모두의연구소

수포자도 도전해 볼 만한

Mathematics in DeepLearning

Lecture5. Probability and Statistics

Juhee Lee Ph.D.
Research Professor @ ewha womans university

Probability Theory

Based on Set theory

- Set, universal set, element

Set Theory & Probability

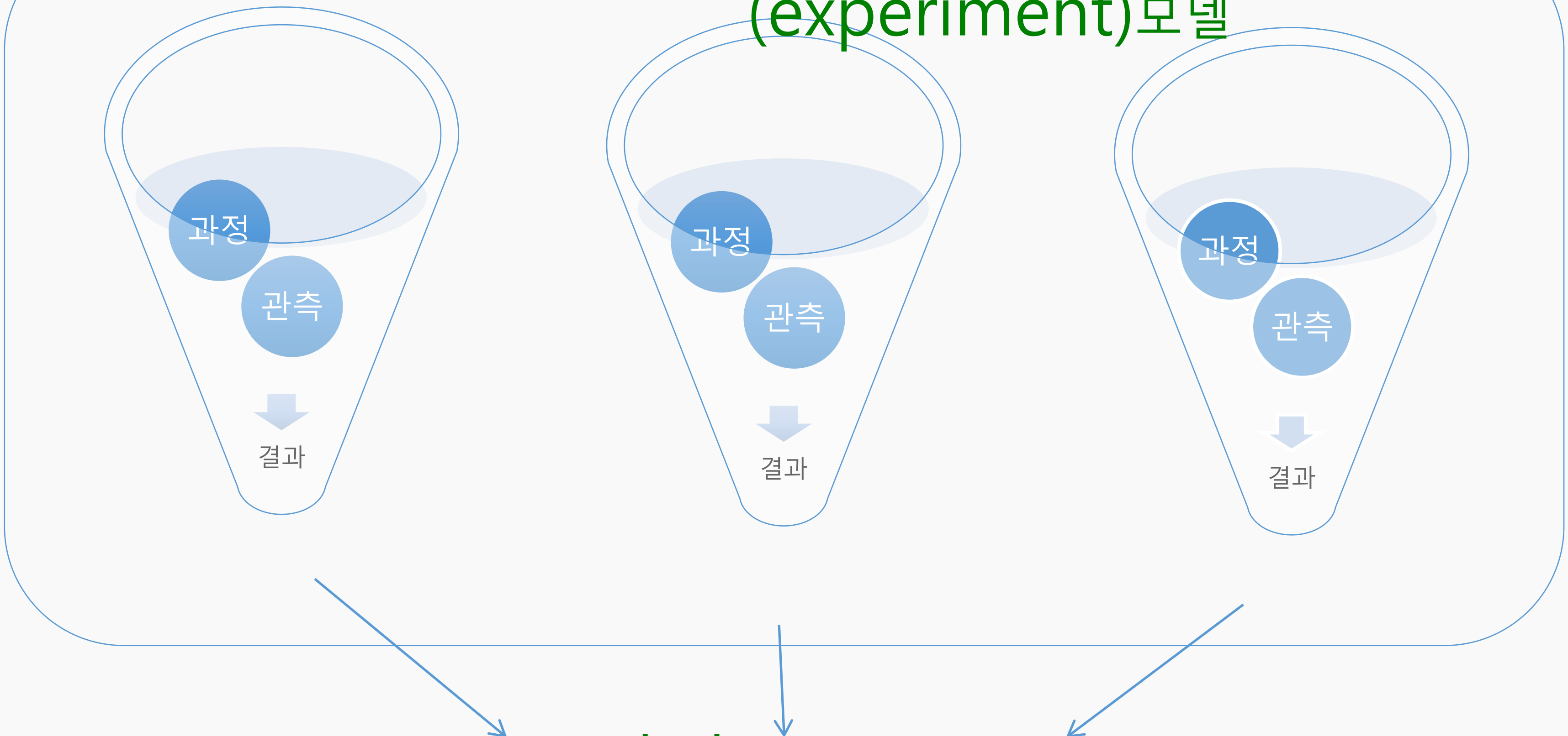
Based on Set theory

집합론	vs	확률론
집합		사건
전체집합		표본공간
원소		결과

- Set, universal set, element
- event, sample space, outcome
  - Event: set of outcomes
  - Outcome: being an observation
  - Experiment(Procedure, observation)

확률은

반복가능한 **확률실험**
(experiment)모델



사건 (event) :

Probability

Sample space : Ω

Vector : $\boldsymbol{a}$

Matrix : $\mathbf{A}$

Tensor : $\boldsymbol{\mathsf{A}}$

Probability

Definition **Sample space (or State space) $\boldsymbol{\Omega}$** is the collection of all possible outcomes under consideration.

Ex. 동전던지기(1개)

$$\Omega = \{H, T\}$$

Ex. 동전던지기(2개)

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$$

Probability

Definition An **Event** is a subset $\mathrm{A} \subset \Omega$ of the sample space

Ex. $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$

$\mathrm{A} = \{(H,T),(T,H)\} \subset \Omega$

Probability

Definition A set function $\mathbf{P}$ defined on the set of subsets of Ω is called a probability measure If it satisfies these 3 conditions for $\mathrm{A}, \mathrm{B} \subset \Omega$

(i) $\mathrm{P}(\mathrm{A}) \geq 0$

(ii) $\mathrm{P}(\emptyset) = 0, P(\Omega) = 1$

(iii) $\mathrm{A} \cap \mathrm{B} = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Review of Real analysis

<u>Definition</u> A collection $\mathcal{F}$ of subsets of a sample space Ω is called a **$\boldsymbol{\sigma}$-algebra**

($\mathcal{F} \subset \wp(\Omega)$)

If

$(i)\ \emptyset \in \mathcal{F}$

(ii) $\mathrm{A^c} = \Omega\backslash\mathrm{A} \in \mathcal{F}\ for\ any\ A \in \mathcal{F}$

$(iii)\ A_n \in \mathcal{F},\ n = 1,2,,\ldots \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

Ex. $\Omega = \{\mathrm{a},\mathrm{b}\}, \mathcal{F} = \{\emptyset,\{a\},\{b\},\{a,b\}\}$

Definition A pair $(\Omega, \mathcal{F})$ is called a **measurable spac
e** if Ω is a non-empty set and $\mathcal{F}$ is a σ-algebra on Ω

A subset $A \subset \Omega$ is said to be measurable if $A \in \mathcal{F}$

$$A_n \in \mathcal{F},\ n = 1,2,,\ldots \ \Rightarrow \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$$

$(\mathcal{F}_i)_{i \in I}$: a family of σ-algebra on Ω

$\bigcap_{i \in I} F_i = \{A \subset \Omega:\ A \in \mathcal{F}_i\ , \forall i \in I\}$ is a σ-algebra on Ω

For any family $\mathcal{O}$ of subsets of Ω

there is a smallest σ-algebra $\sigma(\mathcal{O})$

such that $\mathcal{O} \subset \sigma(\mathcal{O})$

$\mathcal{B}(R^n)$: Borel set in R^n

- the smallest σ-algebra generated by the open sets in R^n
- 측정가능한 집합

Probability space

Definition Let $(\Omega, \mathcal{F})$ be a measurable space

A mapping $\boldsymbol{\mu} : \boldsymbol{\mathcal{F}} \rightarrow \boldsymbol{R} \cup \{\infty\}$ is called a **measure**

If

(i) $\mu(\mathrm{A}) \geq 0$

(ii) $\mu(\emptyset) = 0, P(\Omega) = 1$

(iii) For any $\mathrm{A}_1, \cdots, A_n$ of mutually disjoint sets in $\mathcal{F}$,

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$$

$(\mu, \Omega, \mathcal{F})$: measure space

Random Variable

Definition Let $(\mu, \Omega, \mathcal{F})$: measure space.

A function $\boldsymbol{X} : \boldsymbol{\Omega} \rightarrow \boldsymbol{R^n}$ is said to be **measurable** if $X^{-1}(B) = \{w \in \Omega : \mathrm{X(w)} \in B\} \in \mathcal{F}$

for any $B \in \mathcal{B}(R^n)$

$(\Omega, \mathcal{F}, P)$: probability space

Definition A real measurable function $X : (\Omega, \mathcal{F}) \rightarrow (R^n, \mathcal{B}(R^n))$ is called a **random variable**

$(\Omega, \mathcal{F}, P)$: probability space

Definition Two events $\mathrm{A}, \mathrm{B} \in \mathcal{F}$ are said to be **independent** if

$$P(A \cap B) = P(A)P(B)$$

Ex. $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$, $H,T \rightarrow \frac{1}{2}$

$$A = \{(H,H),(H,T)\}, B = \{(H,T),(T,T)\}$$

$$A \cap B = \{(H,T)\}$$

$$P(A) = \frac{1}{2}, P(B) = \frac{1}{2} \quad P(A \cap B) = \frac{1}{4}$$

$(\Omega, \mathcal{F}, P)$: probability

<u>Definition</u> n events $\mathrm{A}_1, \ldots, \mathrm{A}_\mathrm{n} \in \mathcal{F}$ are said to be **not independent** if

$$P(A \cap A_2 \cap A_n) = P(A)P(A_2)P(A_n)$$

Ex. $\Omega =$ 두개의 주사위던질때

$$A = \text{첫번째 주사위 } \{1,2,3\}$$

$$B = \text{두번재 주사위 } \{3,4,5\}$$

$$\mathrm{C} = \text{두 결과의 합이 } 9$$

$$P(A \cap B \cap C) \neq P(A)P(B)P(C)$$

$(\Omega, \mathcal{F}, P)$: probability

<u>Definition</u> **Let** X and Y be two random variable on Ω

$$P(X \in A, Y \in \mathrm{B}) = P(X \in A)P(Y \in B)$$

$$P(X \in A, Y \in \mathrm{B}) = P(\{X \in A\} \cap \{\mathrm{Y} \in B\})$$
$$\forall A, B \in \mathcal{B}(R)$$

Ex. $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$

$$w = \{w_1, w_2\} \in \Omega, \mathrm{w_i} \in \{H,T\}, i = 1,2$$

$$X(w) = 1 \; if \; \mathrm{w_1} = H, \; 0 \text{ o.w}$$

$$Y(w) = 1 \; if \; \mathrm{w_2} = H, \; 0 \text{ o.w}$$

=> Independent

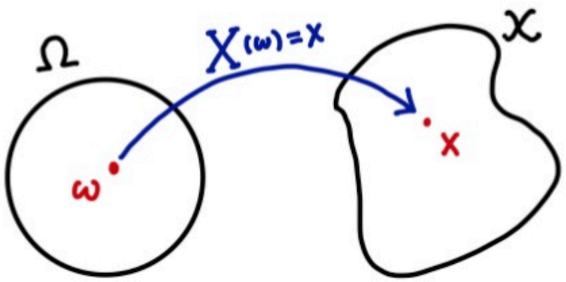
$$P(X=0, Y=1) = P(X^{-1}(1) \cap Y^{-1}(0)) = P(A \cap B) = P(A)P(B) = P(X=0)P(Y=1)$$

$$A = \{(H,H),(H,T)\}, \mathrm{B} = \{ (H,T),,(T,T)\}$$

Check $P(X=0, Y=0)P(X=1, Y=1)P(X=1, Y=0)$

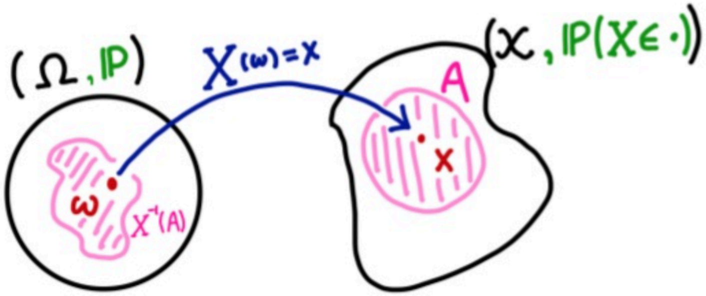
확률변수 X 는 ω 를 $\mathcal{X}$ 상의 원소 x 로 사상(mapping) 하는 함수입니다

$$X : \omega \mapsto x \in \mathcal{X}$$



확률분포는 X 가 어떻게 mapping 되는지 보여주는 $\mathcal{X}$ 상의 측도입니다

$$\mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A)$$



그리고 확률변수 X 는 H 와 T 를 각각 1 과 0 으로 보내는 걸로 정의합니다

$$X(\mathrm{H}) = 1, \quad X(\mathrm{T}) = 0$$

이 때 확률분포는 다음과 같이 $\mathcal{X} = \{0, 1\}$ 위에 똑같이 정의됩니다

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p$$

고로 확률측도는 사실 확률분포와 **전혀** 다르지 않습니다 😃

$$\mathbb{P}(\mathrm{H}) = \mathbb{P}(X = 1), \quad \mathbb{P}(\mathrm{T}) = \mathbb{P}(X = 0)$$

확률분포를 다른 용어로 X 의 **법**(law of X) 이라고도 부릅니다

단지 확률값의 측정을 어떤 공간에서 하는지, 그리고 확률변수가 묵시적으로 정의되어 있다는 점에서 차이가 있을 뿐입니다. 앞으로는 두 용어를 혼용해서 쓰겠습니다.

https://www.slideshare.net/ssuser7e10e4/wasserstein-gan-i

$(\Omega, \mathcal{F}, P)$: probability space

<u>Definition</u> The **conditional probability** $P\,(A|B)$ of A given B is defined by

$$P\,(A|B) = \frac{P(A \cap B)}{P(B)} \text{ for } \forall A \in \mathcal{F}$$

If A and B are independent

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

Distribution (law)

$(\Omega, \mathcal{F}, P)$: probability space

<u>Definition</u> The distribution
(law) $\mathrm{P_X}$ $of\ a\ random\ variable\ X: (\Omega, \mathcal{F}) \to (R^n,$
$\mathcal{B}(R^n))$ *is given by*

$$P_X(B) = P(X \in B) = (\{w \in \Omega: X(w) \in B\})$$

For each Borel set B

P_X is also probability measure on $(\Omega, \mathcal{F})$

The distribution $\mathrm{F_X}$ is defined by

$$F_X(x) = P_X((-\infty, x]) = P(X \leq x), x \in R$$

Ex. $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$

$$\mathrm{Z = X + Y} \in \{0,1,2\}, F_Z(x) = \begin{cases} 0. & , x < 0 \\ \frac{1}{4}, & 0 \leq x < 1 \\ \frac{3}{4}, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

$X_1, X_2, \ldots, X_n$: random variables

<u>Definition</u> The vector of random variables $X = (X_1, X_2, \ldots, X_n\ )$ is called a random vector.

The distribution function F_X is defined by $P(X_1 \leq x_1, \ldots, X_n \leq x_n)$ for $X = (X_1, X_2, \ldots, X_n\ ) \in R\text{^}n$

Theorem. The random variables $X_1, X_2, \ldots, X_n$ *are independent* $\Leftrightarrow F_{X_1}(x_1) \cdots F_{X_n}(x_n)$

Expectation

$(\Omega, \mathcal{F}, P)$: probability space

$\Omega = \{\mathrm{w}_1, \mathrm{w}_2, \ldots, w_n\}, P(w_i) = P_i\ , i = 1, \ldots, n$

$$P(\Omega) = \sum_{i=1} P_i = 1$$

<u>Definition</u> $X\colon \Omega \rightarrow \quad R$ a random vector,

The expectation of X is defined by

$E(X) = \sum_{i=1}^{n} P_i X(w_i) = \quad \sum_{i=1}^{n} P_i x_i \quad .$

$X(w_i) = x_i$ for $i = 1, \ldots, n$

Expectation

Ex. Two coin, X: the number of heads

E(X)=?

- Expectation is a linear functional

$$E(aX + bY) = aE(X) + bE(Y)$$

Expectation

We may assume that $\mathrm{x}_1, \cdots, x_n$ for random variable X *such that* $\mathrm{x}_1 < \cdots < x_n$

The distribution function F_X is the given by

$F_X(x_i) = P_1 + \cdots + P_{i-1}$ for $\mathrm{x}_{\mathrm{i}-1} < x < x_i$

with $F_X(x) = \begin{cases} 0, x < x_1 \\ 1, x \geq x_0 \end{cases}$

In particular, $P_i = F_X(x_i) - F_X(x_{i-1}), i = 2, \cdots n,$
$P_1 = F_X(x_1)$

$$E(X) = \sum_{i=1}^{n} P_i x_i = x_1 F_x(x_1) + x_2(F_x(x_2) - F(x_1)) + \cdots$$

Expectation

$$E(X) = \sum_{i=1}^{n} P_i x_i \quad = x_1 F_x(x_1) + x_2(F_x\,(x_2) - F(x_1))$$

$$+ \cdots + x_n(F_x(x_n) - F_x(x_{n-1}))$$

Take, $x_0 < x_1$ and let $F_X\colon [x_0, X_n] \rightarrow \quad R$ with

$f_X(x_i) = \quad \frac{F_x(x_n) - F_x(x_{n-1})}{x_i - x_{i-1}}$, $F_x(x_0) = 0$

Expectation

$$E(X)$$
$$= x_1 F_X(x_1) + \sum_{i=2}^{n} x_i \frac{F_x(x_n) - F_x(x_{n-1})}{x_i - x_{i-1}} (x_i - x_i)$$
$$= \int x f_x(x) dx$$

Density function

$(\Omega, \mathcal{F}, P)$: probability space , $random\ variable\ X$

<u>Definition</u> If the distribution function F_X is differentiable , then the expectation of X can be written by

$E(X) = \int_{-\infty}^{\infty} x f_x\,(x) dx$ where $f_x\,(x) = \frac{d}{dx} F_X(x)$

We call f_x the **density function** for X

$f_x\,(x)$: non-negative, $\int_{-\infty}^{\infty} x f_x\,(x) dx = 1$

$$F_X(x) = \int_{-\infty}^{x} x f_x\,(x) dx$$

Consider indep. Rv X *and* Y *with finitely many values* $x_1, \cdots, x_n$ *and* $y_1, \cdots, y_m$ *Then*

$$E(XY) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j \, P(X = x_i, Y = y_j) = E(X)E(Y)$$

Definition The variance $var(X)$ *of a random variable* X *is defined by*

$$var(X) = \mathrm{E}\big(\mathrm{X} - \mathrm{E(X)}\big)^2 = \mathrm{E(X^2)} - \mathrm{E(X)}^2$$

X : random variable

$$g(t) = \big(t - E(X)\big)^2$$

$$var(X) = \mathrm{E}\big(\mathrm{X} - \mathrm{E(X)}\big)^2 = E\big(g(x)\big) = \int_{-\infty}^{\infty} g(x) f_x\,(x) dx$$

Higher moments

Definition *a random variable X with density function f*

$$E(x^n) = \int_{-\infty}^{\infty} x^n f_x(x)dx$$

Is called the n-th moment of X, where n∈ N

$$\int_{-\infty}^{\infty} g(x)f_x(x)dx$$

$XE(X)$: $central\ moment\ of\ X, \sigma = \sqrt{Var(X)}$

$Y = \left(\mathrm{X} - \mathrm{E}(\mathrm{X})\right)/\sigma$'s moment : standard moment's

$E(Y) = 0\ E(Y^2) = 1,$

Definition *The covariance*

$Cov(X, Y)$ *of two random variable* $X, Y : \Omega \rightarrow \quad R$ *is defined by*

$$Cov(X, Y) = E((\mathrm{X} - \mathrm{E(X)})(\mathrm{Y} - \mathrm{E(Y)})) = \mathrm{E(XY)} - \mathrm{E(X)E(Y)}$$

X, Y: independent, $Cov(X, Y) = 0$

Statistics

- Probability theory based on Set theory

- A branch of mathematics dealing with the co llection, analysis, interpretation, presentation , and organization of data

Q & A