Notes on Probability Theory

Gyubeom Edward Im*

February 26, 2024

Contents

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

1 Set theory

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.

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.

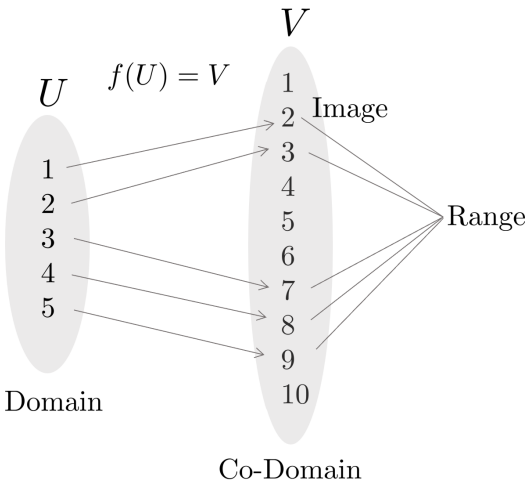
1.1 Cardinality

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.

1.2 Function

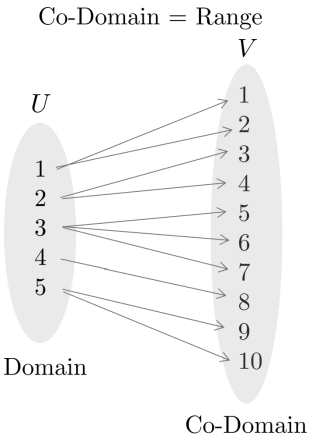


함수 f는 집합 U에서 다른 집합 V로 변환 또는 매핑하는 연산자를 말한다.

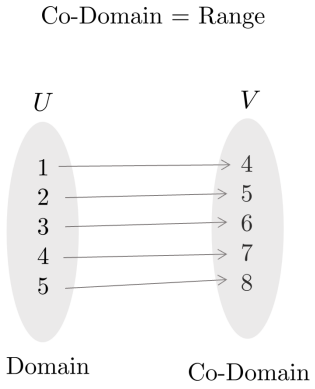
$$f : U \to V \tag{1}$$

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) : $f^{-1}(B) = \{x \in U : f(x) \in B\}, B \subseteq V$



Surjective
Not Injective



Surjective
Injective

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.

2 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.

2.1 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.

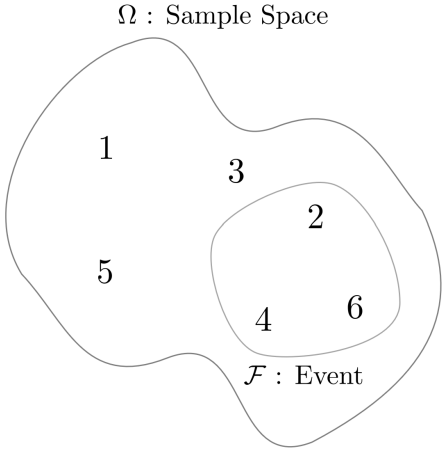
1. $\emptyset \in \mathcal{B}$: 공집합을 포함해야 한다(e.g., 아무 사람의 몸무게도 재지 않는 기준점이 필요하다).
2. $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$: 어떤 집합 A가 σ-field에 속한다면 이의 여집합 A^c 또한 σ-field에 속해야 한다(e.g., 2명에 대해 몸무게를 잴 수 있다면 98명에 대해서도 몸무게를 잴 수 있어야 한다.).
3. $A_i \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{B}$: 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).

2.1.1 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.

2.2 Measurable space

3 Sample space and event



어떤 시행에서 일어날 수 있는 모든 결과들의 모임을 **표본공간 Ω**라고 한다. 예를 들어 주사위를 한번 던지는 시행의 경우 표본공간은 $\{1, 2, 3, 4, 5, 6\}$과 같은 집합이 된다. 표본공간 Ω의 부분집합을 **사건(event) $\mathcal{F}$**라고 한다.

4 Probability function

확률함수 p는 표본공간의 원소를 0과 1사이의 숫자에 대응시키는 함수를 의미한다. 사건 $\mathcal{F}$ 에 대한 확률은 다음과 같이 정의할 수 있다.

$$\forall \mathcal{F} \in \Omega, \quad p(\mathcal{F}) = \sum_{w \in \mathcal{F}} p(w) \tag{2}$$

5 Random variables and distribution

확률변수(random variable)는 표본공간 Ω에 정의된 함수를 의미한다. 이 때 확률변수의 결과값은 항상 실수이다. 분포(distribution)은 확률변수가 가질 수 있는 값들에 대해서 확률들을 나열해 놓은 것을 의미한다. 중요한 점은 어떤 확률변수 x, y가 확률함수 p에 대해 같은 분포를 가져도 둘은 다른 확률변수일 수 있다.

6 Probability axioms

표본공간 Ω에 사건 $\mathcal{F}$가 있을 때, 사건 $\mathcal{F}$의 확률변수 x가 일어날 확률 $p(x)$는 항상 0 이상 1 이하이다.

$$0 \leq p(x) \leq 1 \quad \forall x \in \mathcal{F} \tag{3}$$

표본공간 Ω 전체가 일어날 확률은 1이다.

$$p(\Omega) = 1 \tag{4}$$

7 Continuous r.v. vs discrete r.v.

연속확률변수(continuous random variable)와 이산확률변수(discrete random variable)는 확률론과 통계학에서 확률 분포의 특성을 기반으로 한 두 가지 주요 범주이다.

7.1 Continuous r.v.

온도 측정이나 물체의 길이 측정, 주식 가격 등 연속적인 범위의 값을 가지는 확률변수를 연속확률변수라고 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.

7.2 Discrete r.v.

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

본 문서에서는 이산확률변수 x에 대한 pmf는 $P(x)$, 연속확률변수에 x대한 pdf는 $p(x)$로 나타낸다.

8 Expected value

기대값(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.

$$\mathbb{E}[x] = \sum_{x \in \Omega} x \cdot P(x) \tag{5}$$

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

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x \cdot p(x) dx \tag{6}$$

8.1 Properties of expected value

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.

$$\mathbb{E}[ax + by] = a\mathbb{E}[x] + b\mathbb{E}[y] \tag{7}$$

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

$$\mathbb{E}[L(x)] = L(\mathbb{E}[x]) \tag{8}$$

8.2 Law of total expectation

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

$$\mathbb{E}[x] = \mathbb{E}[\mathbb{E}[x|y]] \tag{9}$$

자세히 표현하면 아래와 같다.

$$\mathbb{E}_x[x] = \mathbb{E}_y[\mathbb{E}_x[x|y]] \tag{10}$$

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

두 연속확률변수 x, y에 대하여 증명은 다음과 같이 할 수 있다.

$$\begin{aligned}
\mathbb{E}_y[\mathbb{E}_x[x|y]] &= \int \left(\int xp(x|y)dx \right) p(y)dy \\
&= \int \int xp(x|y)p(y)dxdy \\
&= \int \int xp(x,y)dxdy \\
&= \int x \left(\int p(x,y)dy \right) dx \\
&= \int xp(x)dx \\
&= \mathbb{E}_x[x]
\end{aligned} \tag{11}$$

9 Variance and standard deviation

가우시안 분포를 따르는 확률변수 x의 분산은 σ^2 또는 var$[x]$라고 표기하고 다음과 같이 정의한다

$$\text{var}[x] = \mathbb{E}[(x - \mathbb{E}(x))^2] \tag{12}$$

또한 아래와 같이 표현할 수도 있다.

$$\text{var}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \tag{13}$$

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

10 Gaussian distribution

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

$$x \sim \mathcal{N}(\mu, \sigma^2) \tag{14}$$

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

이 때, 확률분포함수(pdf) $p(x)$는 다음과 같이 정의된다.

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - \mu)^2 \right)} \tag{15}$$

11 Joint gaussian distribution

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

$$\boxed{p(x,y) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x-\mathbb{E}(x) \\ y-\mathbb{E}(y) \end{bmatrix}\right)^T \boldsymbol{\Sigma}^{-1} \left(\begin{bmatrix} x-\mathbb{E}(x) \\ y-\mathbb{E}(y) \end{bmatrix}\right)\right)} \tag{16}$$

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 $\boldsymbol{\Sigma}$을 일반적으로 공분산(covariance)라고 부른다.**

12 Multivariate gaussian distribution

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

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{17}$$

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.

$$\boxed{p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)} \tag{18}$$

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

13 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.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{19}$$

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.

$$\boxed{\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} + \mathbf{b} \\ &\sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\intercal) \end{aligned}} \tag{20}$$

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

$$\begin{aligned} \mathrm{cov}(\mathbf{A}\mathbf{x}+\mathbf{b}) &= \mathbb{E}((\mathbf{y}-\boldsymbol{\mu}_\mathbf{y})(\mathbf{y}-\boldsymbol{\mu}_\mathbf{y})^\intercal) \\ &= \mathbb{E}((\mathbf{y}-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))(\mathbf{y}-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))^\intercal) \\ &= \mathbb{E}(((\mathbf{A}\mathbf{x}+\mathbf{b})-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))((\mathbf{A}\mathbf{x}+\mathbf{b})-(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}))^\intercal) \\ &= \mathbb{E}([\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})]\,[\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})]^\intercal) \\ &= \mathbb{E}(\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathbf{A}^\intercal) \\ &= \mathbf{A}\mathbb{E}((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal)\mathbf{A}^\intercal \\ &= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\intercal \end{aligned} \tag{21}$$

14 Conditional probability

조건부 확률은 두 사건 x, y가 주어졌을 때, y가 발생했을 때 x가 발생할 확률을 의미한다. 이는 다음과 같이 나타낼 수 있다.

$$P(x|y) = \frac{P(x \cap y)}{P(y)} \tag{22}$$

이를 통해 두 사건이 동시에 발생한 확률은 $P(x \cap y) = P(x)P(y|x)$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 확률을 의미한다.

$$P(y|x) = \frac{P(y \cap x)}{P(x)} \tag{23}$$

만약 두 사건 x, y가 독립사건이면 조건부 확률은 다음과 같다.

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

위 식들은 이산확률변수에 대한 조건부 확률을 의미한다. 연속확률분포에 대한 조건부 확률은 다음과 같다. 연속확률변수가 주어졌을 때, y가 발생했을 때 x가 발생할 확률은 다음과 같다.

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

- $p(x, y)$: 사건 x와 y가 동시에 발생하는 결합 확률밀도함수(joint pdf)

반대로 x가 발생했을 때 y가 발생할 확률은 다음과 같다.

$$p(y|x) = \frac{p(y, x)}{p(x)} \tag{26}$$

이 때, $p(x, y)$와 $p(y, x)$는 동일하다.

15 Marginal probability

주변 확률분포(marginal probability distribution)는 다변수 확률분포에서 한 확률변수의 행동을 단독으로 이해하고자 할 때 주로 사용된다. 예를 들어, 두 이산확률변수 x와 y가 있고, 이들의 결합 확률분포(joint probability distribution)가 주어졌을 때, x의 주변 확률분포는 y가 취할 수 있는 모든 값에 대해 x의 확률을 합하여 얻어진다.

$$P(x) = \sum_{y} P(x, y) \tag{27}$$

연속확률변수로 나타내면 다음과 같다.

$$p(x) = \int p(x, y)dy = \int p(x|y)p(y)dy \tag{28}$$

16 Bayesian rule

Bayesian rule은 다음과 같은 조건부확률 간 관계를 의미한다.

$$\begin{aligned} p(x|y) &= \frac{p(x, y)}{p(y)} \\ &= \frac{p(y|x)p(x)}{p(y)} \end{aligned} \tag{29}$$

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

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.

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z})} = \eta \cdot p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \tag{30}$$

- $p(\mathbf{x}|\mathbf{z})$: 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.

17 Conditional gaussian distribution

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

$$\begin{aligned} p(x|y) &= \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} = \eta \cdot p(y|x)p(x) \\ &\sim \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \end{aligned} \tag{31}$$

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

$$\begin{aligned} \boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(y - \boldsymbol{\mu}_y) \\ \boldsymbol{\Sigma}_{x|y} &= \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}^{\intercal} \end{aligned} \tag{32}$$

18 Exponential family

지수족(exponential family)이란 지수항을 가지는 다양한 분포들의 집합을 의미한다. 지수족에는 Gaussian, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, binomial, multinomial, geometric 분포 등 다양한 분포들이 포함된다. 언급한 분포들을 일반화하여 표현하면 다음과 같다.

$$\boxed{p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\intercal}\mathbf{u}(\mathbf{x})\}} \tag{33}$$

- $\mathbf{x}$: 확률 변수(random variable)
- $h(\mathbf{x})$: $\mathbf{x}$에 대한 임의의 함수
- $\boldsymbol{\eta}$: 자연 파라미터(nature parameters)
- $g(\boldsymbol{\eta})$: 확률의 정의 상 크기를 1로 만들어주는 정규화(normalization) 값
- $\mathbf{u}(\mathbf{x})$: 충분통계량(sufficient statistic)

위 식은 pdf이기 때문에 확률의 정의를 만족한다.

$$\begin{aligned} &\int p(\mathbf{x}|\boldsymbol{\eta}) = 1 \\ &\rightarrow \int h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\intercal}\mathbf{u}(\mathbf{x})\} = 1 \\ &\rightarrow g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^{\intercal}\mathbf{u}(\mathbf{x})\}d\mathbf{x} = 1 \end{aligned} \tag{34}$$

위 식에서 보다시피 지수족은 완비충분통계량(complete sufficient statistic) u(x)를 포함하고 있기 때문에 앞서 설명한 분포들을 지수족의 형태로 인수 분해하면 해당 분포에 대한 완비충분통계량을 쉽게 구할 수 있다. 완비충분통계량을 사용하면 최소분산불편추정값(minimum variance unbiased estimator, MVUE)를 쉽게 구할 수 있으므로 이 때 지수족이 유용하게 사용된다.

18.1 Exponential Family 1 - Bernoulli distribution

베르누이 분포가 지수족에 포함되는지 알아보도록 하자. 베르누이 분포는 다음과 같이 표현할 수 있다.

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{35}$$

위 분포를 (33)와 동일한 형태로 표현할 수 있을까? 양 변에 $\ln, \exp$를 동시에 취해주고 식을 변환하면 다음과 같다.

$$\begin{aligned} p(x|\mu) &= \exp\{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned} \tag{36}$$

위 식에서 $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$이고 η와 μ의 관계를 바꿔서 역함수로 나타내면 $\mu = \sigma(\eta)$ 함수 형태가 된다.

$$\sigma(\eta) = \frac{1}{1+\exp(-\eta)} \tag{37}$$

위 식을 로지스틱 회귀(logistic regression)식이라고 부른다. 따라서 베르누이 분포로부터 식을 적절하게 변형하면 로지스틱 회귀식이 나오는 것을 알 수 있다. (37)는 $1-\sigma(\eta) = \sigma(-\eta)$ 관계를 만족하므로 이를 통해 다음과 같은 지수족 파라미터를 얻을 수 있다.

$$\boxed{\begin{array}{l} p(x|\mu) = \sigma(-\eta)\exp(\eta x) \quad \cdots \text{Bernoulli} \\ \text{where,} \\ \quad u(x) = x \\ \quad h(x) = 1 \\ \quad g(\eta) = \sigma(-\eta) \end{array}} \tag{38}$$

18.2 Exponential Family 2 - Gaussian distribution

다음으로 가우시안 분포가 지수족에 속하는지 알아보자.

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \end{aligned} \tag{39}$$

가우시안 분포는 베르누이 분포와 달리 자체적으로 exponential 항을 포함하고 있는 것을 알 수 있다. 따라서 별도의 유도 과정 없이 바로 지수족의 파라미터를 구할 수 있다.

$$\boxed{\begin{array}{l} p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \quad \cdots \text{Gaussian} \\ \text{where,} \\ \quad \boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \\ \quad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2}\exp(\frac{\eta_1^2}{4\eta_2}) \\ \quad \mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ \quad h(x) = (2\pi)^{-1/2} \end{array}} \tag{40}$$

18.3 Maximum Likelihood Estimator and Sufficient Statistics

지수족에서 자연 파라미터 $\boldsymbol{\eta}$를 추정하는 문제를 살펴보자. 일반적으로 MLE를 사용하여 $\boldsymbol{\eta}$를 추정한다. MLE를 찾기 위해 (34)를 미분하면 다음과 같다.

$$\begin{aligned} &\frac{d}{d\eta}\left(g(\eta)\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\}d\mathbf{x} = 1\right) \\ &\rightarrow \nabla g(\eta)\int h(\mathbf{x})\exp\{\eta^T\mathbf{u}(\mathbf{x})\}d\mathbf{x} + g(\eta)\int h(\mathbf{x})\exp\{\eta^T\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})d\mathbf{x} = 0 \end{aligned} \tag{41}$$

$\int h(\mathbf{x})\exp\{\eta^T\mathbf{u}(\mathbf{x})\}d\mathbf{x} = \frac{1}{g(\eta)}$를 이용하여 위 식을 정리하면 다음과 같다.

$$-\frac{1}{g(\eta)}\nabla g(\eta) = g(\eta)\int h(\mathbf{x})\exp\{\eta^T\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})d\mathbf{x} = \mathbb{E}(\mathbf{u}(\mathbf{x})) \tag{42}$$

좌항 $-\frac{1}{g(\eta)}\nabla g(\eta)$은 $-\nabla \ln g(\eta)$의 미분값이므로 이를 정리하면 다음과 같다.

$$-\nabla \ln g(\eta) = g(\eta)\int h(\mathbf{x})\exp\{\eta^T\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})d\mathbf{x} = \mathbb{E}(\mathbf{u}(\mathbf{x})) \tag{43}$$

$$\boxed{-\nabla \ln g(\eta) = \mathbb{E}(\mathbf{u}(\mathbf{x}))} \tag{44}$$

다음으로 여러 관측 데이터 $\mathbf{x} = [x[0], x[1], \cdots, [N-1]]^\intercal$이 주어진 경우를 생각해보자. 각 $x[n]$들은 서로 독립이며 동일한 확률 분포를 따른다(=i.i.d)고 하자. Likelihood를 보면 다음과 같다.

$$p(\mathbf{x}|\eta) = \left(\prod_{n=1}^{N} h(x[n]) \right) g(\eta)^N \exp\left\{ \eta^T \sum_{n=1}^{N} \mathbf{u}(x[n]) \right\} \tag{45}$$

앞선 경우와 동일하게 미분 후 0이 되는 η 값을 찾으면 이는 곧 MLE가 된다.

$$\boxed{-\nabla \ln g(\eta_{\mathrm{ML}}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(x[n])} \tag{46}$$

위 식에서 $\sum_{n=1}^{N} \mathbf{u}(x[n])$는 충분통계량이다. 만약 데이터가 충분한 경우($N \to \infty$), 우측 항은 큰 수의 법칙(law of large numbers)에 의해 $\mathbb{E}(\mathbf{u}(\mathbf{x}))$가 되고 $\eta_{\mathrm{ML}} \to \eta$가 된다.

19 Various distributions

19.1 Discrete r.v.'s distribution

19.1.1 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 : {success, fail} $\to$ $\{0, 1\}$이다.

$$\begin{aligned} P(x=0) &= 1-p \\ P(x=1) &= p \end{aligned} \tag{47}$$

확률질량함수(pmf) $P(x)$는 다음과 같다.

$$\begin{aligned} &x \sim \mathrm{Ber}(p) \\ &P(x) = p^x (1-p)^{1-x} \quad x = 0, 1 \end{aligned} \tag{48}$$

19.2 Continuous r.v.'s distribution

19.2.1 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.

$$\boxed{y = \sum_{i=1}^{n} x_i^2 \sim \mathcal{X}_n^2} \tag{49}$$

- $x_i \sim \mathcal{N}(0,1),\ i = 1, 2, \cdots, n$: 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.

$$p(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} \exp(-\frac{1}{2}y) & y \geq 0 \\ 0 & y < 0 \end{cases} \tag{50}$$

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

y의 평균과 분산은 다음과 같다.

$$\begin{aligned} \mathbb{E}(y) &= n \\ \mathrm{var}(y) &= 2n \end{aligned} \tag{51}$$

20 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

21 Revision log

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