Notes on Probability Theory

Gyubeom Edward Im*

February 9, 2024

Contents

1 Probability theory

본 포스트는 필자가 SLAM을 위한 확률 이론을 공부하면서 중요하다고 생각한 부분을 정리한 내용이다.

선형대수학에 대해 알고 싶으면 선형대수학 (Linear Algebra) 개념 정리 포스트를 참조하면 된다. 선형 시스템 및 다양한 행렬 분해 방법에 대해 알고 싶으면 선형시스템 행렬 분해 (Linear System Matrix Decomposition) 개념 정리 포스트를 참조하면 된다. Kalman filter (KF, EKF, ESKF)에 대해 알고 싶으면 칼만 필터(Kalman Filter) 개념 정리 포스트를 참조하면 된다. Particle filter에 대해 알고 싶으면 파티클 필터(Particle Filter) 개념 정리 포스트을 참조하면 된다.

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

2 Sample space and event

어떤 시행에서 일어날 수 있는 모든 결과들의 모임을 표본공간 Ω라고 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.

3 Probability function

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

$$\forall \mathcal{F} \in \Omega, \quad p(\mathcal{F}) = \sum_{w \in \mathcal{F}} p(w) \tag{1}$$

4 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.

5 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.

$$0 \leq p(x) \leq 1 \quad \forall x \in \mathcal{F} \tag{2}$$

표본공간 Ω 전체가 일어날 확률은 1이다.

$$p(\Omega) = 1 \tag{3}$$

6 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.

6.1 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.

6.2 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.

7 Expected value

기대값(expected value) $\mathbb{E}[\cdot]$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.

$$\mathbb{E}[x] = \sum_{x \in \Omega} x \cdot P(x) \tag{4}$$

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

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x \cdot p(x)dx \tag{5}$$

7.1 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.

$$\mathbb{E}[ax + by] = a\mathbb{E}[x] + b\mathbb{E}[y] \tag{6}$$

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

$$\mathbb{E}[L(x)] = L(\mathbb{E}[x]) \tag{7}$$

8 Variance and standard deviation

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

$$\text{var}[x] = \mathbb{E}[(x - \mathbb{E}(x))^2] \tag{8}$$

또한 아래와 같이 표현할 수도 있다.

$$\text{var}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \tag{9}$$

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

9 Gaussian distribution

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

$$x \sim \mathcal{N}(\mu, \sigma^2) \tag{10}$$

- $\sim \mathcal{N}(\cdot, \cdot)$: 확률변수가 가우시안 분포(또는 정규 분포)를 따른다는 의미
- μ: x의 평균
- σ^2: x의 분산

가 성립한다. 이 때, 확률분포함수 $p(x)$는 다음과 같이 정의된다.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-1}{2\sigma^2}(\mathbf{x} - \mu)^2) \tag{11}$$

10 Multivariate gaussian distribution

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

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{12}$$

가 성립한다. 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.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \tag{13}$$

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

11 Joint gaussian distribution

확률변수가 두 개 이상일 때는 다변수 확률분포(multivariate probability distribution)를 사용해야한다. 예를 들어 두 개의 확률변수 x, y가 있을 때 다변수 확률분포$p(x, y) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$는 다음과 같이 나타낼 수 있다.

$$p(x,y) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} \exp\Bigg(-\frac{1}{2}\bigg(\begin{bmatrix} x-\mathbb{E}(x) \\ y-\mathbb{E}(y) \end{bmatrix}\bigg)^T \mathbf{\Sigma}^{-1} \bigg(\begin{bmatrix} x-\mathbb{E}(x) \\ y-\mathbb{E}(y) \end{bmatrix}\bigg)\Bigg) \tag{14}$$

이 때 평균은 $\boldsymbol{\mu} = \begin{bmatrix} \mathbb{E}(x) \\ \mathbb{E}(y) \end{bmatrix}$이고 분산은 $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy} \end{bmatrix}$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 $\mathbf{\Sigma}$을 일반적으로 공분산(covariance)라고 부른다.

12 Linear transformation of gaussian random variable

스칼라 랜덤 변수(random variable) x가 주어졌을 때 만약 가우시안 분포를 따른다고 가정하면 다음과 같이 표기할 수 있다.

$$x \sim \mathcal{N}(\mu, \sigma^2) \tag{15}$$

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

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) \tag{16}$$

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

$$\boxed{\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} + \mathbf{b} \\ &\sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\intercal) \end{aligned}} \tag{17}$$

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

$$\begin{aligned} \text{cov}(\mathbf{A}\mathbf{x} + \mathbf{b}) &= \mathbb{E}((\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})^\intercal) \\ &= \mathbb{E}((\mathbf{y} - (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}))(\mathbf{y} - (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}))^\intercal) \\ &= \mathbb{E}(((\mathbf{A}\mathbf{x} + \mathbf{b}) - (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}))((\mathbf{A}\mathbf{x} + \mathbf{b}) - (\mathbf{A}\boldsymbol{\mu} + \mathbf{b}))^\intercal) \\ &= \mathbb{E}([\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})]\,[\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})]^\intercal) \\ &= \mathbb{E}(\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\intercal\mathbf{A}^\intercal) \\ &= \mathbf{A}\mathbb{E}((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\intercal)\mathbf{A}^\intercal \\ &= \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\intercal \end{aligned} \tag{18}$$

13 Conditional probability

조건부 확률은 두 사건 X,Y가 주어졌을 때, Y가 발생했을 때 X가 발생할 확률을 의미한다. 이는 다음과 같이 나타낼 수 있다.

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \tag{19}$$

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.

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)} \tag{20}$$

만약 두 사건 X,Y가 독립사건이면 조건부 확률은 다음과 같다.

$$P(X|Y) = P(X)P(Y) \tag{21}$$

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

$$p(x|y) = \frac{p(x,y)}{p(y)} \tag{22}$$

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

반대로 x가 발생했을 때 y가 발생할 확률은 다음과 같다.

$$p(y|x) = \frac{p(y,x)}{p(x)} \tag{23}$$

이 때, $p(x,y)$와 $p(y,x)$는 동일하다.

14 Bayesian rule

Bayesian rule은 다음과 같은 조건부확률 간 관계를 의미한다.

$$\begin{aligned} p(x|y) &= \frac{p(x,y)}{p(y)} \\ &= \frac{p(y|x)p(x)}{p(y)} \end{aligned} \tag{24}$$

- $p(x|y)$: 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.

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z})} = \eta \cdot p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \tag{25}$$

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

15 Conditional gaussian distribution

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

$$\begin{aligned} p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} &= \eta \cdot p(y|x)p(x) \\ \sim \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \end{aligned} \tag{26}$$

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

$$\begin{aligned} \boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(y - \boldsymbol{\mu}_y) \\ \boldsymbol{\Sigma}_{x|y} &= \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}^{\intercal} \end{aligned} \tag{27}$$

16 References

[1] 평균과 기댓값 - infograph blog

17 Revision log

- 1st: 2024-02-09