Bayesian Deep Neural Networks

Elementary mathematics

Sungjoon Choi

February 9, 2018

Seoul National University

Elementary of mathematics

Figure 1: Elementary of mathematics (copyright to wikipedia).

Elementary of mathematics

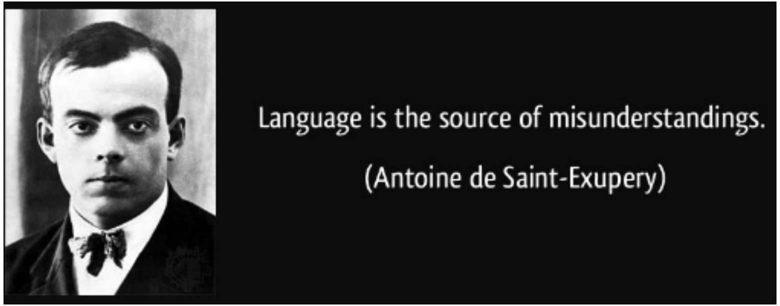


Table of contents

Introduction

Introduction

- Whats Wrong with Probability Notation? [1]
  - Whats Wrong?
    1. overloading $p(\cdot)$ for every probability function.
    2. using bound variables named after random variables.

       0, 1, 관측치 주사위눈 같은 것!

       행위를 할 때마다 random한 객체가 나와야 함. 그것을 RV 라고 함.

둘의 차이?

- Probability Notation is Bad

$$p(x|y) = p(y|x)p(x)/p(y)$$

$Y \sim N$ $X \sim B$

$P(X=0) \neq P(Y=0)$

다음

∴ 분포 종류를 명확히 해줘야 함.

- Random variables don't help.

$$P_{X|Y}(x|y) = P_{Y|X}(y|x)p_X(x)/p_Y(y)$$

- Great expectations

Goal: X vs x

P vs p

차이를 알면 된다.

$$\mathbb{E}[x] = \sum_x x p(x)$$

$$\mathbb{E}[X] = \sum_x x P_X(x)$$

[1] https://lingpipe-blog.com/2009/10/13/whats-wrong-with-probability-notation/

Introduction

- Today, I will introduce
  1. **probability theory** of Kolmogorov
     - set theory
     - measure theory.
  2. basic **functional analysis** 함수 해석학. (kernel의 이해를 위해 필요함)

완벽히 다룸?

- **Caution**
  - Try to get familiar with the terminologies.
  - Some facts could be counterintuitive.
  - No proof will be provided here.

kernel은 최소한 3가지 다른 관점으로 볼 수 있다.

확률론
함수해석학
...

+ Variational Inference에 함수해석학 개념 필요.

Introduction

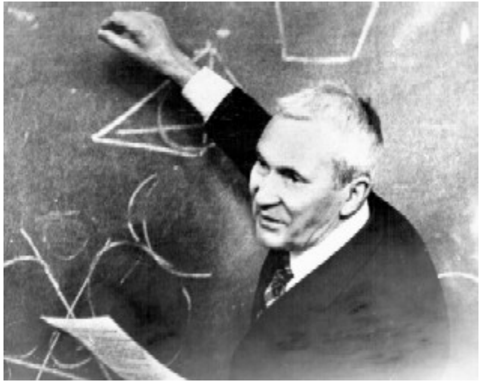


Figure 2: Andrey Kolmogorov

(확률론 제안)

using set & measure theory.

Introduction

- Import questions to have in mind throughout this lecture:
  1. What is probability?
  2. What is a random variable?
  3. What is a random process?
  4. What is a kernel function?

Don’t panic.

Most of the contents are from Prof. Taejeong Kim's slides.

Set theory

Set theory

옷장 옷 옷들의 일부 옷장안의 전체 옷들 ex) 옷장 안의 옷이 7개다, 3000의 개수, 아 옷이 압축되거나, 등 ...

- **set**, **element**, **subset**, **universal set**, **set operations**
- **disjoint** sets: $A \cap B = \emptyset$
- **partition** of A disjoint 하게 나누는 것
  example: $A = \{1, 2, 3, 4\}$, partition of A: $\{\{1, 2\}, \{3\}, \{4\}\}$
- **Cartesian product**: $A \times B = \{(a, b) : a \in A, b \in B\}$ ex) $\mathbb{R}^2$ space
  쌍들의 집합
  - example: $A = \{1, 2\}, B = \{3, 4, 5\}$
  - $A \times B = \{(1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5)\}$
- **power set** 2^A: the set of all the subsets of A.
  - example: $A = \{1, 2, 3\}$ 2^3
  - $2^A = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\}$

정수들의 개수 : infinite
실수들의 개수 : infinite

infinite =? infinite

- **cardinality** $|A|$: finite, infinite, countable, uncountable, denumerable (countably infinite)
  - $|A| = m, |B| = n \Rightarrow |A \times B| = mn$
  - $|A| = n \Rightarrow |2^A| = 2^n$
  - If there exists a one-to-one correspondence between two sets, they have the same cardinality. 두개의 무한대가 같으려면 one-to-one mapping을 보이면 됨. 즉, 다름을 보이기 위해 one-to-one이 아님을 증명하면, 두개의 무한대는 다름.
  - **countable**: There is a one-to-one between the set and a set of natural numbers. (example: set of all integers, set of all rational numbers)

자연수의 집합(뭔가 셀 수 있는 집합)과 1-1 mapping이 가능하면, 그 set을 'countable'이라 본다.

- Are **the set of all integers** and **the set of all rational numbers countable**?

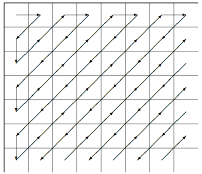
  trivial 분수도 countable.

- Yes. by the following mappings.

자연수

n	z
1	0
2	1
3	-1
4	2
5	-2
6	3
7	-3
$\vdots$	$\vdots$

m/n	1	2	3	4	5	$\cdots$
1	$1/1$	$1/2$	$1/3$	$1/4$	$1/5$	$\cdots$
-1	$-1/1$	$-1/2$	$-1/3$	$-1/4$	$-1/5$	$\cdots$
2	$2/1$	$2/2$	$2/3$	$2/4$	$2/5$	$\cdots$
-2	$-2/1$	$-2/2$	$-2/3$	$-2/4$	$-2/5$	$\cdots$
3	$3/1$	$3/2$	$3/3$	$3/4$	$3/5$	$\cdots$
-3	$-3/1$	$-3/2$	$-3/3$	$-3/4$	$-3/5$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	



직관은 ∞를 쪼갈지만
$\infty == 2 \times \infty$ 를 생각하면
같은 ∞임을 알 수 있다.
또 다른 관점으로 분모, 분자의 set을
정수인 set으로 볼수있고 분수는 원소의
두 정수의 product space 이므로
순서는 매길 수 있음. ∴ countable

- In fact, they are the same.

  cardinality가 같다.

- **denumerable**: "countably infinite" ex) 자연수 set, 정수 set, 분수 set : 한 개씩 순서를 매길 수 있음

  All denumerable sets are of the **same** cardinality, which is denoted by $\aleph_0$, aleph null or aleph naught.

- **uncountable**: not countable[2] 셀 수 없음.

  The smallest known uncountable set is $(0, 1)$ or $\mathbb{R}$, the set of all real numbers, whose cardinality is denoted by $\mathbf{c}$, "continuum."

  we call

  $\mathbf{c} = 2^{\aleph_0}$

[2]Found by Georg Cantor in 1874.

함수해석학 기초

Cantor

- Show that the cardinality of $C = [0, 1]$ is uncountable (Cantor's diagonal argument). 임을 보여라!
- **Proof sketch)**
  1. Suppose that C is countable.
  2. Then, there exists a sequence $S = \{x_1, x_2, \ldots\}$ such that all elements in C are covered. aleph null 개
  3. We can represent each x_i using a binary system. 0과 1 사이의 수는 이진법으로 표현가능

$$x_1 = 0.d_{11}d_{12}d_{13}\ldots$$
$$x_2 = 0.d_{21}d_{22}d_{23}\ldots$$
$$x_3 = 0.d_{31}d_{32}d_{33}\ldots$$

  where $d_{ij} \in \{0, 1\}$. 이렇게 모든 실수들의 집합을 cover할 수 있다면 countable 이라고 볼 수 있을 것 같다.

  4. Define $x_{new} = 0.\bar{d}_1\bar{d}_2\bar{d}_3\ldots$ such that $\bar{d}_i = 1 - d_{ii}$.
  5. Clearly, x_{new} does not appear in S, which is a contraction. So C must be uncountable.

반례로 증명

- Then what is the number of real numbers between 0 and 1?
- **Proof sketch)**
  1. We can represent a real number between 0 and 1 using a binary system.

$$r_1 = 0.d_{11}d_{12}d_{13}...$$
$$r_2 = 0.d_{21}d_{22}d_{23}...$$
$$r_3 = 0.d_{31}d_{32}d_{33}...$$

where $\boxed{d_{ij} \in \{0,1\}}$. 2^1

  2. To fully distinguish a real number r_i, we need $\aleph_0$ bits where $\aleph_0$ is the number of all integers.
  3. Consequently, $\mathbf{c} = 2^{\aleph_0}$ (uncountable). (continuum)

∴ 정수 set < [0, 1] 의 실수

Set theory

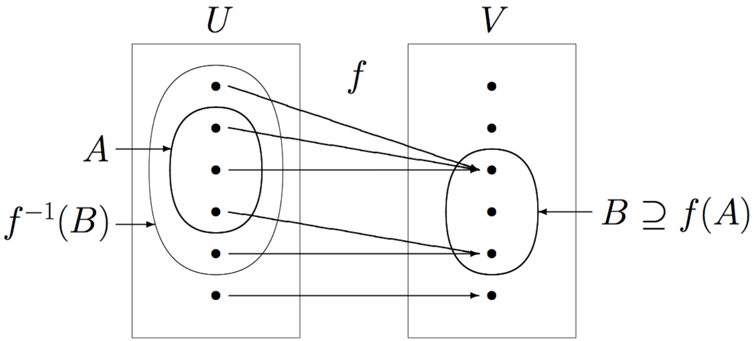
- **function** or **mapping** $f : \overset{\text{set}}{U} \to \overset{\text{set}}{V}$
- **domain** U, **codomain** V
- **image** $f(A) = \{f(x) \in V : x \in A\}$, $A \subseteq U$ codomain의 subset
- **range** $f(\underline{U})$ universal set에 들어갔을 때의 해당하는 codomain의 공간
- **inverse image** or **preimage** ML과 반대
  $f^{-1}(B) = \{x \in U : f(x) \in B\}$, $B \subseteq V$

Reality: label $\longrightarrow$ input.

ML: input $\longrightarrow$ $\underline{\underline{\text{label}}}$

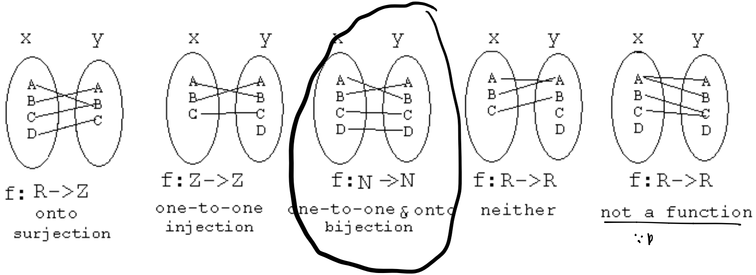
pre image

U는 당연히 f를 통과할 수 있다.



$f(U)$가, 즉 U가 모든 V를 cover 할 필요는 없다.

- **one-to-one** or **injective**: $f(a) = f(b) \Rightarrow a = b$
- **onto** or **surjective**: $f(U) = V$: codomain 전체 커버
- ★ **invertible**: one-to-one and onto (bijection)


x
y
A
B
C
D
A
B
C
f: R->Z
onto
surjection
x
y
A
B
C
A
B
C
D
f:Z->Z
one-to-one
injection
y
A
B
C
D
A
B
C
D
f:N ->N
one-to-one & onto
bijection
x
y
A
B
C
A
B
C
D
f:R->R
neither
x
y
A
B
C
D
A
B
C
D
f:R->R
not a function

Measure theory

Given a universal set U, a measure assigns a nonnegative real number to each subset of U.

Measure theory

- **set function**: a function assigning a number of a set (example: cardinality, length, area). set에 대해 어떤 숫자 하나를 매기는 것.
- σ-**field** $\mathcal{B}$: a collection of subsets of U such that (axioms)
  1. $\emptyset \in \mathcal{B}$ (empty set is included.)
  2. $B \in \mathcal{B} \Rightarrow B^c \in \mathcal{B}$ (closed under set complement.)
  3. $B_i \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} B_i \in \mathcal{B}$ (closed under countable union.)

σ: 면적을 주기위한 최소한의 '단위'로 이해하자.

가장 대표적인 σ-field: power set

σ-field 안에 들어가 있는 어떤 subset들은 우리가 measure를 부여할 수 있는 최소단위가 된다.

Measure theory

앞의 axiom을 통해 증명가능

- properties of σ-**field** $\mathcal{B}$
  1. $U \in \mathcal{B}$ (entire set is included.) $\phi^c = U \in \mathcal{B}$ $\because \phi^c \in \mathcal{B}$
  2. $B_i \in \mathcal{B} \Rightarrow \cap_{i=1}^{\infty} B_i \in \mathcal{B}$ (closed under countable intersection)
  3. 2^U is a σ-field. power set
  4. $\mathcal{B}$ is either finite or uncountable, never denumerable. (countable infinite 일 수 없어.)
  5. $\mathcal{B}$ and $\mathcal{C}$ are σ-fields $\Rightarrow \mathcal{B} \cap \mathcal{C}$ is a σ-field but $\mathcal{B} \cup \mathcal{C}$ is not.
     - $\mathcal{B} = \{\emptyset, \{a\}, \{b, c\}, \{a, b, c\}\}$
     - $\mathcal{C} = \{\emptyset, \{a, b\}, \{c\}, \{a, b, c\}\}$
     - $\mathcal{B} \cap \mathcal{C} = \{\emptyset, \{a, b, c\}\}$
       (this is a σ-field)
     - $\mathcal{B} \cup \mathcal{C} = \{\emptyset, \{a\}, \{c\}, \{a, b\}, \{b, c\}, \{a, b, c\}\}$
       (this is not a σ-field as $\{a, c\} = \{a\} \cap \{c\}$ is not included.)

ex) set의 cardinality가 10이면, powerset은 2^{10} $\therefore$ countable
integer set의 경우 $2^{\aleph_0}$
$2^{\aleph_0}$는 uncountable (continuum)

- $\sigma(\mathcal{C})$ is called the σ-field **generated** by $\mathcal{C}$.

이제, measure를 써야봐.
measure란? subset들을 어떤 숫자 하나로 (0보다 큰 양수) 보내는 함수 (즉 set function)

measure를 주기위해 σ-field 가 정의된다.

A σ-field is designed to define a measure.

kolmo의 가장 중요한 기여가 무엇이냐는?

바로 확률 P 가 measure 이다.

If the element is not inside a σ-field, it cannot be measured.

Measure theory

이 measure를 씌울 수 있으면 measurable space라 한다.

- A set U and a σ-field of subsets of U form a **measurable space** $(U, \mathcal{B})$.
- A **measure** μ defined on a measurable space $(U, \mathcal{B})$ is a set function $\mu : \mathcal{B} \to [0, \infty]$ such that
  1. $\mu(\emptyset) = 0$
  2. For disjoint B_i and $B_j \Rightarrow \mu(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mu(B_i)$ (countable additivity)
- Probability is a measure such that $\mu(U) = 1$, i.e., normalized measure. 모든 경우의 수를 다 더하면 1이다.
- A measurable space $(U, \mathcal{B})$ and a measure μ defined on it together form a measure space $(U, \mathcal{B}, \mu)$.

$\mathcal{B}$: σ-field

Probability

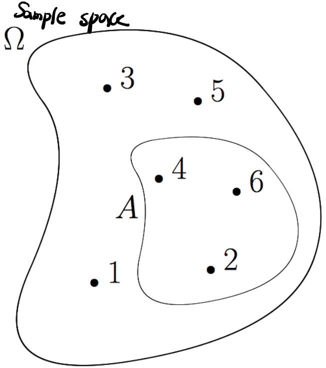
Probability

What is probability?

Probability

- Toss a fair dice and observe the outcomes.

모든 가능한 종류 (subset)에 대해 확률을 계산할 수 있어야 해서 power set의 개념이 쓰임.



대문자 P임 주의

확률은 sample space의 면적이다.

$\mu = P$

- $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = 1/6$
- $P(A) = P(2, 4, 6) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 1/2$

sample space에 대한 measure를 capital P로 표현함.

- The **random experiment** should be well defined.
- The **outcomes** are all the possible results of the random experiment each of which canot be further divided. 나올 수 있는 모든 값을 outcome이라고 한다
- The **sample point** w: a point representing an outcome. $\in \Omega$
- The **sample space** Ω: the set of all the sample points. S.S.에서 random하게 하나가 골라짐.

모든 sample point의 집합

sample space $\neq$ outcome

$\because$ S.S.에 없는 outcome이 나온다.

Ω의 원소들이 outcome이 X

Probability

- Definition (**probability**)
  - P defined on a measurable space $(\Omega, \mathcal{A})$ is **a set function** $P : \mathcal{A} \to [0, 1]$ such that (probability axioms).

measure와 같음.

1. $P(\emptyset) = 0$
2. $P(A) \geq 0 \; \forall A \subseteq \Omega$
3. For disjoint sets A_i and $A_j \Rightarrow P(\cup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i)$ (countable additivity)

조건 한개 추가

4. $P(\Omega) = 1$

Capital P 의미 중요

Probability

How do we assign **probability** to each event in A in such a way as to satisfy the axioms?

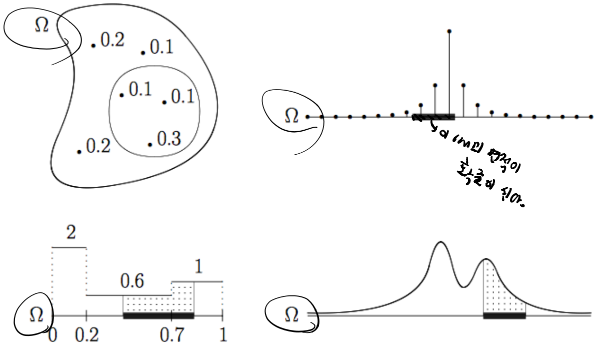
4가지 axiom을 만족하는 함수를 찾다보니 probability allocation function이 등장함.

- **probability allocation function** (PMF, PDF)
  - For discrete Ω:
    small
    $p : \Omega \to [0,1]$ such that $\sum_{w \in \Omega} p(w) = 1$ and $P(A) = \sum_{w \in A} p(w)$.
  - For continuous Ω:
    $f : \Omega \to [0, \infty)$ such that $\int_{w \in \Omega} f(w)dw = 1$ and
    $P(A) = \int_{w \in A} f(w)dw$.
- Recall that probability P is a set function $P : \mathcal{A} \to [0,1]$ where $\mathcal{A}$ is a σ-field.

P : set function

p : Ω 안의 원소가 들어가는 어떤 probability allocation function.

Examples of probability allocation functions:



이 (x의) 면적이 확률이 됨.

Ω에 대해 정의됨

Conditional probability

- **conditional probability** of A given B:

A는 set

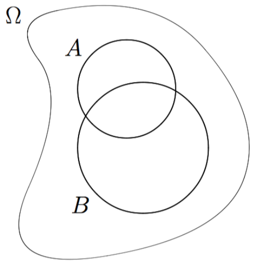
교집기호 ?

$$P(A|B) \underset{\text{def.}}{\triangleq} \frac{P(A \cap B)}{P(B)}.$$

large P 는 measure 이고,

A, B 는 set이고,

- Again, recall that **probability** P is a set function, i.e., $P : \mathcal{A} \to [0, 1]$.



Conditional probability

- From the definition of conditional probability, we can derive:

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

- **chain rule**:
  - $P(A \cap B) = P(A|B)P(B)$
  - $P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C)$
- **total probability law**:

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^C) \\ &= P(A|B)P(B) + P(A|B^C)P(B^C) \end{aligned}$$

Bayes' rule

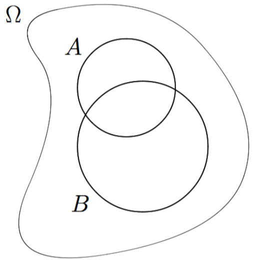
- **Bayes' rule**

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

- When B is the event that is considered and A is an observation,
  - $P(B|A)$ is called **posterior probability**.
  - $P(B)$ is called **prior probability**.

Independence

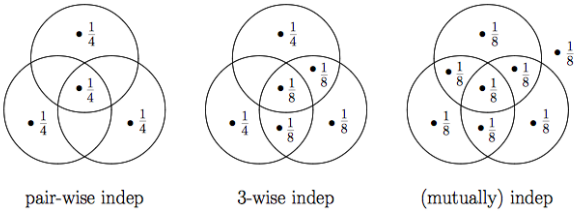
- **independent events** A and B: $P(A \cap B) = P(A)P(B)$



- independent $\neq$ disjoint, mutually exclusive

Independence

Example:



Summary : set이 있다 → σ-filed를 만든다 → σ-filed 하나하나마다 숫자를 매기는 measure를 줬다. → 이때 전체 set에 measure를 할때 1이 되는 normalized measure로 probility space를 정의를 했다. → 이 공간이 있다.

Random variable

우리가 np.randInt로 랜덤한 수 하나가 뽑힐 때,

이걸 어떻게 할수 있는 것일까? Random variable

- **random variable**:
  A random variable is a real-valued function defined on Ω that is measurable w.r.t. the probability space $(\Omega, \mathcal{A}, P)$ and the Borel measurable space $(\mathbb{R}, \mathcal{B})$, i.e.,

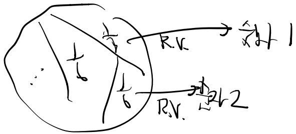
  → 우리가 가진 어떤 범위의 결합 집합을 만들어지는 σ-field를 Borel set이라고 함. $(\mathbb{R}, \mathcal{B})$

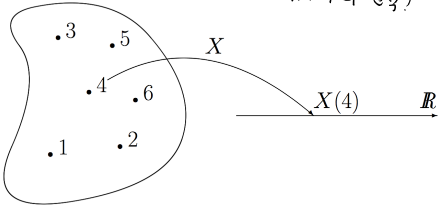
  우리의 관심

$$\underbrace{X : \Omega \to \mathbb{R}}_{\text{Random variable}} \text{ such that } \forall B \in \mathcal{B}, X^{-1}(B) \in \mathcal{A}.$$

숫자 '1'의 면적?

Ω가 공간에 (정수의) 숫자가 있는 것이 X





확률 : Ω의 σ-field에 정의된 set function

R.V. : Ω에 정의된 real line 혹은 discrete set으로 가는 함수

- What is random here?

  Ω에서 하나를 뽑는 것이 random이다.

- What is the result of carrying out the random experiment?

  반복시행하다가 주어지는 것.

Random variable

- Random variables are real numbers of our interest that are associated with the outcomes of a random experiment.
- $X(w)$ for a specific $w \in \Omega$ is called a **realization**. 즉 sampling
- The set of all realizations of X is called the **alphabet** of X.
- We are interested in $P(X \in B)$ for $B \in \mathcal{B}$:

X라는 r.v.를 통해 나올 수 있는 realization의 set은 alphabet이라고 부른다.

ex) Gaussian의 alphabet은 $\mathbb{R}$
주사위의 alphabet은 $1, \dots, 6$

정의 $$P(X \in B) \triangleq P(X^{-1}(B)) = P(\{w : X(w) \in B\})$$

우리는 sample space에서
사건이 일어나는지가 궁금하다!

Random variable

X라는 r.v가 있는데, 그것의 결과가
x_i가 나올 면적의 크기

- **discrete random variable**: There is a discrete set $\{x_i : i = 1, 2, \cdots\}$ such that $\sum P(X = x_i) = 1$.
- **probability mass function**: $p_X(x) \triangleq P(X = x)$ that satisfies
  1. $0 \leq p_X(x) \leq 1$
  2. $\sum_x p_X(x) = 1$
  3. $P(X \in B) = \sum_{x \in B} p_X(x)$

Random variable

- example: three fair-coin tosses
  - X = number of heads
  - probability mass function (pmf)

$$p_X(x) = \begin{cases} 1/8, & x = 0 \\ 3/8, & x = 1 \\ 3/8, & x = 2 \\ 1/8, & x = 3 \\ 0, & \text{else} \end{cases}$$

  - $P(X \geq 1) = \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = \frac{7}{8}$

해석: X가 Bernoulli distribution을 따를 때 k가 나올 확률

- *Bernoulli* $p_X(k) = \begin{cases} 1-p, & k=0 \\ p, & k=1 \\ 0, & \text{else} \end{cases}$

- *uniform* $p_X(k) = \begin{cases} 1/(m-l+1), & k=l, l+1, l+2, \cdots, m \\ 0, & \text{else} \end{cases}$

- *geometric* $p_X(k) = \begin{cases} (1-p)p^k, & k=0,1,2,\cdots \\ 0, & \text{else} \end{cases}$

Random variable

- **continuous random variable**
  There is an integrable function $f_X(x)$ such that
  $P(X \in B) = \int_B f_X(x)dx$.
- **probability density function**
  $f_X(x) \triangleq \lim_{\Delta x \to 0} \frac{P(x<X\leq x+\Delta x)}{\Delta x}$ that satisfies
  1. $f_X(x) > 1$ is possible.
  2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  3. $P(X \in B) = \int_{x \in B} f_X(x)dx$

Random variable

- **uniform** $f_X(k) = \begin{cases} 1/(b-a), & a \leq x \leq b \\ 0, & \text{else} \end{cases}$
- **exponential** $f_X(k) = \begin{cases} \lambda e^{\lambda x}, & x \geq 0 \\ 0, & \text{else} \end{cases}$
- **Laplace** $f_X(k) = \frac{\lambda}{2} e^{\lambda |x|}$
- **Gaussian** $f_X(k) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$
- **Cauchy** $f_X(k) = \frac{\lambda}{\pi(\lambda^2+x^2)}$

Expectation

definition

$$EX \triangleq \begin{cases} \sum_x x p_X(x), & \text{discrete} X \\ \int_{\infty}^{\infty} x f_X(x) dx, & \text{continuous} X \end{cases}$$

Conditional expectation

일반적인 expectation이라는 다르다.

Cexp.는 given에 따라 바뀌는 function이다.

∴ r.v.이다.

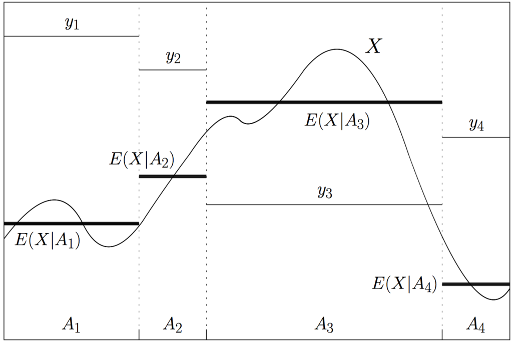
- Conditional expectation $E(X|Y)$
  - Expectation $E(X)$ of random variable X is $EX = \int x f_X(x)dx$ and is a deterministic variable.
  - $E(X|Y)$ is a function of Y and hence a random variable.
  - For each y, $E(X|Y)$ is X average over the event where $Y = y$.

$E(X|Y)$에서 X와 Y의 σ-field가 다르기 때문에, measure 함부로 정의할 수 없음.

∴ 몇 가지 조건이 필요함.

Conditional expectation

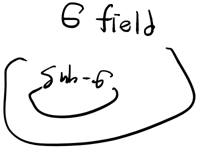
- Conditional expectation $E(X|Y)$



Assume that the probability is uniformly allocated over Ω.

X : continuous

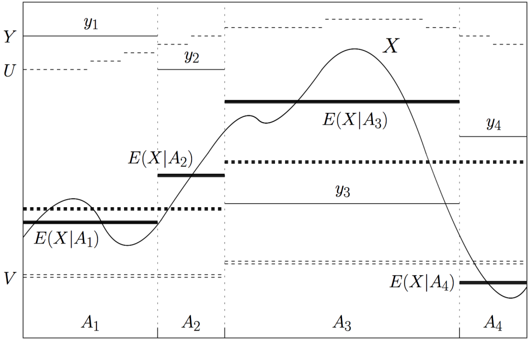
X|Y : 4



- Definition **(conditional expectation)**
  - Given a random variable Y with $\mathbb{E}|Y| < \infty$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and some sub-σ-field $\mathcal{G} \subset \mathcal{A}$ we will define the **conditional expectation** as the almost surely unique random variable $\mathbb{E}(Y|\mathcal{G})$ which satisfies the following two conditions
    1. $(Y|\mathcal{G})$ is $\mathcal{G}$-measurable.
    2. $\mathbb{E}(YZ) = \mathbb{E}(\mathbb{E}(Y|\mathcal{G}Z)$ for all Z which are bounded and $\mathcal{G}$-measurable.

Conditional expectation

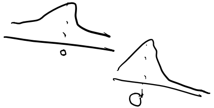
- Conditional expectation $E(X|Y)$ with different σ-fields.



Assume that the probability is uniformly allocated over Ω.

Moment

mean만으로는 분포를 판단하기 힘듭니다.



모든 moment가 같아야지 두 분포가 같다고 한다.

GNN (GIN)에서도 꽤 많이 등장한다

- n-th **moment** EX^n
- **mean** $m_X = EX$
- **variance** $\sigma_X^2 = var(X) = E(X - m_X)^2$
- **skewness** $\frac{E(X-m_X)^3}{\sigma_X^3}$
- **kurtosis** $\frac{E(X-m_X)^4}{\sigma_X^4}$

Joint moment

- **correlation** EXY
- **covariance** $cov(X, Y) = E(X - m_X)(Y - m_Y)$
- **correlation coefficient** $\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$
- **uncorrelated** $EXY = EXEY$
  - independent $\Rightarrow$ uncorrelated
  - uncorrelated $\not\Rightarrow$ independent
- **orthogonal** $EXY = 0$

Sampling.

- Gaussian

- Multivariate Gaussian

- GAN : 128차원 정도.

- 무한차원은?

Random process

무한 차원의 r.v.을 정의하고 싶을 때 쓰는게

Random process이다.

Random process

보통 함수는 무한차원의 vector라고 얘기한다.

- We would like to extend random vectors to infinite dimensions. That is, we would like to mathematically describe an infinite number of random variables simultaneously, e.g., infinite trials of tossing a die.

5차원 벡터 : element 5개

그럼 무한차원은?

무한차원은 내가 몇 번째 element를 가지고 있는지가 변수로 표현된다.

즉, input이 real line인 함수가 된다.

수학적으로 함수와 무한차원 벡터는 같다.

∴ Random process는 함수들의 공간에 확률을 부여한 것이라고 볼 수 있다

Random process

I : Index set

- **random process** $X_t(w)$, $t \in I$: 즉 t에 따라 달라지는 r.v.

  time

  1. random sequence, random function, or random signal:

     $$X_t : \Omega \to \text{the set of all sequences or functions}$$ 함수들의 공간으로 가는 mapping

  2. indexed family of infinite number of random variables:

     $$X_t : I \to \text{set of all random variables defined on } \Omega$$

  3. $X_t : \Omega \times I \to \mathbb{R}$ → 이것이 random process
  4. If t is fixed, then a random process becomes a random variable.

Randomness는 어디서 오는가?

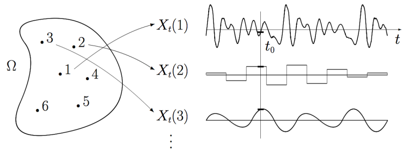
Ω에서 온다.

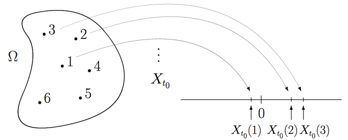
I는 index set이므로 Randomness가 없다.

즉, 우리의 시간(t 혹은 I)이 고정되었다면

random process는 random variable이 된다.

반대로 I가 열려있으면 함수를 뽑는 것임.





Random process

- A random process X_t is completely characterized if the following is known.
    - $P((X_{t_1}, \cdots, X_{t_k}) \in B)$ for any B, k, and $t_1, \cdots, t_k$

: 가변적으로 길이가 변화하는 것에 대해 확률을 정의 (분포)할 수 있으면, 랜덤프로세스라 할 수 있다.
여기서 길이는 t의 위치이며, 길이가 ∞일 때도 가능.

- Note that given a random process, only 'finite-dimensional' probabilities or probability functions can be specified.

Random process

- For a fixed $t \in \mathcal{T}$, $X_t(w)$ is a random variable. t가 fix이면 r.v이다.
- For a fixed $w \in \Omega$, $X_t(w)$ is a deterministic function of t, which is called a **sample path**.

random process에서의 sampling은 sample path.

- types of random processes
  1. discrete-time
  2. continuous-time

  } Index

  3. discrete-valued
  4. continuous-valued

  } output

4가지의 combination으로 나뉘진다.

r.V.는 종류 2가지 (discrete, continuous)

r.p.는 Index set I 까지 있으므로 종류가 더 많음.

- Example: Brownian motion

Random process

- **Moment** 더이상 하나의 value를 보는 것이 아니고 함수를 나타내기 때문에 mean도 function이 된다
  - mean function

$$m_X(t) \triangleq EX_t = \begin{cases} \sum_x x p_{X_t}(x), & \text{discrete-valued} \\ \int x f_{X_t}(x)dx, & \text{continuous-valued} \end{cases}$$

  - auto-correlation function, acf

≈ kernel in Gaussian process

$$R_X(t, s) \triangleq EX_tX_s$$

입력

  - auto-covariance function, acvf

$$C_X(t, s) \triangleq E(X_t - m_X(t))(X_s - m_X(s))$$

  - cross-covariance function, acvf

$$R_{XY}(t, s) \triangleq E(X_t - m_X(t))(Y_s - m_Y(s))$$

Gaussian process에서

ML에서 나오는 kernel과 acf는 완벽히 동일하다.

Gaussian의 분포는 zero mean이라고 했을 때 acf 만 알면 모든 것을 정의할 수 있는 특징이 있다.

- **Stationarity** Random process (그냥 한마디로 부드러운 random process)
  - **(strict-sense) stationary, sss**

τ만큼 shift를 시켜. $$\underbrace{P((X_{t_1+\tau}, \cdots, X_{t_k+\tau}) \in B)}_{\text{shift invariance.}} = \underline{P((X_{t_1}, \cdots, X_{t_k}) \in B)}$$

- If X_t is strict-sense stationary,
  - mean function $m_X(t+\tau) = m_X(t)$: 즉 mean이 constant이다. ∴ 0으로 봐도 상관이 없다.
  - Correlation $R_X(t+\tau, s+\tau) = R_X(t,s)$ ∵ stationary 자체가 shift invariant 하므로!
  - $C_X(t+\tau, s+\tau) = C_X(t,s)$
- If X_t is wide-sense stationary,
  - $m_X(t+\tau) = m_X(t)$
  - $R_X(t+\tau, s+\tau) = R_X(t,s)$

Random process

두 출력값 사이의 관계가 두 입력 사이의 차이에만 dependent하다

- If X_t is wide-sense stationary,
  - $m_X(t) = m_X$
  - $R_X(t, s) = R_X(t - s) = R_X(\tau)$
  - $C_X(t, s) = C_X(t - s) = C_X(\tau)$

시간의 차이만 고려 의 random process를 말하는 것. → stationary random process

- In (general) Gaussian processes, wss is assumed.
- $R_X(t, s)$ corresponds to a kernel function, i.e., $k(t, s)$.

이러한 r.p.를 특별히 stationary random process 이다.

kernel function : 두 입력 값이 가까우면 큰 값 → 즉, 변수들의 correlation 이라고 볼 수 있다.
〃 〃 멀면 작은 값

두 시간 사이의 거리가 가까우면 correlation ↑ → 두 r.p.의 출력값이 비슷하다.

이를 수리적으로 표현하는 게 Gaussian process의 auto correlation 혹은 kernel function.

Functional analysis

- 더하기, 곱셈이 정의됨
- 순서의 필요 X,

ex) 함수: $\sin(\cdot) \times 2 = 2\sin(\cdot)$

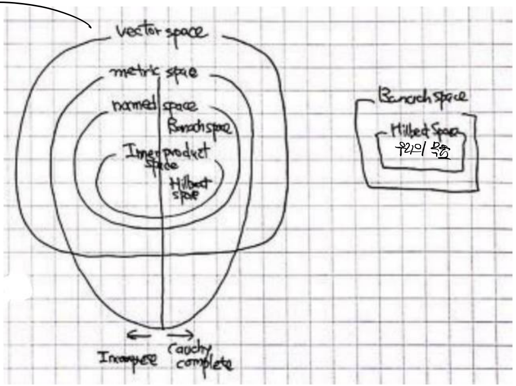
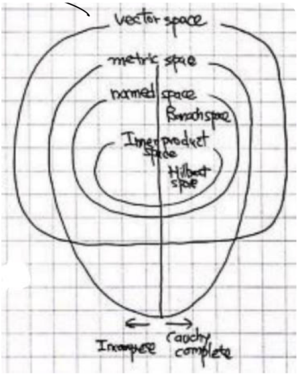


Figure 3: Mathematical spaces (copyright to Kyungmin Noh).



- Vector space: space with algebraic structures (addition, scalar multiplication, ...)
- Metric space: space with a metric (distance)

  두 개가 들어가 사이의 거리를 잰다!

  ① 0보다 크다
  ② 같을 때 0이다
  ③ 삼각부등식 (triangle inequality)
- Normed space: space with a norm (size)
- Inner-product space: space with an inner-product (similarity)

  Normed-space에 포함됨.
- Hilbert space: complete space

  Inner-product 중에서 complete space

  - 우리가 아는 수학적 연산 대부분 가능

  (유리수로 표현하지 못 하는 값이 있다면 incomplete space)

  위의 모든 구조 + Complete

Functional analysis

- We will show a bunch to terminologies and theorems.
  1. Inner product
  2. Hilbert space
  3. Kernel 수학에선 3가지, but 연관성X
  4. Positive definite
  5. Eigenfunction and eigenvalue
  6. Mercer's theorem
  7. Bochner's theorem
  8. Reproducing kernel Hilbert space (RKHS)
  9. Moore-Aronszajn theorem
  10. Representer theorem

- Definition (**inner product**)
  - Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if

만족해야 함.

    1. Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 f_1 \cdot g + \alpha_2 f_2 \cdot g$
    2. Symmetric $\langle f, g \rangle_{\mathcal{H}} = \langle g, h \rangle_{\mathcal{H}}$
    3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Hibert space $\subset$ Vector space

$\therefore$ Vector space의 property에 따라 이것들이 만족된다.

  - Note that norm can be naturally defined from the inner product:

$$\|f\|_{\mathcal{H}} \triangleq \sqrt{\langle f, f \rangle_{\mathcal{H}}}$$

Don’t panic.

- Definition (**Hilbert space**) 즉 Complete 하며,
  - Inner product space containing Cauchy sequence limits.

    = 공간이 내 set에 포함됨.

    $\Rightarrow$ Complete space

    $\Rightarrow$ Always possible to *fill all the holes*.

    $\Rightarrow$ $\mathbb{R}$ is complete, $\mathbb{Q}$ is not complete.

    유리?

- Definition (**Kernel**)
  - Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,
$$k(x, x') \triangleq \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

조건이 까다롭지 않음.

- Note that there is almost no condition on $\mathcal{X}$.

ex)
- Sum of kernels or product of kernels are also a kernel.
- Kernels can be defined in terms of sequences in $\phi \in l_2$, i.e., $\sum_{i=1}^{\infty} \phi_i^2(x) \leq \infty$.

  뒤에서 계속.
- Theorem
  - Given a sequence of functions $\{\phi_i(x)\}_{i\geq 1}$ in l_2, where $\phi_i : \mathcal{X} \to \mathbb{R}$ is the i-th coordinate of $\phi(x)$. Then

    coordinate의 값.

    매우 높은 차원

    $$k(x, x') \triangleq \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x').$$

    각각을 곱해서 내적을 한 것이 커널이다.
- This is often used as an intuitive interpretation of a kernel function.

kernel SVM

∴ 일반적으로 나눌 수 없는 것을 더 큰 차원 'ϕ'로 보내서 나눠버림. separable

→ 실제로 커널을 통해 training dataset을 linear line으로 모두 나눌 수 있다. ∵ 무한차원으로 확장가능

but, overfitting 이라는 맹점이 있으므로 그렇게 하지 않는다.

Functional analysis

function₁ → function₂

유한차원의 공간이 필요. ∵ function은 무한차원의 벡터니까.

A 함수를 eigen function으로 쓸 수 있고 eigen function들의 coefficient로 분리할 수 있다고 가정해보자. ⇒ 이것을 mapping 할 수 있다.

※.

우리가 $\vec{V_1}$ 에서 $\vec{V_2}$ 로 projection을 A로 쓰는 연산은 'inner product'

why? 그 방향의 coefficient를 구하고 싶기 때문!

우리는 이제 function₁ 과 function₂를 가지고 inner product를 할 것이고, 이를 위해 Hilbert space가 필요하다.

- Let T_k be an operator defined as

$$(T_k f)(x) = \int_{\mathcal{X}} k(x,x')f(x')d\mu(x')$$

함수

k라는 function 이 정의 되면 $T_k f$는 function을 새로운 function으로 만들어주는 mapping

where $\mu(\cdot)$ denotes a measure $(d\mu(x') \to dx')$.

- T_k can be viewed as a mapping between spaces of functions:

$$T_k : L_2(\mathcal{X},\mu) \to L_2(\mathcal{X},\mu).$$

함수의 domain에 대해 제곱하여 적분한 값은 유한한 공간 → 즉 무한대는 보고싶지 않아!

$T_k f$: func. → func.

- Once a kernel $k(\cdot,\cdot)$ is defined, the mapping T_k is defined accordingly.

① function을 basis로 표현

(이를 위해 coefficient를 계산해야 함.)

(이를 위해 함수 사이의 내적을 정의해야 함.)

(또한 무한차원으로 만들어진 함수가 원래 함수로 수렴하는 것을 보이기 위해, Complete 해야 함.)

∴ 내적 정의 + complete ⇔ Hilbert

Functional analysis

- Definition (**positive definite**)
  - A kernel is said to be positive definite if

$$\int k(x,x')f(x)f(x')d\mu(x)d\mu(x') \geq 0$$

for all $f \in L_2(x,\mu)$.

Functional analysis

- Definition (**Eigenfunction and eigenvalue**)
  - Given a kernel function $k(\cdot, \cdot)$ and

$$\int k(x, x')\phi(x)d\mu(x) = \lambda\phi(x').$$

Then, $\phi(x)$ and λ are eigenfunction and eigenvalue of a kernel $k(\cdot, \cdot)$.

eigenfunction eigenvalue

- Theorem (**Mercer**)
  - Let $(\mathcal{X}, \mu)$ be a finite measurable space and $k \in L_\infty(\mathcal{X}^2, \mu^2)$ be a kernel such that $T_k : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ is positive definite.
  - Let $\phi_i \in L_2(\mathcal{X}, \mu)$ be the normalized eigenfunctions of T_k associated with the eigenvalues $\lambda_i > 0$. Then:

    kernel function 에 대한 eigen function

    1. The eigenvalues $\{\lambda_i\}_{i=1}^{\infty}|$ are absolutely summable.
    2. $$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x')$$
       holds μ^2 almost everywhere, where the series converges absolutely and uniformly μ^2 almost everywhere.

결국 k : f → f

k는 eigen function 과 coefficient로 분해할 수 있다.

- Absolutely summable is more important than it seems. 유한개로 나눠서 무한합하는 것은 되고 길구 ㅇㅇ/a.
- SB: Mercer's theorem can be interpreted as an infinite dimensional SVD.

무한차원의 SVD로 볼 수 있다.

Functional analysis

- Theorem (Kernels are positive definite)
  - Let $\mathcal{H}$ be a Hilbert space, $\mathcal{X}$ be a non-empty set, and $\phi : \mathcal{X} \to \mathcal{H}$. Then $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ is positive definite.
- Reverse also holds:
  Positive definite $k(x, x')$ is an inner-product in $\mathcal{H}$ between $\phi(x)$ and $\phi(x')$.

Functional analysis

- Theorem (**Bochner**)
  - Let f be a bounded continuous function on $\mathbb{R}^d$. Then f is positive semidefinite iff. it is the (inverse) Fourier transform of a nonnegative and finite Borel measure *mu*, i.e.,

$$f(x) = \int_{\mathbb{R}^d} e^{iw^T x} \mu(dw).$$

X라는 공간에서 만족하는 성질이 주파수대역 공간에서도 만족한다는 것.

- What does this mean?

- Corollary (**Bochner**) P.S를 증명할 때 쓰이는 theorem
  - If we have an isotropic kernel function function, i.e.,

$$k(x, x') = k_I(t = |x - x'|),$$

어떤 kernel function이 두 입력 사이의 거리(차)로 표현이 되고, 이것이 positive semidefinite (PS)인 경우 (보이는 것이)

showing the non-negativeness of a Fourier series of $k_I(t)$ is equivalent to showing the positive definiteness of $k(x, x')$.

Fourier series의 값이 non-negative 임을 보이면 된다.

  - Example:

$$k(x, x') = \underset{\text{양수}}{\underline{\underline{\cos}}} \left(\frac{\pi}{2} |x - x'| \right).$$

Gaussian process에서 나오는 kernel과 여기는 다르지만, 또한 positive definite의 조건은 만족시켜야 한다.

RKHS

- Definition (**reproducing kernel Hilbert space**)
  - Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel on $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space if

reproducing kernel Hilbert space

1. $\forall x \in \mathcal{X}$

   $k(x, \cdot) \Leftrightarrow$ $k(\cdot, x) \in \mathcal{H}$ 하나의 입력을 고정시켜도 in H

   ∵ symmetric
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$

   $\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (reproducing property)

   얘만 okay. 결국 eigen decomposition 된다.
3. $\forall x, x' \in \mathcal{X}$

   $k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}}$

- What does this indicates?

- Suppose we have a RKHS $\mathcal{H}$, $f(\cdot) \in \mathcal{H}$, and $k(\cdot, x) \in \mathcal{H}$.

2번 조건

- Then the reproducing property indicates that evaluation of $f(\cdot)$ at x, i.e., $f(x)$ is the inner-product of $k(\cdot, x)$ and $f(\cdot)$ itself, i.e.,

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

- Recall Mercer's theorem $k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x')$. Then,

$$
\begin{aligned}
f(x) &= \left\langle f, \sum_{i=1}^{\infty} \lambda_i \phi_i(\cdot) \phi_i(x) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{\infty} \lambda_i \langle f, \phi_i(\cdot) \rangle_{\mathcal{H}} \phi_i(x) \\
&= \sum_{i=1}^{\infty} \bar{\lambda}_i \phi_i(x)
\end{aligned}
$$

① 함수로 나눌 수 있다

② coefficient를 함수로 나눈 내적을 통해 정의할 수 있다.

③ basis function 과 그것의 coefficient로 나눠서 표현할 수 있음

결국 커널이 무한 차원의 eigen vector와 value로 decomposition이 가능하다.

where $\bar{\lambda}_i = \lambda_i \langle f, \phi_i(\cdot) \rangle_{\mathcal{H}}$.

- Theorem (**Moore-Aronszajn**)
  - Let $\mathcal{X}$ be a non-empty set. Then, for every positive-definite function $k(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS and vice versa.
- This indicates:
  reproducing kernels $\Leftrightarrow$ positive definite function $\Leftrightarrow$ RKHS

kernel function 하나를 정의하면 RKHS를 하나 정의하는 것과 같다.

즉 공간은 하나 고정된 것.

내가 smooth or stiff한 kernel을 정의하면
raw data가 있는 공간이 smoother, more stiff 하게 변다.

- Definition (**another view of RKHS**)
  - Consider the space of function $\mathcal{H}$ defined as

입력이 들어오니까, 각 data 근접시킴.

$$\mathcal{H} = \{f(x) = \sum_{i=1}^{n} \alpha_i \underline{k(x, x_i)} : n \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}\}.$$

입력된거 basis로 kernel로 잡음

  - Let $g(x) = \sum_{j=1}^{n'} \alpha'_j k(x, x'_j)$, then we define the inner-product

이 공간에 아직 내적은 정의하지 않았으므로, 남은 같이 just define 함.

$$\langle f, h \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \alpha'_j k(x_i, x'_j)$$

재현되게 이 property 만족!

  - We can easily demonstrate the reproducing property:

$$\langle k(\cdot, x), f(\cdot) \rangle_{\mathcal{H}} = \langle k(\cdot, x), \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \rangle_{\mathcal{H}}$$

$$= \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

$$= f(x).$$

결국 함수 f가 어떤 basis function으로 span된다.

Functional analysis

- Theorem (**Representer**)
  - Let $\mathcal{X}$ be a nonempty set and $k(\cdot, \cdot)$ be a positive definite kernel with corresponding RKHS $\mathcal{H}_k$. Given training samples $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$, a strictly monotonically increasing real-valued function $g : [0, \infty) \to \mathbb{R}$, and an arbitrary empirical risk function $E : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$, then for any $f^* \in \mathcal{H}_k$ satisfying

$$f^* = \underset{f \in \mathcal{H}_k}{\arg\min} \{E(\mathcal{D}) + g(\underbrace{\|f\|}_{\text{norm}})\}$$

f^* admits a representation of the form:

우리가 fitting 하고 싶은 함수

$$f^*(\cdot) = \sum_{i=1}^{\overset{\text{train data 개수}}{n}} \alpha_i k(\cdot, \underbrace{x_i}_{\text{train data}})$$

k의 input 고정시킨 것으로 span 되는데, at most n개로 span 된다

f가 이렇게 표현된다는 theorem

where $\alpha_i \in \mathbb{R}$.

즉 우리가 찾고 싶은 f가 n개의 coefficient를 찾는 문제로 바뀜.

Functional analysis

- Example
  1. Given $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, solve

$$\min_{f \in \mathcal{H}_k} \frac{1}{2} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \gamma \|f\|_{\mathcal{H}}^2. \tag{1}$$

  2. From the representer theorem, solving (1) becomes:

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_j k(x_i, x_j) - y_i \right)^2 + \gamma \|f\|_{\mathcal{H}}^2. \tag{2}$$

  3. Represent (2) with a matrix form:

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \|K_{XX}\alpha - Y\|_2^2 + \gamma \alpha^T K_{XX} \alpha. \tag{3}$$

  4. $\nabla_\alpha (3) = K_{XX}(K_{XX}\alpha - Y) + \gamma K_{XX}\alpha = 0$
  5. Finally, $\alpha = (K_{XX} + \gamma I)^{-1} Y$ where $f(x) = \sum_{i=1}^n \alpha_i(x, x_i)$.
- Note that the form of this solution is identical to the mean function of Gaussian process regression.

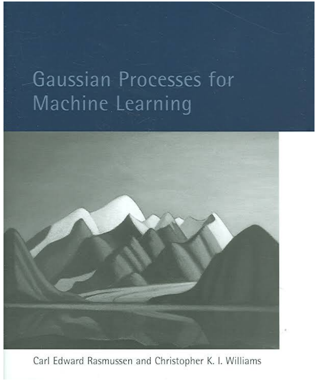
나중에

Gaussian process에서

또 보자

Questions?

Text book


Gaussian Processes for
Machine Learning
Carl Edward Rasmussen and Christopher K. I. Williams

References i