Notes on Probability Theory

Gyubeom Edward Im*

February 26, 2024

Contents

*blog: alida.tistory.com, email: criterion.im@gmail.com

Tip

NOMENCLATURE of Probability Theory

- 확률(probability)는 $Pr(\cdot)$으로 표기한다.
- 사건(event)은 대문자로 표기힌다. e.g., A, B
- 이산 확률질량함수(pmf)와 연속 확률밀도함수(pdf)는 각각 $P(\cdot)$와 $p(\cdot)$으로 표기한다.
- 확률변수(random variable)는 소문자로 표기한다. e.g., x, y
- 확률의 파라미터는 사건이고 pdf, pmf의 파라미터는 확률변수이다. e.g., $Pr(A), P(x), p(x)$

1 Introduction

현대 확률론의 수학적인 정의는 20세기 수학자 Andrey Kolmogorov에 의해 정립되었다. 이번 섹션에서는 확률론을 설명하기 위한 기반 이론이 되는 set theory와 measure theory을 설명한다. 해당 이론에 대한 대부분의 내용은 [[6]]를 참고하여 작성하였다.

2 Set theory

집합론(set theory)은 수학의 기본적인 개념인 집합과 그 집합들 간의 관계, 연산 등을 연구하는 수학의 한 분야이다. 집합론은 수학의 거의 모든 분야에 걸쳐 기초적인 언어와 도구를 제공한다. 다양한 집합론의 용어를 먼저 정의해보자. 옷장과 옷으로 비유하여 생각해보면

- **집합(set)**이란 옷장을 의미하고

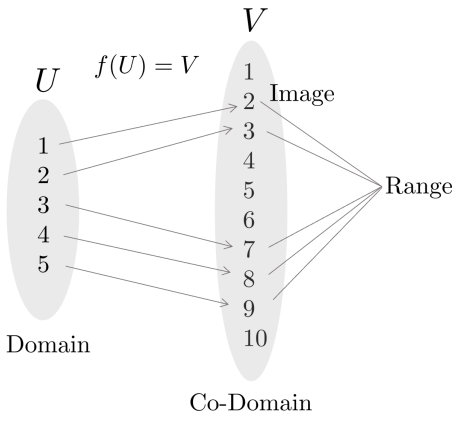
- **원소(element)**란 옷을 의미한다.
- **부분집합(subset)**이란 옷들 중 일부분을 의미하며
- **전체집합(universal set)**은 옷장의 모든 옷을 의미한다.
- **집합 연산자(set operator)**는 옷으로 할 수 있는 연산(e.g., 옷이 몇개 있는가)을 의미한다.
- **서로소 집합(disjoint set)**은 청바지와 코트처럼 교집합이 없는 집합을 의미한다($A \cap B = \emptyset$)
- **분할(partition of A)**는 집합 A를 서로소 집합으로 나누는 것을 의미한다. $A = \{1, 2, 3, 4\} \rightarrow \{\{1, 2\}, \{3\}, \{4\}\}$
- **곱집합(Cartesian product, 데카르트 곱)**은 두 집합 A,B가 있을 때 각각의 집합에서 한 개씩 가져와서 쌍(pair)를 이루는 것을 말한다. $A \times B = \{(a, b) : a \in A, b \in B\}$. 왼쪽 예시는 2차원 벡터 공간 $\mathbb{R}^2$의 예시이다.
- **멱집합(power set)**은 집합 A의 모든 부분집합의 집합을 의미하며 2^A로 표기한다. $A = \{1, 2, 3\}$인 경우 멱집합은 $\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$과 같이 8개가 된다.

2.1 Cardinality

집합 A의 크기(cardinality)는 $|A|$와 같이 표기한다. $|A| = m$이고$|B| = n$이면 둘의 곱집합은 $|A \times B| = mn$이 된다.

- 멱집합의 cardinality는 $|A| = n$인 경우 $|2^A| = 2^n$이 된다.
- 만약 두 집합이 일대일 대응(one-to-one correspondence)을 보인다면 두 집합의 cardinality는 동일하다.
- **가산집합(countable set)**은 자연수와 일대일 대응을 이루는 함수로 셀 수 있는(countable) 집합을 말한다. 셀 수 있다고 반드시 유한할 필요는 없다.
- 자연수 집합과 분수들의 집합과 같이 셀 수 있으나 크기가 무한인(countable infinite) 집합을 **가부번(denumerable) 집합**이라고 한다. 또는 aleph null($\aleph_0$)이라고도 부른다.
- **비가산집합(uncountable set)**이란 가산집합과 달리 셀 수 없는 집합을 말하며 $\mathbf{c}$(continuum)라고 부르거나 $\mathbf{c} = 2^{\aleph_0}$라고 표기한다. 예를 들면 0과 1 사이의 실수의 개수들의 집합이 비가산 집합에 해당한다.

2.2 Function



함수 f는 집합 U에서 다른 집합 V로 변환 또는 매핑하는 연산자를 말한다.

$$f : U \rightarrow V \tag{1}$$

이 때, U를 **정의역(domain)**이라고 부르며 V를 **공역(co-domain)**이라고 한다. **상(image)**이란 주어진 입력 U에 대해 매핑된 출력 V를 의미한다. **치역(range)**이란 정의역 내에 있는 입력 U들에 의해 매핑된 모든 출력 V의 집합을 의미한다.

- domain U, co-domain V
- image: $f(A) = \{f(x) \in V : x \in A\}, A \subseteq U$
- range: $f(U)$
- inverse image(=preimage) : $f^{-1}(B) = \{x \in U : f(x) \in B\}, B \subseteq V$

‘cm‘cm

Onto는 전사함수(surjective)라고도 불리며 공역이 치역과 같은 경우를 의미한다. 이는 co-domain의 모든 원소들이 사영된 것을 의미한다. **One-to-one은 일대일함수(injective)**라고도 불리며 정의역의 원소와 공역의 원소가 하나씩 대응되는 함수를 의미한다. **함수 f가 역함수 f^{-1}를 가지기 위해서는(invertible)** 전사함수이면서 동시에 일대일 함수이어야 한다.

3 Measure theory

측도론(measure theory)은 크기, 길이, 면적, 부피 등을 일반화한 측도(measure)의 개념을 다루는 수학의 분야이다. 이 이론은 특히 확률론과 함수해석학에서 중요한 역할을 한다. 측도론의 기본적인 아이디어는 집합에 숫자를 할당하여 그 집합의 크기를 측정하는 것이다.

예를 들어, 실수 집합의 부분집합에 대해 길이를 할당할 수 있고, 이를 통해 무한대의 집합이나 아주 작은 집합의 크기를 정량화할 수 있다. **집합 함수(set function)**란 하나의 집합에 하나의 값(measure)을 할당하는 함수를 말한다. 이는 앞서 말한 것처럼 크기, 길이, 면적, 부피 등이 될 수 있다.

3.1 σ-field

σ-field $\mathcal{B}$란 '측정 가능한 집합'들을 정의하기 위한 집합들의 컬렉션을 의미한다. 100명의 사람들의 몸무게를 재는 것으로 비유하여 σ-field를 만족하기 위한 세 가지 정의를 살펴보자.

1. $\emptyset \in \mathcal{B}$: 공집합을 포함해야 한다(e.g., 아무 사람의 몸무게도 재지 않은 기준점이 필요하다).
2. $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$: 어떤 집합 A가 σ-field에 속한다면 이의 여집합 A^c 또한 σ-field에 속해야 한다(e.g., 2명에 대해 몸무게를 잴 수 있다면 98명에 대해서도 몸무게를 잴 수 있어야 한다.).
3. $A_i \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{B}$: σ-field에 속하는 임의의 집합 시퀀스 $A_i, i = 1, 2, \cdots$에 대하여 이 집합들의 합집합도 σ-field에 속해야 한다(e.g., a라는 사람의 몸무게를 잴 수 있고 b라는 사람의 몸무게를 잴 수 있다면 a+b 둘을 합쳤을 때도 몸무게를 잴 수 있어야 한다).

3.1.1 Properties of σ-field

- 전체집합 U을 포함한다. : $U \in \mathcal{B}$
- 가산 합집합(countable union)에 대하여 닫혀있다. : $A_i \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{B}$
- 가산 교집합(countable intersection)에 대하여 닫혀있다. : $A_i \in \mathcal{B} \Rightarrow \cap_{i=1}^{\infty} A_i \in \mathcal{B}$
- 멱집합 2^U는 가장 잘게 나눌 수 있는 σ-field이다.
- 유한할 수 있고(finite) 셀 수 없을 수 있으나(uncountable) 가부번할 수 없다(never denumerable).
- $\mathcal{B}, \mathcal{C}$가 σ-field인 경우 $\mathcal{B} \cap \mathcal{C}$은 σ-field이지만 $\mathcal{B} \cup \mathcal{C}$은 σ-field가 아니다.
- 집합 A에 대하여 만들어진 σ-field는 $\sigma(A)$와 같이 표기한다.

3.2 Measurable space

임의의 집합 U와 U의 부분집합으로 이루어진 σ-field $\mathcal{B}$가 주어졌다면 **가측공간(measurable space)**는 $(U, \mathcal{B})$ 같이 정의할 수 있다.

측도(measure) μ란 가측공간 $(U, \mathcal{B})$에서 σ-field의 원소를 사용하여 $[0, \infty]$의 값을 반환하는 일종의 집합함수(set function)을 말한다.

$$\mu : \mathcal{B} \to [0, \infty] \tag{2}$$

- $\mu(\emptyset) = 0$: 공집합에 대한 측도는 0이다.
- 서로소 집합들 A_i, A_j들에 대하여 $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$가 성립한다.

  - 100명의 몸무게를 잴 때 1 10번(A) 사람의 몸무게를 잰 것과 65-75번(B) 사람의 몸무게를 잰 것은 서로 다르지만 이 두 집합을 뭉쳐서(A+B) 한 번에 몸무게를 잰 것과 각 집합(A,B)들의 몸무게를 잰 값을 더한 값은 서로 같아야 한다.

- 가측공간과 측도를 합하여 $(U, \mathcal{B}, \mu)$와 같이 표기하기도 한다.

4 Probability

측도론에서 확률(probability)이란 전체집합 U에 대하여 $\mu(U) = 1$의 크기를 만족하도록 정규화된 측도(normalized measure)를 의미한다.

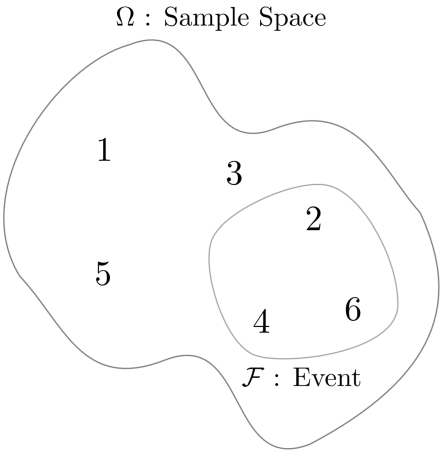
4.1 Random experiment

1부터 6까지 나올 확률이 모두 동일한 주사위(fair dice)가 주어졌다고 하자. 전체 집합 $U = \{1, 2, 3, 4, 5, 6\}$에 대한 σ-field $\mathcal{B}$는 어떻게 정할까? 확률 문제는 일반적으로 다음과 같은 질문을 묻는다.

- 주사위가 1이 나올 확률은 얼마인가?
- 주사위가 4 또는 6이 나올 확률은 얼마인가?
- 주사위가 2,3,5 중 하나가 나올 확률은 얼마인가?

이와 같이 집합 U에 대한 모든 부분집합을 사용할 수 있어야 되기 때문에 $\mathcal{B}$는 보통 멱집합 2^U을 사용한다. **확률에서 가측공간 $(U, \mathcal{B}, \mu)$는 특별히 확률공간(probability space)라고 하며 일반적으로 $(\Omega, \mathcal{F}, Pr)$로 표기한다.** 각 기호들의 설명은 다음과 같다.

- Ω: 표본공간(sample space)라고 하며 이름은 공간이 들어가지만 나올 수 있는 원소들의 전체 집합(=set)을 의미한다.
- $\mathcal{F}$: 표본공간의 부분집합을 의미하며 사건(event)이라고 부른다.
- Pr: 측도(measure)를 수행하는 연산자로써 표본공간의 원소에 확률을 부여하는 역할을 한다.



어떠한 현상으로부터 결과를 얻기 위해서는 **실험(experiment)**를 해야하고 이 때 우리는 **실현값(outcome)**을 얻을 수 있다. 만약 실험의 결과가 매 번 다르고 우리가 실현값의 결과를 하나의 표본공간으로 정의할 수 있다고 하면 이는 **확률 실험(random experiment)**라고 부른다.

따라서 주사위 던지기는 확률 실험의 일종이며 우리는 표본공간 $\Omega = \{1, 2, 3, 4, 5, 6\}$에서 확률 $Pr(\cdot)$이라는 측도(measure)를 사용하여 공간 내 원소의 크기를 측정한 실현값(outcome)을 얻을 수 있다.

$$\begin{aligned} &Pr(1) = Pr(2) = Pr(3) = Pr(4) = Pr(5) = Pr(6) = 1/6 \\ &Pr(A) = Pr(2, 4, 6) = Pr(2) + Pr(4) + Pr(6) = 1/2 \end{aligned} \tag{3}$$

4.2 Probability axioms

확률공간 $(\Omega, \mathcal{F}, Pr)$에 정의된 확률 $Pr(\cdot)$은 하나의 집합 함수(set function)이며 다음과 같이 정의된다.

$$Pr(\mathcal{F}) \rightarrow [0, 1] \tag{4}$$

확률의 공리(axiom)는 다음과 같다.

- $Pr(\emptyset) = 0$: 공집합에 대한 측도는 0이다.
- $Pr(A \geq 0) \quad \forall A \subseteq \Omega$: 확률은 항상 양의 값을 가진다.
- 서로소 집합들 A_i, A_j들에 대하여 $Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} Pr(A_i)$가 성립한다.
- $Pr(\Omega) = 1$: **전체 표본공간에 대한 확률의 총량은 1이다.** 위 조건에서 보다시피 **확률이란 기존 측도(measure)의 세가지 조건에 마지막 조건(총합=1)이 추가된 특수한 버전으로 해석할 수 있다.**

4.3 Probability allocation function

지금까지 확률공간 $(\Omega, \mathcal{F}, Pr)$을 정의하였고 표본공간의 부분집합인 $\mathcal{F}$의 크기를 측도(measure)함으로써 확률을 측정할 수 있는 연산자 Pr를 정의하였다. 이 때, 표본공간 Ω은 확률 실험이 이산확률인지 연속확률인지에 따라 달라진다. 서로 다른 표본공간에서 대하여 각각 측도 Pr를 수행할 수 있는 함수 $P(\cdot), p(\cdot)$를 정의할 수 있고 이를 확률할당함수(probability allocation function)이라고 한다.

4.3.1 Discrete sample space Ω:

$$\boxed{P : \Omega \rightarrow [0, 1]}$$

such that

$$\begin{aligned} &\sum_{w \in \Omega} P(w) = 1 \\ &Pr(A) = \sum_{w \in A} P(w) \end{aligned} \tag{5}$$

4.3.2 Continuous sample space Ω:

$$\boxed{p : \Omega \rightarrow [0, \infty)}$$

such that

$$\begin{aligned} &\int_{w \in \Omega} p(w) dw = 1 \\ &Pr(A) = \int_{w \in A} p(w) dw \end{aligned} \tag{6}$$

일반적으로 전자를 **확률질량함수(probability mass function, pmf)**라고 하며 후자를 **확률밀도함수(probability density function, pdf)**라고 한다.

[그림]

4.4 Independence

두 사건(event) A, B이 서로 독립 사건임을 보이기 위해서는 다음의 정의를 만족해야 한다.

$$\begin{aligned} Pr(A \cap B) &= Pr(A)Pr(B) \\ Pr(A|B) &= Pr(A) \end{aligned} \tag{7}$$

위 식에서 보다시피 두 사건의 공집합이 없다면 이는 더 이상 독립이 아니다. 따라서 두 사건이 서로소(disjoint)이거나 상호 배제(mutually exclusive)한 상황에서는 사건의 독립을 보일 수 없음을 유의해야 한다.

4.5 Joint probability

결합 확률(joint probability)이란 두 사건 A, B가 동시에 발생할 확률(=교집합이 발생할 확률)을 의미한다.

$$Pr(A \cap B) \tag{8}$$

4.6 Marginal probability

주변 확률(marginal probability)는 여러 사건들 사이에서 하나의 사건만을 고려하는 것을 말한다. 예를 들어, 두 사건 A와 B가 있고, 이들의 결합 확률(joint probability)가 주어졌을 때, A의 주변 확률은 B가 취할 수 있는 모든 값에 대해 A의 확률을 합하여 얻어진다.

$$Pr(A) = \sum_{B} Pr(A \cap B) \tag{9}$$

4.7 Conditional probability

조건부 확률은 두 사건 A, B에 대하여 B가 발생했을 때 A가 발생할 확률을 의미한다.

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} \tag{10}$$

이를 통해 두 사건이 동시에 발생한 확률은 $Pr(A \cap B) = Pr(A)Pr(B|A)$와 같이 나타낼 수 있다. 이 때, $Pr(A \cap B) = Pr(B \cap A)$이므로 A, B 순서를 바꿔도 공식이 성립한다. 이는 A가 발생했을 때 B가 발생할 확률을 의미한다.

$$Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)} \tag{11}$$

만약 두 사건 A, B가 독립이면 조건부 확률은 다음과 같다.

$$Pr(A|B) = Pr(A)Pr(B) \tag{12}$$

4.8 Bayesian rule

Bayesian rule은 다음과 같은 조건부확률 간 관계를 의미한다.

$$\begin{aligned} Pr(A|B) &= \frac{Pr(A \cap B)}{Pr(B)} \\ &= \frac{Pr(B|A)Pr(A)}{Pr(B)} \end{aligned} \tag{13}$$

- $Pr(A|B)$: posterior probability
- $Pr(B|A)$: likelihood
- $Pr(A)$: prior probability

5 Random variables

확률변수(random variable) x는 표본공간 Ω에 정의된 함수를 의미한다. 이 함수는 확률공간 $(\Omega, \mathcal{F}, Pr)$의 한 원소를 Borel 가측공간 $(\mathbb{R}, \mathcal{B})$의 원소로 변환하는 역할을 수행한다.

$$\begin{gathered} \boxed{x : \Omega \to \mathbb{R}} \\ \text{such that} \\ \forall B \in \mathcal{B} \\ x^{-1}(B) \in \mathcal{F} \end{gathered} \tag{14}$$

- Borel 가측공간(mesurable space): 실수들의 집합으로 만들어진 공간을 Borel 가측공간이라고 하며 이 때 σ-field를 Borel set이라고 한다.

- 확률(probability)은 표본공간의 부분집합인 σ-field를 하나의 실수값으로 측정해주는 연산자(=measure)인 반면에, 확률변수(random variable)는 표본공간의 하나의 원소를 하나의 실수값으로 변환해주는 함수(=function)를 말한다. **즉, 확률과 달리 확률변수는 하나의 원소에 대해서만 변환이 가능하다.**
- 확률변수가 함수라면 무엇이 **무작위성(randomness)**이 있다는 것일까? 표본공간 Ω에서 하나의 원소를추출할 때 무작위로 하나를 선택한 후 하나의 실수값으로 변환하기 때문에 일반적으로 확률변수가 무작위성이 있다고 한다.
- 표본공간 Ω의 하나의 원소 w가 있을 때 $x(w)$를 **실현(realization)**이라고 한다. 간단히 말하자면 **샘플링**을 의미한다(e.g., 가우시안 샘플링).
- x에 대한 모든 실현값들의 집합을 알파벳(alphabet of x)이라고 한다.
- 우리는 확률변수 x 자체보다 x의 확률 $Pr(x)$에 관심이 있다. 실현값이 $x \in B, B \in \mathcal{B}$일 때 다음과 같이 정의한다.
-

$$Pr(x \in B) \triangleq Pr(x^{-1}(B)) = Pr(\{w : x(w) \in B\}) \tag{15}$$

5.1 Discrete random variable

주사위 굴리기나 동전 던지기 같이 값이 **유한하거나 셀 수 있는 무한의 값들을 가지는 확률변수를 이산확률변수**라고 한다. 확률질량함수(probability mass function, pmf) $P(\cdot)$를 사용하여 각 값에 대한 확률을 나타내며 각각의 개별 값에 대해 명확한 확률을 할당할 수 있다.

5.2 Continuous random variable

온도 측정이나 물체의 길이 측정, 주식 가격 등 **연속적인 범위의 값을 가지는 확률변수를 연속확률변수**라고 한다. 확률밀도함수(probability density function, pdf) $p(\cdot)$을 사용하여 값의 범위에 대한 확률을 나타내며 개별 값에 대한 확률을 표현할 수 없으나 범위에 대한 확률을 표현할 수 있는 특징이 있다.

6 Probability distribution

확률분포(probability distribution)은 확률변수가 가질 수 있는 모든 값들과 그에 대응하는 확률들이 어떻게 분포하고 있는지 정의해놓은 함수를 말한다. 어떤 확률변수 x, y가 확률함수 p에 대해 같은 분포를 가져도 둘은 다른 확률변수일 수 있음에 유의한다.

6.1 Discrete probability distribution

6.1.1 Bernoulli distribution

확률 변수의 값이 성공 혹은 실패로 나타나는 경우의 확률 분포를 베르누이 분포라고 한다. 확률 실험의 결과 값이 성공 혹은 실패로 나타나는 실험을 베르누이 실험(Bernoulli experiment)라고 한다. 성공 확률이 p인 베르누이 실험에서 성공의 횟수를 나타내는 확률 분포이다. 따라서 확률 변수의 영역이 $x : \{\text{success, fail}\} \to \{0, 1\}$이다.

$$\begin{aligned} P(x = 0) &= 1 - p \\ P(x = 1) &= p \end{aligned} \tag{16}$$

확률질량함수(pmf) $P(x)$는 다음과 같다.

$$\boxed{\begin{aligned} &x \sim \text{Ber}(p) \\ &P(x) = p^x(1-p)^{1-x} \quad x = 0, 1 \end{aligned}} \tag{17}$$

6.2 Continuous probability distribution

6.2.1 Gaussian distribution

스칼라 확률변수 x가 가우시안 분포를 따른다고 하면 일반적으로 다음과 같이 표기한다.

$$x \sim \mathcal{N}(\mu, \sigma^2) \tag{18}$$

- $x \sim \mathcal{N}(\mu, \sigma^2)$: 확률변수 x가 평균이 μ이고 분산이 σ^2인 가우시안 분포를 따른다는 의미

이 때 pdf $p(x)$는 다음과 같이 정의된다.

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - \mu)^2 \right)} \tag{19}$$

6.2.2 Chi-square distribution

카이스퀘어(chi-square) 분포는 관측 데이터 $[x[1], x[2], \cdots, x[n]]^\intercal$가 서로 독립이며 동일한 분포를 갖고 있을 때(=i.i.d), 다음과 같이 나타낼 수 있다.

$$y = \sum_{i=1}^{n} x_i^2 \sim \mathcal{X}_n^2 \tag{20}$$

- $x_i \sim \mathcal{N}(0, 1),\ i = 1, 2, \cdots, n$: 평균이 0이고 분산이 1인 표준정규분포를 따른다
- $\mathcal{X}_n^2$: 자유도가 n인 카이스퀘어 분포

즉, 확률 변수 x_i의 제곱의 합은 카이스퀘어 분포를 따른다. y의 pdf는 다음과 같이 나타낼 수 있다.

$$\boxed{p(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} \exp(-\frac{1}{2}y) & y \geq 0 \\ 0 & y < 0 \end{cases}} \tag{21}$$

- $\Gamma(u) = \int_0^\infty t^{u-1} e^{-t} dt$: 감마 적분 함수. 자연수 n에 대하여 $\Gamma(n) = (n-1)!$이 성립한다.

y의 평균과 분산은 다음과 같다.

$$\begin{aligned} \mathbb{E}(y) &= n \\ \text{var}(y) &= 2n \end{aligned} \tag{22}$$

6.3 Joint probability distribution

결합확률분포(joint probability distribution)란 두 확률변수 x, y가 동시에 발생할 확률(=교집합이 발생할 확률)을 의미한다.

$$\begin{aligned} &P(x \cap y) \quad \cdots \text{ for discrete probability} \\ &p(x, y) \quad \cdots \text{ for continuous probability} \end{aligned} \tag{23}$$

6.4 Marginal probability distribution

주변확률분포(marginal probability)는 여러 확률변수들 사이에서 한 확률변수의 행동만을 고려하는 것을 말한다. 예를 들어, 두 이산확률변수 x와 y가 있고, 이들의 결합확률분포(joint probability distribution)가 주어졌을 때, x의 주변확률분포는 y가 취할 수 있는 모든 값에 대해 x의 확률을 합하여 얻어진다. 이산확률변수를 보면 다음과 같다.

$$P(x) = \sum_y P(x \cap y) \tag{24}$$

연속확률변수로 나타내면 다음과 같다.

$$p(x) = \int p(x, y) dy = \int p(x|y) p(y) dy \tag{25}$$

6.5 Conditional probability distribution

조건부 확률분포(conditional probability distribution)은 두 확률변수 x, y에 대하여 y가 발생했을 때 x가 발생할 확률을 의미한다. 이산확률변수에 대한 조건부 확률분포는 다음과 같다.

$$P(x|y) = \frac{P(x \cap y)}{P(y)} \tag{26}$$

이를 통해 두 확률변수가 동시에 발생한 확률는 $P(x \cap y) = P(x)P(y|x)$와 같이 나타낼 수 있다. 이 때, $P(x \cap y) = P(y \cap x)$이므로 x, y 순서를 바꿔도 공식이 성립한다. 이는 x가 발생했을 때 y가 발생할 확률분포를 의미한다.

$$P(y|x) = \frac{P(y \cap x)}{P(x)} \tag{27}$$

만약 두 확률변수 x, y가 독립이면 조건부 확률분포는 다음과 같다.

$$P(x|y) = P(x)P(y) \tag{28}$$

연속확률분포에 대한 조건부 확률분포은 다음과 같다. 연속확률변수가 주어졌을 때, y가 발생했을 때 x가 발생할 확률분포는 다음과 같다.

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{29}$$

반대로 x가 발생했을 때 y가 발생할 확률분포는 다음과 같다.

$$p(y|x) = \frac{p(y, x)}{p(x)} \tag{30}$$

이 때, $p(x, y)$와 $p(y, x)$는 동일하다.

6.6 Bayesian rule

Bayesian rule은 다음과 같은 조건부확률분포 사이의 관계를 의미한다.

$$\begin{aligned} p(x|y) &= \frac{p(x, y)}{p(y)} \\ &= \frac{p(y|x)p(x)}{p(y)} \end{aligned} \tag{31}$$

- $p(x|y)$: posterior pdf
- $p(y|x)$: likelihood
- $p(x)$: prior pdf

예를 들어, 로봇의 위치를 $\mathbf{x}$, 로봇의 센서를 통해 관측한 값을 $\mathbf{z}$이라고 했을 때 주어진 관측 데이터를 바탕으로 현재 로봇이 $\mathbf{x}$에 위치할 확률 $p(\mathbf{x}|\mathbf{z})$는 다음과 같이 나타낼 수 있다.

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z})} = \eta \cdot p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \tag{32}$$

- $p(\mathbf{x}|\mathbf{z})$: 관측값 $\mathbf{z}$이 주어졌을 때 로봇이 $\mathbf{x}$에 위치할 확률 (posterior)
- $p(\mathbf{z}|\mathbf{x})$: $\mathbf{x}$ 위치에서 관측값 $\mathbf{z}$가 나올 확률 (likelihood) - $p(\mathbf{x})$: 로봇이 $\mathbf{x}$ 위치에 존재할 확률 (prior)
- $\eta = 1/p(\mathbf{z})$: 전체 확률분포의 넓이가 1이 되어 확률분포의 정의를 유지시켜주는 normalization factor 이다. 주로 η로 치환하여 표현한다.

7 Momentum

모멘텀(momentum, 또는 적률)은 확률분포의 특징을 설명해주는 지표를 의미한다. 1차 적률은 확률분포의 **평균(mean)**을 의미하고 2차 적률은 **분산(variance)**를 의미하며 3차 적률은 **왜도(skewness)**, 4차 적률은 **첨도(kurtosis)**를 의미한다. 왜도는 확률분포의 비대칭성을 나타내는 지표이고 첨도는 확률분포의 뾰족한 정도를 나타내는 지표이다.

$$\boxed{\begin{aligned}
&\mu = \mathbb{E}(x) && \cdots \text{1st moment} \\
&\sigma^2 = \text{var}(x) = \mathbb{E}((x-\mu)^2) && \cdots \text{2nd moment} \\
&\text{skewness} = \frac{\mathbb{E}((x-\mu)^3)}{\sigma^3} && \cdots \text{3rd moment} \\
&\text{kurtosis} = \frac{\mathbb{E}((x-\mu)^4)}{\sigma^4} && \cdots \text{4th moment}
\end{aligned}} \tag{33}$$

임의의 두 분포 A, B가 있을 때 두 분포가 같은지 판단하려면 일반적으로 두 분포의 모멘텀이 같은지를 판단하면 된다.

7.1 Covariance and correlation

두 확률변수 x, y가 주어졌을 때 공분산(covarinace)이란 두 분포가 어떤 상관관계를 가지는지 나타내는 값을 말한다.

$$\boxed{\text{cov}(x,y) = \mathbb{E}[(x-\mu_x)(y-\mu_y)]} \tag{34}$$

- μ_x, μ_y: x와 y의 평균

이를 전개해보면 다음과 같다.

$$\begin{aligned}
\text{cov}(x,y) &= \mathbb{E}[(x-\mu_x)(y-\mu_y)] \\
&= \mathbb{E}[xy - \mu_x y - \mu_y x + \mu_x \mu_y] \\
&= \mathbb{E}(xy) - \mu_y \mathbb{E}(x) - \mu_x \mathbb{E}(y) + \mu_x \mu_y \\
&= \mathbb{E}(xy) - \mu_x \mu_y
\end{aligned} \tag{35}$$

- $\text{cov}(x,y) > 0$: 두 확률변수가 양의 상관관계를 갖는다.
- $\text{cov}(x,y) < 0$: 두 확률변수가 음의 상관관계를 갖는다.
- $\text{cov}(x,y) = 0$: 두 확률변수의 상관관계가 없다(uncorrelated).

만약 두 확률변수가 상관관계가 없거나(uncorrelated) x, y가 독립(independence)이라면 $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y) = \mu_x \mu_y$가 되어 $\text{cov}(x,y) = 0$이 된다. **따라서 두 개념이 동일하다고 생각할 수 있으나 둘은 다른 개념이다. 두 확률변수가 상관관계가 없다고 하더라도(uncorrelated) 두 확률변수는 독립(independence)이 아닐 수 있다.** 상관관계가 없다는 말은 선형 상관관계가 없다는 의미이므로 비선형적으로 상관관계가 있을 수 있음을 암시한다. 따라서 독립성의 개념이 조금 더 강한 개념으로 서로의 확률에 어떠한 영향도 주지 않음을 의미한다. 이 둘의 관계를 헷갈리지 않도록 유의한다.

$$\begin{aligned}
&\text{independence} \;\Rightarrow\; \text{uncorrelated} \\
&\text{uncorrelated} \;\not\Rightarrow\; \text{independence}
\end{aligned} \tag{36}$$

7.1.1 Correlation coefficient

상관계수 ρ_{xy}는 공분산이 단위의 영향을 받는 점을 고려하여 이를 정규화한 값을 말한다.

$$\rho_{xy} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \tag{37}$$

상관계수는 $-1 \le \rho_{xy} \le 1$의 범위를 갖는다.

7.2 Orthogonal

두 확률변수 x, y가 직교한다(orhtogonal)는 의미는 두 확률변수의 곱의 기대값이 0임을 의미한다.

$$\mathbb{E}(xy) = 0 \tag{38}$$

7.3 Expectation

기대값(expectation, expected value) $\mathbb{E}$란 확률적 사건에 대한 평균을 의미하며 사건이 벌어졌을 때 이득과 그 사건이 벌어질 확률을 곱한 것을 합한 값을 말한다. 표본공간 Ω에서 정의된 확률변수 x가 있을 때 확률함수 p에 대한 x의 기대값은 $\mathbb{E}[x]$라고 하고 다음과 같은 식으로 나타낸다.

$$\boxed{\mathbb{E}[x] = \sum_{x \in \Omega} x \cdot P(x)} \tag{39}$$

위 식은 이산확률변수에 대한 기대값을 의미한다. 연속확률변수에 대한 기대값은 다음과 같다.

$$\boxed{\mathbb{E}[x] = \int_{-\infty}^{\infty} x \cdot p(x) dx} \tag{40}$$

7.3.1 Properties of expectation

기대값은 선형성(Linearity)라는 성질을 가지고 있다. 수학에서 선형성에 대한 정의는 다음과 같다. 임의의 함수 f에 대해

임의의 수 x,y에 대해 f(x+y) = f(x)+f(y)가 항상 성립하고 임의의 수 x와 a에 대해 f(ax) = af(x)가 항상 성립하면 함수 f는 선형이라고 한다. 따라서 임의의 확률변수 x, y와 임의의 실수 a, b에 대해서 다음 식이 성립하게 된다.

$$\mathbb{E}[ax + by] = a\mathbb{E}[x] + b\mathbb{E}[y] \tag{41}$$

그리고 선형인 함수 $L(x)$에 대해서 기대값과 함수의 계산순서를 바꿀 수 있다.

$$\mathbb{E}[L(x)] = L(\mathbb{E}[x]) \tag{42}$$

7.3.2 Conditional expectation

연속확률변수 x에 대한 기대값 $\mathbb{E}(x) = \int xp(x)dx$는 하나의 결정된 값을 의미하며 더 이상 확률성을 띄고 있지 않다. **하지만 두 확률변수 x, y의 조건부 기대값 $\mathbb{E}(x|y)$는 기대값을 취하더라도 여전히 y에 대한 확률변수가 되며 이 부분이 기대값과 가장 다른 부분이다.**

7.3.3 Law of total expectation

확률 변수 x, y가 주어졌을 때 총 기대값의 법칙(law of total expectation)은 다음과 같이 정의한다.

$$\mathbb{E}[x] = \mathbb{E}[\mathbb{E}[x|y]] \tag{43}$$

자세히 표현하면 아래와 같다.

$$\mathbb{E}_x[x] = \mathbb{E}_y[\mathbb{E}_x[x|y]] \tag{44}$$

- $\mathbb{E}_x$: 확률 변수 x에 대한 기대값
- $\mathbb{E}_y$: 확률 변수 y에 대한 기대값

두 연속확률변수 x, y에 대하여 증명은 다음과 같이 할 수 있다.

$$\begin{aligned}
\mathbb{E}_y[\mathbb{E}_x[x|y]] &= \int \left(\int xp(x|y)dx \right) p(y)dy \\
&= \int \int xp(x|y)p(y)dxdy \\
&= \int \int xp(x,y)dxdy \\
&= \int x \left(\int p(x,y)dy \right) dx \\
&= \int xp(x)dx \\
&= \mathbb{E}_x[x]
\end{aligned} \tag{45}$$

7.4 Variance and standard deviation

확률변수 x의 **분산(variance)**은 σ^2 또는 $\mathrm{var}[x]$라고 표기하고 다음과 같이 정의한다

$$\boxed{\mathrm{var}[x] = \mathbb{E}[(x - \mathbb{E}(x))^2]} \tag{46}$$

또한 아래와 같이 표현할 수도 있다.

$$\mathrm{var}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \tag{47}$$

분산의 제곱근을 **표준편차(standard deviation)**이라고 하며 σ로 표기한다.

8 More on Gaussian distribution

8.1 Joint gaussian distribution

두 개의 확률변수 x, y가 주어졌을 때 두 확률이 동시에 발생할 결합확률분포(joint probability distribution)는 다음과 같이 나타낼 수 있다.

$$\boxed{p(x,y) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix}\right)^T \boldsymbol{\Sigma}^{-1} \left(\begin{bmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(y) \end{bmatrix}\right)\right)} \tag{48}$$

평균은 $\boldsymbol{\mu} = \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(y) \end{bmatrix}$이고 분산은 $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}$이다.

이 때, 분산은 여러 변수에 대한 분산을 의미하고 대각성분들은 하나의 변수에 대한 분산을 의미하며 대각성분이 아닌 성분들은 두 변수 간 상관관계를 의미한다. **이러한 다변수 확률분포에서 분산 Σ을 일반적으로 공분산(covariance)라고 부른다.**

8.2 Multivariate gaussian distribution

벡터 확률변수 $\mathbf{x} \in \mathbb{R}^n$가 가우시안 분포를 따른다고 하자.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{49}$$

평균 $\boldsymbol{\mu} \in \mathbb{R}^n$은 벡터이고 공분산 $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$은 행렬이다. 이 때, 확률분포함수 $p(\mathbf{x})$는 다음과 같이 정의된다.

$$\boxed{p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)} \tag{50}$$

- $|\boldsymbol{\Sigma}|$: $\boldsymbol{\Sigma}$의 행렬식(determinant)
- $\boldsymbol{\Sigma}^{-1}$: information matrix $\boldsymbol{\Omega}$라고도 표현한다.

8.3 Linear transformation of gaussian random variable

벡터 랜덤 변수 $\mathbf{x} \in \mathbb{R}^n$가 가우시안 분포를 따를 때는 다음과 같이 표기할 수 있다.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{51}$$

만약 x를 선형 변환(linear transformation)한 새로운 랜덤변수 $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$가 주어졌다고 하면 y는 아래와 같은 확률 분포를 따른다.

$$\boxed{\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} + \mathbf{b} \\ &\sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\intercal) \end{aligned}} \tag{52}$$

공분산 $\mathrm{cov}(\mathbf{Ax}+\mathbf{b})$는 다음과 같이 유도할 수 있다.

$$\begin{aligned}
\mathrm{cov}(\mathbf{Ax}+\mathbf{b}) &= \mathbb{E}((\mathbf{y}-\boldsymbol{\mu}_\mathbf{y})(\mathbf{y}-\boldsymbol{\mu}_\mathbf{y})^\intercal) \\
&= \mathbb{E}((\mathbf{y}-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))(\mathbf{y}-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))^\intercal) \\
&= \mathbb{E}(((\mathbf{Ax}+\mathbf{b})-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))((\mathbf{Ax}+\mathbf{b})-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))^\intercal) \\
&= \mathbb{E}([\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})]\,[\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})]^\intercal) \\
&= \mathbb{E}(\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathbf{A}^\intercal) \\
&= \mathbf{A}\mathbb{E}((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal)\mathbf{A}^\intercal \\
&= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\intercal
\end{aligned} \tag{53}$$

8.4 Conditional gaussian distribution

두 개의 확률변수 x, y가 주어졌을 때 조건부 확률분포 $p(x|y)$가 가우시안 분포를 따른다고 하면

$$\begin{aligned}
p(x|y) &= \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} = \eta \cdot p(y|x)p(x) \\
&\sim \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y})
\end{aligned} \tag{54}$$

가 된다 이 때 평균 $\boldsymbol{\mu}_{x|y}$과 분산 $\boldsymbol{\Sigma}_{x|y}$은 아래와 같다.

$$\boxed{\begin{aligned}
\boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(y-\boldsymbol{\mu}_y) \\
\boldsymbol{\Sigma}_{x|y} &= \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}^\intercal
\end{aligned}} \tag{55}$$

9 Random Process

랜덤 프로세스(random process)는 확률변수(random variable)을 무한차원으로 확장한 버전으로 생각하면 된다. 지금까지 배운 확률변수 x는 표본공간 Ω에서 **하나의** 표본을 **하나의** 실수로 변환해주는 연산자의 역할을 수행하였다. 벡터 확률변수 $\mathbf{x} \in \mathbb{R}^n$을 생각해보면 표본공간 Ω에서 n**개**의 표본을 n**개**의 실수 벡터로 변환하는 연산자의 역할을 수행하였다. 만약 n을 무한으로 확장하면 어떻게 될까?

$n \to \infty$가 된다면 벡터 확률변수 $\mathbf{x}$는 표본공간 Ω에서 ∞**개**의 표본을 ∞**개**의 실수 벡터로 변환하는 연산자가 될 것이다. **이는 표본공간 Ω에서 하나의 함수를 변환하는 연산자로 볼 수 있을 것이다. 함수해석학적으로 봤을 때 함수 $y=f(x)$는 x를 넣으면 y가 나오는 무한차원의 벡터로 해석할 수 있다.** 예를 들어, 입출력이 실수라면 x도 무한개 y도 무한개인 벡터가 된다.

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_\infty \end{bmatrix}\right) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\infty \end{bmatrix} \tag{56}$$

9.1 Definition of random process

랜덤 프로세스는 다음과 같이 정의한다.

$$x_t(w), \quad \text{where, } t \in \mathcal{I} \tag{57}$$

- $\mathcal{I}$: 인덱스 집합(Index set). 일반적으로 시간(t)으로 간주한다.

이는 앞서 정의한 확률변수 x와 t의 존재만 제외하고는 동일하다. t는 일반적으로 시간으로 간주한다. 랜덤 프로세스는 랜덤 시퀀스(random sequence), 랜덤 함수(random function), 랜덤 신호(random signal) 등으로 해석할 수 있으며

$$x_t : \Omega \to \text{ the set of all sequences or functions} \tag{58}$$

와 같이 표기하기도 한다.

10 Gaussian Process

11 Gaussian Process Regression

12 References

[1] (Blog) 평균과 기댓값

[2] (Blog) PRLM - 4. The Exponential Family

[3] (Blog) [수리통계학] 38. 지수족

[4] (Wiki) Law of total expectation

[5] (Blog) 2 장 확률변수와 확률분포

[6] (Lecture) Bayesian Deep Learning - 최성준

[7] Williams, Christopher KI, and Carl Edward Rasmussen. Gaussian processes for machine learning. Vol. 2. No. 3. Cambridge, MA: MIT press, 2006.

13 Revision log

- 1st: 2024-02-09
- 2nd: 2024-02-24
- 3rd: 2024-02-26