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

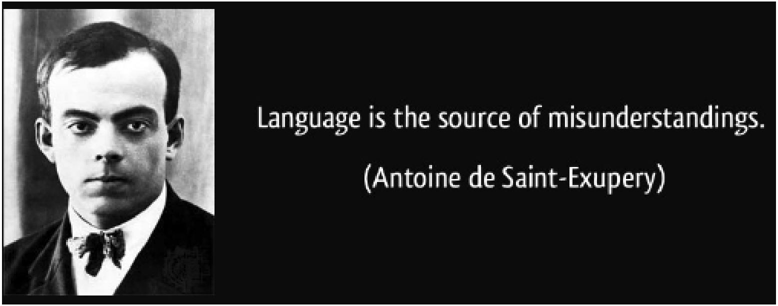


Table of contents

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.
  - Probability Notation is Bad

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

  - Random variables don't help.

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

  - Great expectations

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

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

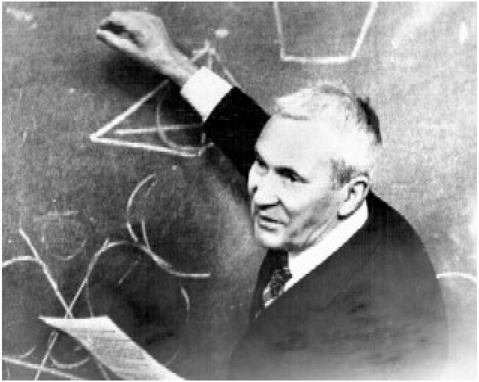


Figure 2: Andrey Kolmogorov

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

- **set**, **element**, **subset**, **universal set**, **set operations**
- **disjoint** sets: $A \cap B = \emptyset$
- **partition** of A

  example: $A = \{1, 2, 3, 4\}$, partition of A: $\{\{1, 2\}, \{3\}, \{4\}\}$
- **Cartesian product**: $A \times B = \{(a, b) : a \in A, b \in B\}$
  - 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: $A = \{1, 2, 3\}$
  - $2^A = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\}$

Set theory

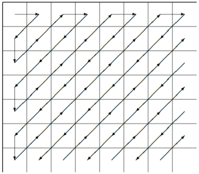
- **cardinality** $|A|$: finite, infinite, countable, uncountable, denumerable (countably infinite)
  - $|A| = m, |B| = n \Rightarrow |A \times B| = mn$
  - $|A| = n \Rightarrow |2^A| = 2^n$
  - If there exists a one-to-one correspondence between two sets, they have the same cardinality.
  - **countable**: 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)

Set theory

- Are **the set of all integers** and **the set of all rational numbers countable**?
- Yes. by the following mappings.

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



- In fact, they are the same.

Set theory

- **denumerable**: countably infinite
  All denumerable sets are of the **same** cardinality, which is denoted by $\aleph_0$, aleph null or aleph naught.
- **uncountable**: not countable[2]
  The smallest known uncountable set is $(0, 1)$ or $\mathbb{R}$, the set of all real numbers, whose cardinality is denoted by **c**, continuum.
  $\mathbf{c} = 2^{\aleph_0}$

[2] Found by Georg Cantor in 1874.

Set theory

- Show that the cardinality of $C = [0, 1]$ is uncountable (Cantor's diagonal argument).
- **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.
  3. We can represent each x_i using a binary system.

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

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

- 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.

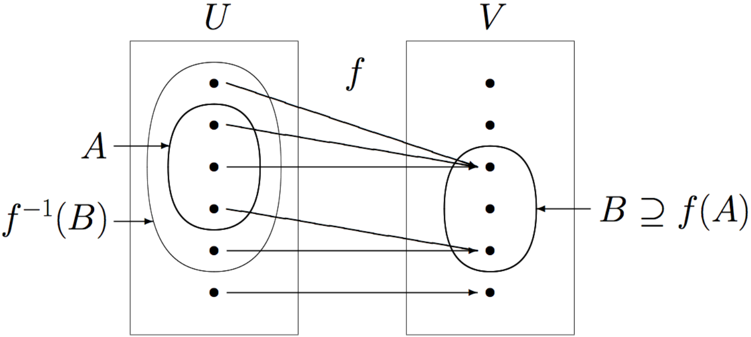
$$r_1 = 0.d_{11}d_{12}d_{13}...$$
$$r_2 = 0.d_{21}d_{22}d_{23}...$$
$$r_3 = 0.d_{31}d_{32}d_{33}...$$

     where $d_{ij} \in \{0, 1\}$.
  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).

Set theory

- **function** or **mapping** $f : U \to V$
- **domain** U, **codomain** V
- **image** $f(A) = \{f(x) \in V : x \in A\}$, $A \subseteq U$
- **range** $f(U)$
- **inverse image** or **preimage**
  $f^{-1}(B) = \{x \in U : f(x) \in B\}$, $B \subseteq V$

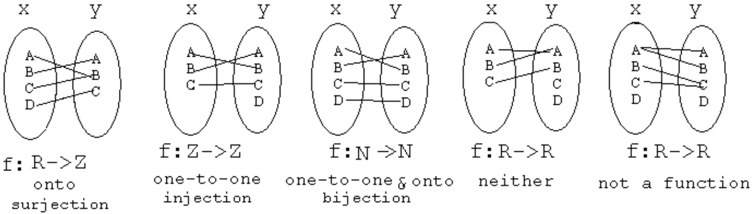
Set theory



Set theory

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

Set theory


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
x
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

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).
- 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.)

- properties of σ-**field** $\mathcal{B}$
  1. $U \in \mathcal{B}$ (entire set is included.)
  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.
  4. $\mathcal{B}$ is either finite or uncountable, never denumerable.
  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.)
- $\sigma(\mathcal{C})$ is called the σ-field **generated** by $\mathcal{C}$.

A σ-field is designed to define a measure.

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

Measure theory

- 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.
- A measurable space $(U, \mathcal{B})$ and a measure μ defined on it together form a measure space $(U, \mathcal{B}, \mu)$.

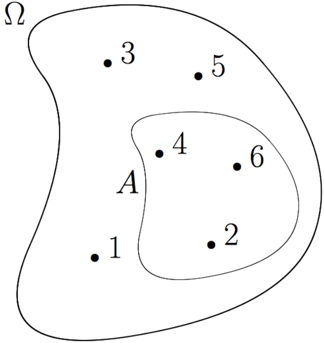
Probability

Probability

What is probability?

Probability

- Toss a fair dice and observe the outcomes.



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

Probability

- 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.
- The **sample point** w: a point representing an outcome.
- The **sample space** Ω: the set of all the sample points.

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).
    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)
    4. $P(\Omega) = 1$

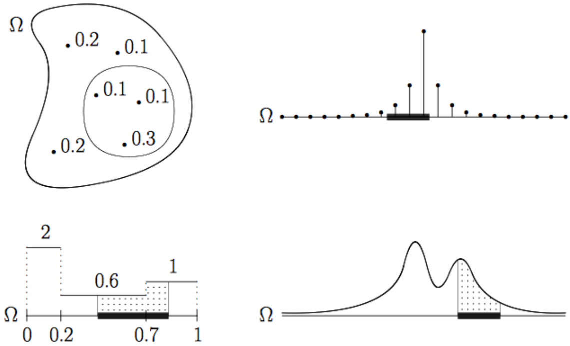
Probability

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

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

Probability

Examples of probability allocation functions:

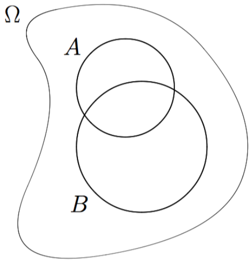


Conditional probability

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

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

- 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

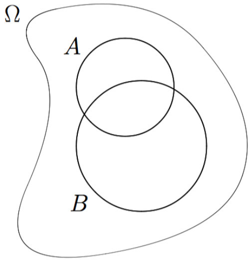
- **Bayes' rule**

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

- 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

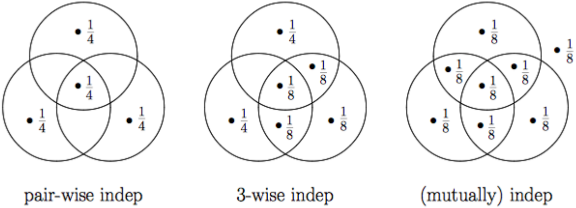
- **independent events** A and B: $P(A \cap B) = P(A)P(B)$



- independent $\neq$ disjoint, mutually exclusive

Independence

Example:

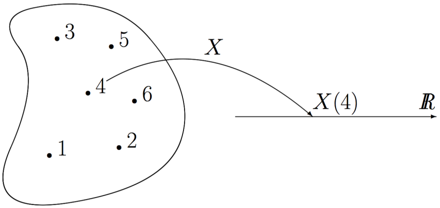


Random variable

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.,

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



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

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**.
- The set of all realizations of X is called the **alphabet** of X.
- We are interested in $P(X \in B)$ for $B \in \mathcal{B}$:

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

Random variable

- **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)

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

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

Random variable

- *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

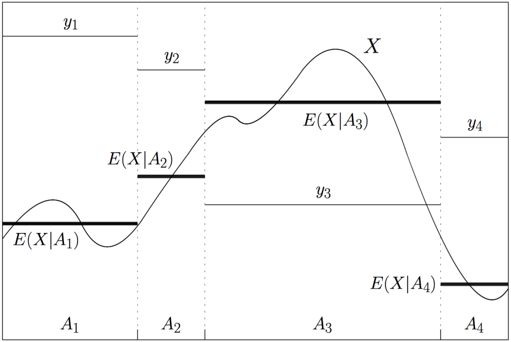
$$EX \triangleq \begin{cases} \sum_x x p_X(x), & \text{discrete} X \\ \int_\infty^\infty x f_X(x) dx, & \text{continuous} X \end{cases}$$

Conditional expectation

- 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 $Y = y$.

Conditional expectation

- Conditional expectation $E(X|Y)$



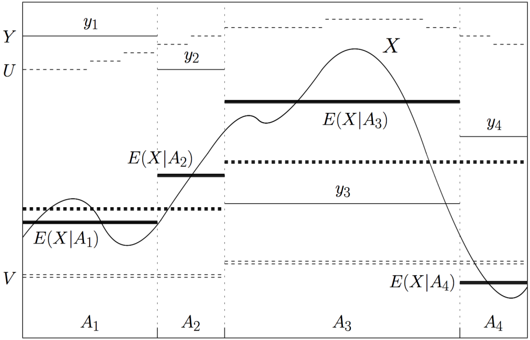
Assume that the probability is uniformly allocated over Ω.

Conditional expectation

- 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

- 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** $\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$
- **uncorrelated** $EXY = EXEY$
  - independent $\Rightarrow$ uncorrelated
  - uncorrelated $\not\Rightarrow$ independent
- **orthogonal** $EXY = 0$

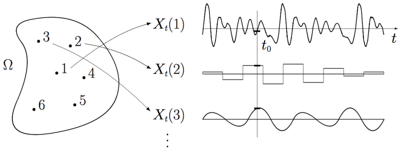
Random process

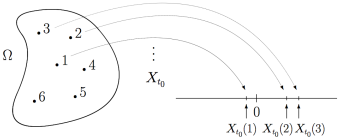
Random process

- 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.

Random process

- **random process** $X_t(w), t \in I$:
  1. random sequence, random function, or random signal:
     $X_t : \Omega \to$ the set of all sequences or functions
  2. indexed family of infinite number of random variables:
     $X_t : I \to$ set of all random variables defined on Ω
  3. $X_t : \Omega \times I \to \mathbb{R}$
  4. If t is fixed, then a random process becomes a random variable.





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 $t_1, \cdots, t_k$
- 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.
- For a fixed $w \in \Omega$, $X_t(w)$ is a deterministic function of t, which is called a **sample path**.
- types of random processes
  1. discrete-time
  2. continuous-time
  3. discrete-valued
  4. continuous-valued
- Example: Brownian motion

Random process

- **Moment**
  - **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**

$$R_X(t,s) \triangleq EX_t X_s$$

  - **auto-covariance function, acvf**

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

  - **cross-covariance function, acvf**

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

Random process

- **Stationarity**
  - **(strict-sense) stationary, sss**

$$P((X_{t_1+\tau}, \cdots, X_{t_k+\tau}) \in B) = P((X_{t_1}, \cdots, X_{t_k}) \in B)$$

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

- If X_t is wide-sense stationary,
    - $m_X(t) = m_X$
    - $R_X(t, s) = R_X(t - s) = R_X(\tau)$
    - $C_X(t, s) = C_X(t - s) = C_X(\tau)$
- In (general) Gaussian processes, wss is assumed.
- $R_X(t, s)$ corresponds to a kernel function, i.e., $k(t, s)$.

Functional analysis

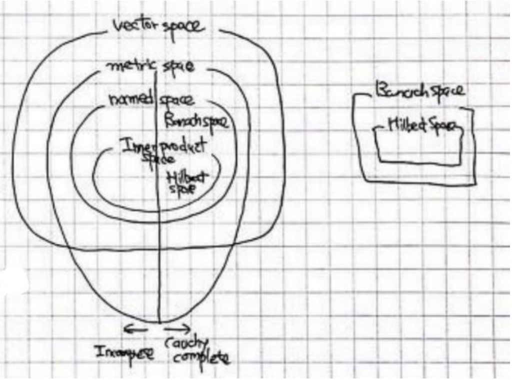


Figure 3: Mathematical spaces (copyright to Kyungmin Noh).

Functional analysis

- Vector space: space with algebraic structures (addition, scalar multiplication, ...)
- Metric space: space with a metric (distance)
- Normed space: space with a norm (size)
- Inner-product space: space with an inner-product (similarity)
- Hilbert space: complete space

Functional analysis

- We will show a bunch to terminologies and theorems.
  1. Inner product
  2. Hilbert space
  3. Kernel
  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
    1. Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}}$
    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 $f = 0$.
  - 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**)
  - Inner product space containing Cauchy sequence limits.
    $\Rightarrow$ Complete space
    $\Rightarrow$ Always possible to *fill all the holes*.
    $\Rightarrow$ $\mathbb{R}$ is complete, $\mathbb{Q}$ is not complete.

- 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 $\forall x, x' \in \mathcal{X}$,
  $$k(x, x') \triangleq \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$
- Note that there is almost no condition on $\mathcal{X}$.

Functional analysis

- 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., $\sum_{i=1}^{\infty} \phi_i^2(x) \leq \infty$.
- 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

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

- This is often used as an intuitive interpretation of a kernel function.

Functional analysis

- Let T_k be an operator defined as

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

  where $\mu(\cdot)$ denotes a measure $(d\mu(x') \rightarrow dx')$.
- T_k can be viewed as a mapping between spaces of functions:

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

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

- 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 $k(\cdot, \cdot)$.

Functional analysis

- 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:
    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.
- Absolutely summable is more important than it seems.
- SB: Mercer's theorem can be interpreted as an infinite dimensional SVD.

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.,

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

- What does this mean?

- Corollary (**Bochner**)
  - If we have an isotropic kernel function function, i.e.,

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

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

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

Functional analysis

- 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
    1. $\forall x \in \mathcal{X}$
       $k(\cdot, x) \in \mathcal{H}$
    2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$
       $\langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (reproducing property)
    3. $\forall x, x' \in \mathcal{X}$
       $k(x, x') = \langle k(\cdot, x), k(\cdot, x')\rangle_{\mathcal{H}}$
- What does this indicates?

Functional analysis

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

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

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

Functional analysis

- 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

Functional analysis

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

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

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

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

  - We can easily demonstrate the reproducing property:

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

- 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 $\mathcal{D} = (x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$, 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

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

f^* admits a representation of the form:

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

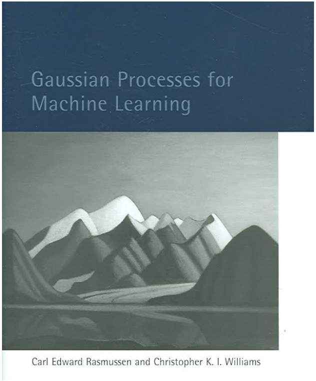
where $\alpha_i \in \mathbb{R}$.

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.

Questions?

Text book


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

References i