Notes on Probability Theory

Gyubeom Edward Im*

February 24, 2024

Contents

1 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.

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

2 Probability function

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

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

3 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.

4 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.

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

5 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.

5.1 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.

5.2 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.

6 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.

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

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

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

6.1 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.

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

7 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.

8 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.

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

9 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.

10 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$p(x,y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$는 다음과 같이 나타낼 수 있다.

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

이 때 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.

11 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.

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

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

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

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

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

12 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.

13 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.

14 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.

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

15 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.

$$\boxed{p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\intercal}\mathbf{u}(\mathbf{x})\}} \tag{28}$$

- $\mathbf{x}$: 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.

$$\begin{aligned} & \int p(\mathbf{x}|\boldsymbol{\eta}) = 1 \\ & \rightarrow \int h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\} = 1 \\ & \rightarrow g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\} d\mathbf{x} = 1 \end{aligned} \tag{29}$$

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.

15.1 Exponential Family 1 - Bernoulli distribution

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

$$p(x|\mu) = \text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \tag{30}$$

위 분포를 (28)와 동일한 형태로 표현할 수 있을까? 양 변에 $\ln, \exp$를 동시에 취해주고 식을 변환하면 다음과 같다.

$$\begin{aligned} p(x|\mu) &= \exp\{x \ln \mu + (1-x) \ln(1-\mu)\} \\ &= (1-\mu) \exp\left\{ \ln\left(\frac{\mu}{1-\mu}\right) x \right\} \end{aligned} \tag{31}$$

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

$$\sigma(\eta) = \frac{1}{1+\exp(-\eta)} \tag{32}$$

위 식을 로지스틱 회귀(logistic regression)식이라고 부른다. 따라서 베르누이 분포로부터 식을 적절하게 변형하면 로지스틱 회귀식이 나오는 것을 알 수 있다. (32)는 $1-\sigma(\eta) = \sigma(-\eta)$ 관계를 만족하므로 이를 통해 다음과 같은 지수족 파라미터를 얻을 수 있다.

$$\boxed{\begin{aligned} & p(x|\mu) = \sigma(-\eta)\exp(\eta x) \quad \cdots \text{Bernoulli} \\ & \text{where,} \\ & \quad u(x) = x \\ & \quad h(x) = 1 \\ & \quad g(\eta) = \sigma(-\eta) \end{aligned}} \tag{33}$$

15.2 Exponential Family 2 - Gaussian distribution

다음으로 가우시안 분포가 지수족에 속하는지 알아보자.

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2 \right\} \end{aligned} \tag{34}$$

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.

$$\boxed{\begin{aligned} & p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \quad \cdots \text{Gaussian} \\ & \text{where,} \\ & \qquad \boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \\ & \qquad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp(\frac{\eta_1^2}{4\eta_2}) \\ & \qquad \mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ & \qquad h(x) = (2\pi)^{-1/2} \end{aligned}} \tag{35}$$

15.3 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 (29)를 미분하면 다음과 같다.

$$\begin{aligned} & \frac{d}{d\eta}\left(g(\eta)\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\}d\mathbf{x} = 1\right) \\ & \rightarrow \nabla g(\eta)\int h(\mathbf{x})\exp\{\eta^T \mathbf{u}(\mathbf{x})\}d\mathbf{x} + g(\eta)\int h(\mathbf{x})\exp\{\eta^T \mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})d\mathbf{x} = 0 \end{aligned} \tag{36}$$

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

$$-\frac{1}{g(\eta)}\nabla g(\eta) = g(\eta)\int h(\mathbf{x})\exp\{\eta^T \mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})d\mathbf{x} = \mathbb{E}(\mathbf{u}(\mathbf{x})) \tag{37}$$

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

$$-\nabla \ln g(\eta) = g(\eta)\int h(\mathbf{x})\exp\{\eta^T \mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})d\mathbf{x} = \mathbb{E}(\mathbf{u}(\mathbf{x})) \tag{38}$$

$$\boxed{-\nabla \ln g(\eta) = \mathbb{E}(\mathbf{u}(\mathbf{x}))} \tag{39}$$

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.

$$p(\mathbf{x}|\eta) = \left(\prod_{n=1}^{N} h(x[n])\right) g(\eta)^N \exp\left\{\eta^T \sum_{n=1}^{N} \mathbf{u}(x[n])\right\} \tag{40}$$

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

$$\boxed{-\nabla \ln g(\eta_{\text{ML}}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{u}(x[n])} \tag{41}$$

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

16 References

[1] (Blog) 평균과 기댓값

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

[3] (Blog) [수리통계학] 38. 지수족

17 Revision log

- 1st: 2024-02-09
- 2nd: 2024-02-24