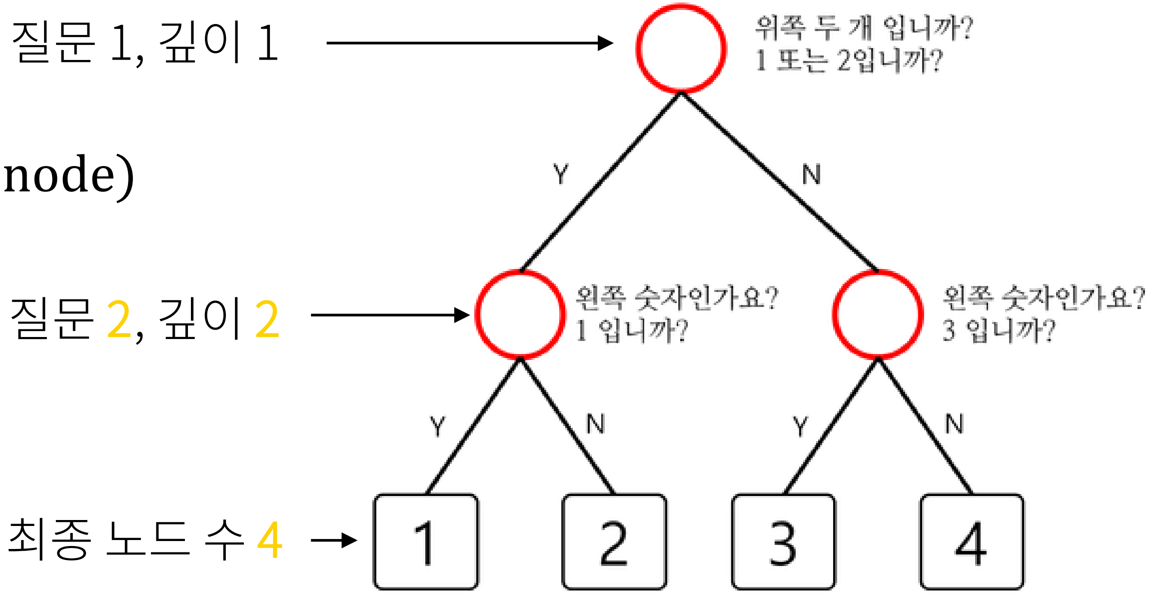
좋은 질문, 나쁜 질문

- 막 질문
  - 1입니까? 2입니까? 3입니까?
  - 만약 4가 나왔다면 질문을 3번이나 해야 된다.
- 좋은 질문
  - 첫번째: 윗쪽 번호 둘 중 하나인가요? 두번째: 왼쪽 번호인가요?
  - 어떤 경우라도 질문 두 번에 1001번째 숫자를 맞춘다.



질문수 표현

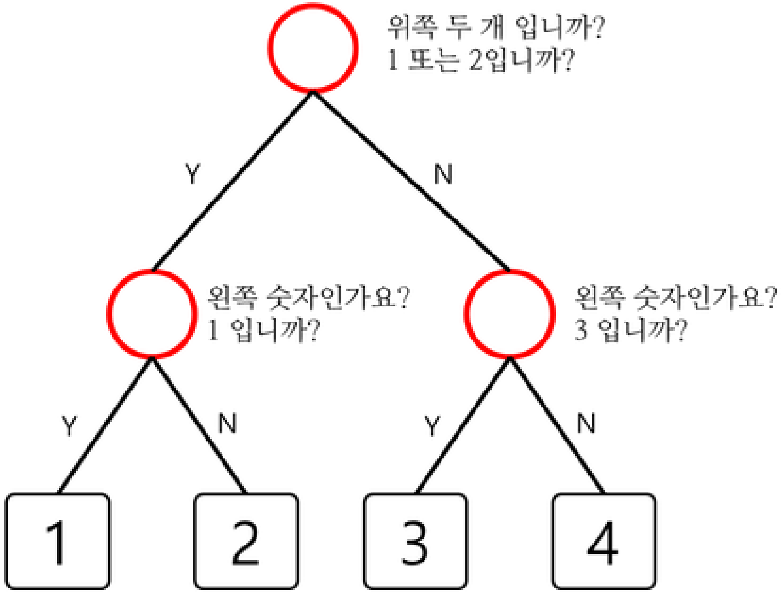
- 이진 트리 깊이: $\log_2(\#node)$



평균 질문수

- 좋은 질문
  - 어떤 경우라도 질문 두 번에 1001번째 숫자를 맞춘다.
  - 1을 맞추기 위한 질문 수: q_1
  - 2를 맞추기 위한 질문 수: q_2
  - $q_1 = 2 = \log_2 4$, $q_2 = 2 = \log_2 4$, $q_3 = 2 = \log_2 4$, $q_4 = 2 = \log_2 4$

$$\mathbb{E}(q) = \frac{250 \times q_1 + 250 \times q_2 + 250 \times q_3 + 250 \times q_4}{1000}$$
$$= \frac{250}{1000} q_1 + \frac{250}{1000} q_2 + \frac{250}{1000} q_3 + \frac{250}{1000} q_4$$



엔트로피

- 확률분포가 가지는 정보의 불확실성 또는 정보량

$$\mathbb{E}(q) = \frac{250 \times q_1 + 250 \times q_2 + 250 \times q_3 + 250 \times q_4}{1000}$$
$$= \frac{250}{1000}q_1 + \frac{250}{1000}q_2 + \frac{250}{1000}q_3 + \frac{250}{1000}q_4$$

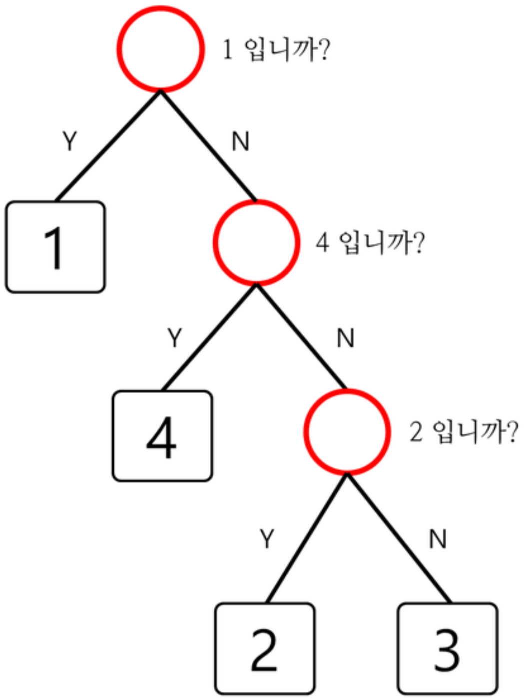
$$\mathbb{E}(q) = P(X=1)q_1 + P(X=2)q_2 + P(X=3)q_3 + P(X=4)q_4$$

$$\mathbb{E}(q) = P(X=1)\log_2 4 + P(X=2)\log_2 4 + P(X=3)\log_2 4 + P(X=4)\log_2 4$$
$$= \sum_{i=1}^{4} P(X=x_i)\log_2 4$$
$$= -\sum_{i=1}^{4} P(X=x_i)\log_2 \frac{1}{4} \quad \to P(X=x_i)$$

$$\boxed{H(X) = -\sum_{i=1}^{4} P(X=x_i)\log_2 P(X=x_i)}$$

엔트로피 비교

- 1000번 관찰한 결과
  - 1: 500번, 2: 125번, 3: 125번, 4: 250번
  - $Q(X=1)=0.5,\ Q(X=2)=0.125,$ $Q(X=3)=0.125,\ Q(X=4)=0.25$
- 질문수: $q_1=1=\log_2 2,\ q_2=3=\log_2 8,$ $q_3=3=\log_2 8,\ q_4=2=\log_2 4$



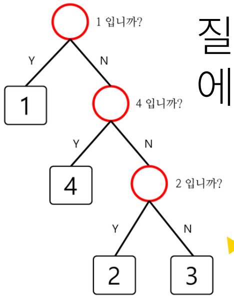
줄어든 엔트로피

$$
\begin{aligned}
\mathbb{E}(q) &= \frac{500 \times q_1 + 125 \times q_2 + 125 \times q_3 + 250 \times q_4}{1000} \\
&= \frac{500}{1000} q_1 + \frac{125}{1000} q_2 + \frac{125}{1000} q_3 + \frac{250}{1000} q_4 \\
&= \frac{500}{1000} \times 1 + \frac{125}{1000} \times 3 + \frac{125}{1000} \times 3 + \frac{250}{1000} \times 2 \\
&= Q(X=1)\log_2 2 + Q(X=2)\log_2 8 + Q(X=3)\log_2 8 + Q(X=4)\log_2 4 \\
&= \left(-Q(X=1)\log_2 \frac{1}{2}\right) + \left(-Q(X=2)\log_2 \frac{1}{8}\right) + \left(-Q(X=3)\log_2 \frac{1}{8}\right) + \left(-Q(X=4)\log_2 \frac{1}{4}\right) \\
&= -Q(X=1)\log_2 Q(X=1) - Q(X=2)\log_2 Q(X=2) - Q(X=3)\log_2 Q(X=3) - Q(X=4)\log_2 Q(X=4) \\
&= -\sum_{i=1}^{4} Q(X=x_i)\log_2 Q(X=x_i) = 1.75
\end{aligned}
$$

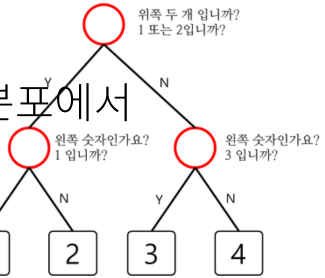
엔트로피 줄어듦, 예측이 쉬움, 뻔한 시스템, 얻을 정보가 별로 없음

크로스 엔트로피

$$
\begin{aligned}
\mathbb{E}(q) &= \frac{250 \times q_1 + 250 \times q_2 + 250 \times q_3 + 250 \times q_4}{1000} \\
&= \frac{250}{1000} q_1 + \frac{250}{1000} q_2 + \frac{250}{1000} q_3 + \frac{250}{1000} q_4 \\
&= \frac{250}{1000} \times 1 + \frac{250}{1000} \times 3 + \frac{250}{1000} \times 3 + \frac{250}{1000} \times 2 \\
&= P(X=1)\boxed{\log_2 2} + P(X=2)\boxed{\log_2 8} + P(X=3)\boxed{\log_2 8} + P(X=4)\boxed{\log_2 4} \\
&= \left(-P(X=1)\log_2 \frac{1}{2}\right) + \left(-P(X=2)\log_2 \frac{1}{8}\right) + \left(-P(X=3)\log_2 \frac{1}{8}\right) + \left(-P(X=4)\log_2 \frac{1}{4}\right) \\
&= -\boxed{P(X=1)}\log_2 Q(X=1) - \boxed{P(X=2)}\log_2 Q(X=2) - \boxed{P(X=3)}\log_2 Q(X=3) - \boxed{P(X=4)}\log_2 Q(X=4) \\
&= -\sum_{i=1}^{4} P(X=x_i)\log_2 Q(X=x_i) = 2.25
\end{aligned}
$$



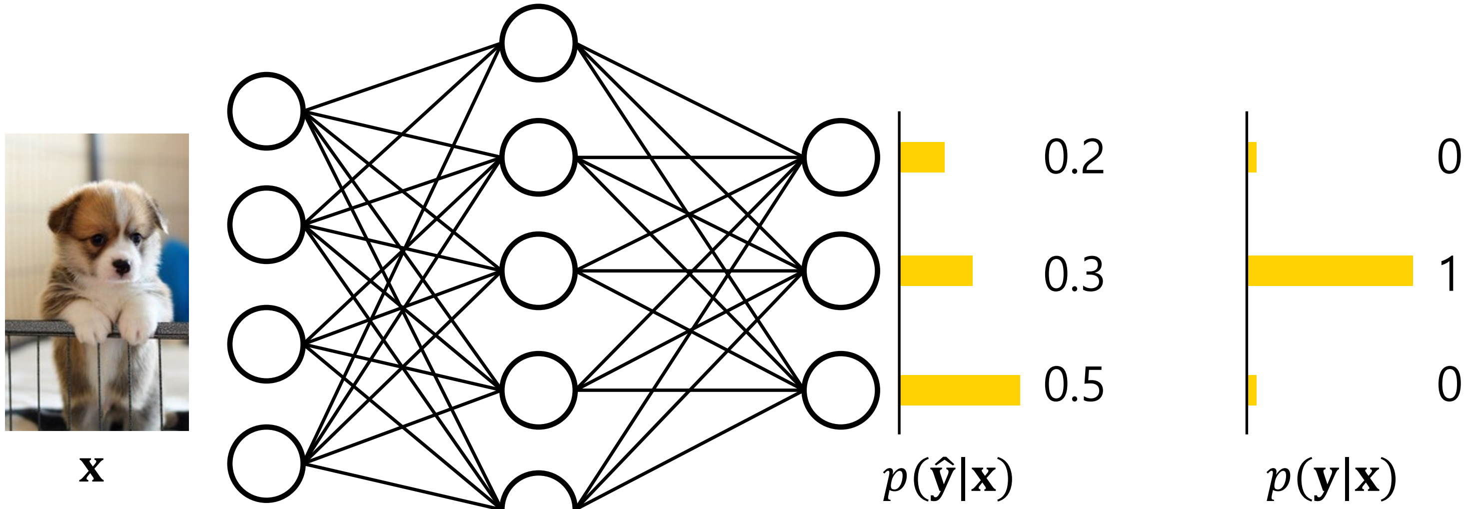
질문 방식 Q분포
에 알맞게 하고



평균은 P분포에서
계산하면

엔트로피 늘어남, $P(X) = Q(X)$일 때 최소

크로스 엔트로피 목적함수



$$\mathcal{L}(\mathbf{W}) = \sum_i p_i(\mathbf{y}|\mathbf{x}) \log p_i(\hat{\mathbf{y}}|\mathbf{x};\mathbf{W})$$