

9. 2 치대가능도 추정법

모멘트 방법으로 추정한 모수는 그 숫자가 가장 가능성 높은 값이라는 이론적 보장이 없다.
이 절에서는 이론적으로 가장 가능성이 높은 모수를 찾는 방법인 치대가능도 추정법을
알아본다.

— 모든 추정방법 중 가장 널리 사용되는 방법!

가능도함수

여러 확률분포 X 에 대한 확률밀도함수 또는 확률질량함수를
다음과 같이 대표하여 쓰기로 한다.

$$p(x; \theta)$$

↓

확률분포가
가지두는 실수값
(스칼라와 벡터)

확률분포 높수 대응자로
(스칼라와 벡터)

$$\begin{cases} \text{Ex - 베르누이 확률분포} & \theta = \mu \\ \text{Ex - 이항분포} & \theta = (N, \mu) \\ \text{Ex - 정규분포} & \theta = (\mu, \sigma^2) \end{cases}$$

for 확률밀도함수) θ 를 알고 있고 x 가 변수
for 가능도함수) x 를 알고 있고 θ 가 변수
확률밀도함수에서 모수를 변수로 보면 경우, 이 함수를 가능도함수(likelihood
function)이라고 한다.

$$L(\theta; x) = p(x; \theta)$$

가능도함수로 표현 확률밀도함수로 표현

예제

- 정규분포

정규분포의 확률밀도함수는 다음과 같은 단변수함수다.

$$p(x; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)$$

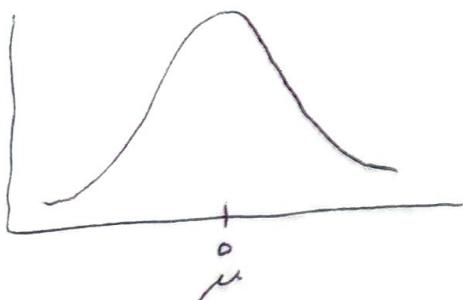
모수가 삼수라는 것을 강조하기 위해 아래첨자를 붙였다.

이제 가능도함수는 다음과 같이 입력변수가 2개인 다변수함수가 된다.

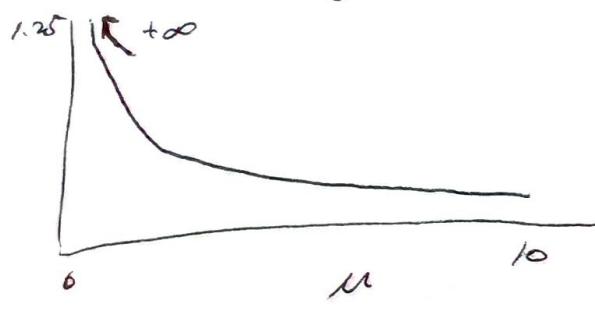
$$L(\mu, \sigma^2; x_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_0-\mu)^2}{2\sigma^2}\right)$$

* 수식은 같지만 함수의 변수가 다른 것에 주의

(정규분포 $p(x; \mu_0, \sigma_0^2)$ pdf)



(가능도함수 $L(\mu, \sigma^2; x_0)$)



예제

- 베르누이분포

베르누이분포의 확률밀도함수 (입력기는 0 또는 1만 가능)

$$p(x; \mu_0) = \mu_0^x (1-\mu_0)^{1-x}$$

베르누이분포의 가능도함수 (0부터 1까지의 연속적인 실수값을 입력받음)

$$L(\mu; x_0) = \mu^{x_0} (1-\mu)^{1-x_0}$$

가능도함수를 수식으로 나타내면, 수식 자체는 확률밀도함수의 수식과 같다. 하지만 가능도함수는 확률분포함수가 아니라는 점에 주목해야 한다. 확률밀도함수는 가능한 모든 표본값 x 에 대해 적분하면 전체면적이 1이 되지만,

$$\int_{-\infty}^{\infty} p(x; \theta) dx = 1$$

가능도함수는 가능한 모든 모수값 θ 에 대해 적분했을 때 1이 되는 보장이 없다.

$$\int_{-\infty}^{\infty} L(\theta; x) d\theta = \int_{-\infty}^{\infty} p(x; \theta) d\theta \neq 1$$

확률밀도함수	가능도함수
θ 값은 이미 알고 있음	x 가 이미 발생했으므로 x 를 알고 있음
θ 상수, x 변수	x 상수, θ 변수
θ 가 정해진 상황에서의 x 의 상대적 확률	x 가 정해진 상황에서의 θ 의 상대적 확률
적분하면 전체면적 = 1	적분하면 전체면적이 1이 아닐 수도 있음

최대가능도 추정법

(Maximum Likelihood Estimation, MLE)

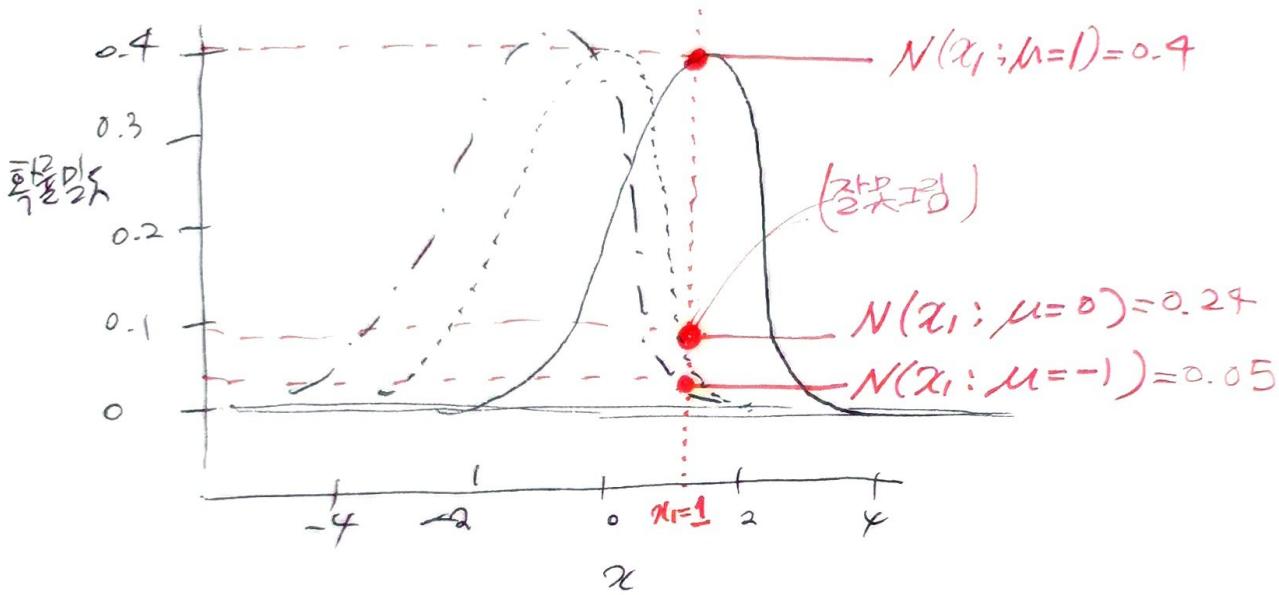
- 주어진 표본에 대해 가능도를 가장 크게 하는 모수 θ 를 찾는 방법.
- 이 방법으로 찾은 모수는 기호로 $\hat{\theta}_{MLE}$ 와 같이 표시한다.

$$\hat{\theta}_{MLE} = \arg \max_{\theta} L(\theta; x)$$

(예제)

"정규분포를 가지는 확률변수의 분산 $\sigma^2=1$ 은 알고 있으나, 평균 μ 를 모르고 있다. 이를 추정해야 하는 문제!"

- 갖고 있는 확률변수 표본 $x_1 = 1$
- $\mu = -1, 0, 1$ 중에 어느 평균부터 $x_1 = 1$ 이 나올 확률이 가장 클까?



(복수의 표본 데이터가 있는 경우의 가능도함수)

추정을 위해 확보한 확률변수 표본의 수가 복수개

$$\{x_1, x_2, \dots, x_N\}$$

가능도함수는 복수 표본 값에 대한 결합 확률밀도 $P_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \theta)$ 가 된다.

표본 데이터 x_1, x_2, \dots, x_N 은 같은 확률밀도에서 나온 독립적인 값들인므로,

결합 확률밀도함수는 다음과 같이 곱으로 표현된다.

$$L(\theta; x_1, x_2, \dots, x_N) = p(x_1, x_2, \dots, x_N; \theta) = \prod_{i=1}^N p(x_i; \theta)$$

$$\leftarrow \prod_{i=1}^N p(x_i; \theta)$$

(상수) 복수

(예제) - 정규분포

정규분포로부터 다음 세 개의 표본 데이터를 얻었다.

$$\begin{matrix} \{1, 0, -3\} \\ x_1 \quad x_2 \quad x_3 \end{matrix}$$

이 경우의 가능도함수는 다음과 같다.

$$L(\theta; x_1, x_2, x_3)$$

$$= N(x_1, x_2, x_3; \theta)$$

$$= N(\underline{x}_1; \theta) \cdot N(\underline{x}_2; \theta) \cdot N(\underline{x}_3; \theta)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(1-\mu)^2}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(0-\mu)^2}{2\sigma^2}\right) \cdot$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(-3-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^3} \exp\left(-\frac{\mu^2 + (1-\mu)^2 + (-3-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^3} \exp\left(-\frac{3\mu^2 + 4\mu + 10}{2\sigma^2}\right)$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^3} \exp\left(-\frac{3(\mu + \frac{2}{3})^2 + \frac{26}{3}}{2\sigma^2}\right)$$

가능도함수는 2차 항수이므로, 미분하지 않아도 최대값의 위치를 알 수 있다.

가장 가능도를 높게 하는 모든 μ 의 값은 $\hat{\mu}_{MLE} = -\frac{2}{3}$ 이다.

예제

- 배운 내용으로부터 다음 표본 데이터를 얻었다고 하자.

$$\{1, 0, 1\}$$

이때 가능도함수는 다음과 같다.

$$\begin{aligned}
 L(\mu; x_1=1, x_2=0, x_3=1) \\
 &= p(x_1=1, x_2=0, x_3=1; \mu) \\
 &= p(x=1; \mu) p(x=0; \mu) p(x=1; \mu) \\
 &= \mu^1 (1-\mu)^{1-1} \cdot \mu^0 (1-\mu)^{1-0} \cdot \mu^1 (1-\mu)^{1-1} \\
 &= \mu \cdot (1-\mu) \cdot \mu \\
 &= -\mu^3 + \mu^2
 \end{aligned}$$

이 가능도함수를 최대화하는 모수의 값을 찾기 위해,
미분한 도함수가 0이 되는 위치를 찾는다.

$$\frac{dL}{d\mu} = -3\mu^2 + 2\mu = -3\mu \left(\mu - \frac{2}{3}\right) = 0$$

모수의 값이 0이면 표본값으로 1이 나올 수 없으므로,
가능도함수를 최대화하는 모수는 $\hat{\mu}_{MLE} = \frac{2}{3}$ 이다.

로그 가능도 함수

일반적으로 최대가능도 추정법을 사용하여 가능도가 최대가 되는 θ 를 계산하는데
수치적 최적화 (numerical optimization)를 해야 한다.

$$\hat{\theta}_{ML} = \arg \max_{\theta} L(\theta; \{x_i\})$$

그런데 보통은 가능도를 직접 사용하는 것이 아니라,

로그 변환한 로그 가능도 함수 $L = \log L(\theta)$ 을 사용하는 경우가 많다.

$$\hat{\theta}_{ML} = \arg \max_{\theta} \log L(\theta; \{x_i\})$$

이유는 다음과 같다.

1. 로그 변환에 의해서는 최대값의 위치가 변지 않는다.
2. 반복시행으로 인한 복수 표본 데이터인 경우, 결합 확률밀도함수가 동일한 항수의 곱으로 나타나는 경우가 많는데, 이때 로그 변환에 의해 곱셈이 덧셈이 되어 계산이 단순해진다.

예제

위 예제 같이 정규분포로부터 얻은 표본값이 다음과 같은 경우

$$\{1, 0, -3\}$$

로그 변환을 하면 최대값의 위치가 $-2/3$ 이라는 것을 쉽게 구할 수 있다.

$$\begin{aligned} & \log L(\mu; x_1, x_2, x_3) \\ &= \log \left(\frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} \exp \left(-\frac{3\mu^2 + 4\mu + 10}{2\sigma^2} \right) \right) \\ &= \log \left(\frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} \right) - \frac{3\mu^2 + 4\mu + 10}{2\sigma^2} \\ &= \log \left(\frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} \right) - \frac{3(\mu + \frac{2}{3})^2 + \frac{26}{3}}{2\sigma^2} \end{aligned}$$

연습문제 9.2.1

베르누이 분포로부터 다음과 같은 표본을 얻었다. 이 확률변수의 평균 μ 를 최대가능도 추정법을 사용하여 구하라.

$$\{1, 0, 1, 1\} \quad \text{B(1,1)방법} \rightarrow \frac{1+0+1+1}{4+0+1+1} = \frac{3}{7}$$

$$\begin{aligned} & \log L(\mu; x_1=1, x_2=0, x_3=1, x_4=1) \\ &= \log p(x_1=1, x_2=0, x_3=1, x_4=1; \mu) \\ &= \log (p(x=1; \mu)p(x=0; \mu)p(x=1; \mu)p(x=1; \mu)) \\ &= \log (\mu \cdot (1-\mu) \cdot \mu \cdot \mu) \\ &= \log (\mu^3 (1-\mu)) \end{aligned}$$

$$\frac{d \log L}{d\mu} = \frac{3\mu^2 - 4\mu^3}{\mu^3 (1-\mu)} = 0$$

$$\mu = \frac{4}{3}$$

연습문제 9.2.2

$K=4$ 인 카테고리분포로부터 다음과 같은 표본을 얻었다.

- i) 확률변수의 모수 μ 를 최대기능도 추정법을 사용하여 구하라.

$$\{1, 4, 1, 2, 4, 2, 3, 4\}$$

(사망률 주사위 단지기)



$$0 \leq \mu_i \leq 1$$

$$\sum \mu_i = 1$$

$$\mu = (\mu_1, \mu_2, \mu_3, \mu_4) = \left(\frac{2}{8}, \frac{2}{8}, \frac{1}{8}, \frac{3}{8}\right) \text{ by moment method}$$

<최대기능도 추정법>

$K=4$ 인 카테고리분포의 확률질량함수

$$P(x; \mu) = \mu_1^{x_1} \mu_2^{x_2} \mu_3^{x_3} \mu_4^{x_4}$$

따라서 전체 데이터에 대한 32가능도 합수는

$$\begin{aligned} & \log L(\mu; x_1=1, x_2=4, x_3=1, x_4=2, x_5=4, x_6=2, x_7=3, x_8=4) \\ &= \log P(x_1=1, x_2=4, x_3=1, x_4=2, x_5=4, x_6=2, x_7=3, x_8=4) \\ &= \log (p(x=1; \mu)p(x=4; \mu)p(x=1; \mu)p(x=2; \mu)p(x=4; \mu)p(x=2; \mu) \\ &\quad p(x=3; \mu)p(x=4; \mu)) \end{aligned}$$

$$= \log \mu_1 \mu_4 \mu_1 \mu_2 \mu_4 \mu_2 \mu_3 \mu_4$$

$$= \log \mu_1^2 \mu_2^2 \mu_3^1 \mu_4^3$$

$$= 2\log \mu_1 + 2\log \mu_2 + \log \mu_3 + 3\log \mu_4$$

$$\text{또가 만족해야 하는 제한조건: } \mu_1 + \mu_2 + \mu_3 + \mu_4 = 1$$

제한조건을 만족하면서 32가능도 합수를 최대화하는 모든의 값을 찾기 위한 새 목적함수 J

$$J = 2\log \mu_1 + 2\log \mu_2 + \log \mu_3 + 3\log \mu_4 + \lambda(\mu_1 + \mu_2 + \mu_3 + \mu_4 - 1)$$

목적함수 J 를 모수와 대량주승수로 미분하여
gradient = 0인 값을 찾는다.

$$\frac{\partial J}{\partial \mu_1} = \frac{2}{\mu_1} + \lambda = 0 \quad \dots \quad ①$$

$$\frac{\partial J}{\partial \mu_2} = \frac{2}{\mu_2} + \lambda = 0 \quad \dots \quad ②$$

$$\frac{\partial J}{\partial \mu_3} = \frac{1}{\mu_3} + \lambda = 0 \quad \dots \quad ③$$

$$\frac{\partial J}{\partial \mu_4} = \frac{3}{\mu_4} + \lambda = 0 \quad \dots \quad ④$$

$$\frac{\partial J}{\partial \lambda} = \mu_1 + \mu_2 + \mu_3 + \mu_4 - 1 = 0 \quad \dots \quad ⑤$$

$\mu_1, \mu_2, \mu_3, \mu_4, \lambda$ 에 대한 5개의 시립방정식을 풀면

$$\mu_1 = \frac{1}{4}, \mu_2 = \frac{1}{4}, \mu_3 = \frac{1}{8}, \mu_4 = \frac{3}{8}$$

(베르누이 분포의) 전개가능도 모수 추정

보통가지로 베르누이 분포의 확률밀량함수는 다음과 같다.

$$p(x; \mu) = \text{Bern}(x; \mu) = \mu^x (1-\mu)^{1-x}$$

그런데 N 번의 반복시행으로 표본데이터가 x_1, \dots, x_N 이 있는 경우에는 모두 독립이므로 전체 확률밀량함수는 각각의 확률밀량함수의 곱과 같다.

$$L(\mu; x_1, \dots, x_N) = p(x_1, \dots, x_N; \mu) = \prod_{i=1}^N \mu^{x_i} (1-\mu)^{1-x_i}$$

미분을 쉽게 하기 위해 로그변환을 하여 로그가능도를 구하면 다음과 같다.

$$\begin{aligned} \log L &= \log p(x_1, \dots, x_N; \mu) \\ &= \sum_{i=1}^N \{x_i \log \mu + (1-x_i) \log (1-\mu)\} \\ &= \sum_{i=1}^N x_i \log \mu + (N - \sum_{i=1}^N x_i) \log (1-\mu) \end{aligned}$$

$x=1$ (성공) 또는 $x=0$ (실패) 이므로 성공회수와 실패회수를 다음과 같이 N_1, N_0 이라고 표기하도록 하자.

$$N_1 = \sum_{i=1}^N x_i, \quad N_0 = N - \sum_{i=1}^N x_i$$

로그가능도는 다음과 같아진다.

$$\log L = N_1 \log \mu + N_0 \log (1-\mu)$$

i) 목적함수를 모수로 미분한 값이 0이 되게 하는 보수값을 구하면 다음과 같다.

$$\frac{\partial \log L}{\partial \mu} = \frac{\partial}{\partial \mu} \{ N_1 \log \mu + N_0 \log (1-\mu) \} = 0$$
$$= \frac{N_1}{\mu} - \frac{N_0}{1-\mu} = 0 \quad \text{chain rule}$$

$$\therefore \frac{N_1}{\mu} = \frac{N_0}{1-\mu}$$

$$\frac{1-\mu}{\mu} = \frac{N_0}{N_1} = \frac{N-N_1}{N_1}$$

$$\frac{1}{\mu} - 1 = \frac{N}{N_1} - 1$$

$$\mu = \frac{N_1}{N} = \begin{matrix} \text{이 나올 확률} \\ \text{moment method} \\ \text{estimation} \end{matrix}$$

전체사행 확률

모멘트 방법과 동일한 결과,
but 다음 논리적인 문제가 있음.

카테고리 분포의 최대가능도 모수 추정

모수가 $\mu = (\mu_1, \dots, \mu_K)$ 인 카테고리 분포의 확률밀도함수는 다음과 같다.

$$p(x; \mu_1, \dots, \mu_K) = \text{Cat}(x; \mu_1, \dots, \mu_K)$$
$$= \prod_{k=1}^K \mu_k^{x_k}$$

$$\sum_{k=1}^K \mu_k = 1$$

위 식에서 x 는 모든 k 개의 원소를 가지는 원핫인코딩(one-hot-encoding) 벡터다. 그런데 N 번의 반복시행으로 표본 데이터가 x_1, \dots, x_N 이 있는 경우 모두 독립으로 전체 확률밀도함수는 각각의 확률밀도함수의 곱과 같다.

$$L(\mu_1, \dots, \mu_k; x_1, \dots, x_N) = \prod_{i=1}^N \prod_{k=1}^K \mu_k^{x_{i,k}}$$

위 식에서 $x_{i,k}$ 는 i 번째 시행결과인 x_i 의 k 번째 원소를 뜻한다.
마분을 쉽게 하기 위해 토그변환을 한 로그가능도를 구하면 다음과 같다.

$$\begin{aligned} \log L &= \log p(x_1, \dots, x_N; \mu_1, \dots, \mu_k) \\ &= \sum_{k=1}^K \sum_{i=1}^N (\log \mu_k \cdot x_{i,k}) \quad \text{곱하기-더하기 치환가능 } (\log) \\ &= \sum_{k=1}^K \left(\log \mu_k \left(\sum_{i=1}^N x_{i,k} \right) \right) \end{aligned}$$

k 번째 원소가 나온 횟수를 N_k 라고 표기하자.

$$N_k = \sum_{i=1}^N x_{i,k}$$

그리면 로그가능도가 다음과 같아지며,
이 합수를 친대화하는 모수의 값을 찾어야 한다.

$$\log L = \sum_{k=1}^K (\log \mu_k \cdot N_k)$$

그럼에 모수를 아우 제한조건을 만족해야 한다.

$$\sum_{k=1}^K \mu_k = 1$$

따라서 라그朗주 승수법을 사용하여,
최대가능도에 제한조건을 추가한 새로운 목적함수를 생각할 수 있다.

$$J = \sum_{k=1}^K \log \mu_k N_k + \lambda \left(1 - \sum_{k=1}^K \mu_k \right)$$

이 목적함수를 모수로 미분한 값이 0이 되는 값을 구한다.

$$\frac{\partial J}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \left\{ \sum_{k=1}^K \log \mu_k N_k + \lambda \left(1 - \sum_{k=1}^K \mu_k \right) \right\} = 0 \quad (k=1, \dots, K)$$

$$\frac{\partial J}{\partial \lambda} = \frac{\partial}{\partial \lambda} \left\{ \sum_{k=1}^K \log \mu_k N_k + \lambda \left(1 - \sum_{k=1}^K \mu_k \right) \right\} = 0$$

따라서 다음과 같이 모수를 추정할 수 있다.

$$\frac{N_1}{\mu_1} = \frac{N_2}{\mu_2} = \dots = \boxed{\frac{N_k}{\mu_k}} = \lambda$$

$$N_k = \lambda \mu_k \quad \text{대입}$$

$$\sum_{k=1}^K N_k = \lambda \sum_{k=1}^K \mu_k = \boxed{\lambda = N}$$

$$\mu_k = \frac{N_k}{N} \quad \begin{array}{l} \text{각 범주값} \\ \text{전체사행} \end{array}$$

결론:

최대가능도 추정법에 의한 카테고리별로의 모수는 각 범주값이 나온
횟수와 실제 시행 횟수의 비율이다.

정규분포의 최대가능도 모수 추정

정규분포의 확률밀도함수는 다음과 같다. 여기에서 σ 는 스칼라 값이다.

$$p(x; \theta) = N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

그런데 N 번의 반복시행으로 표본 데이터가 x_1, \dots, x_N 이 있는 경우에는 모두 독립적으로 전체 확률밀도함수는 각각의 확률밀도함수의 곱과 같다.

$$L(\mu; x_1, \dots, x_N) = p(x_1, \dots, x_N; \mu)$$

$$= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

미분을 쉽게 하기 위해, 로그변환을 한 로그가능도를 구하면 다음과 같다.

여기에서 상수부분은 모아서 C로 표기했다.

$$\log L = \log p(x_1, \dots, x_N; \mu)$$

$$= \sum_{i=1}^N \left\{ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} \right\} = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i-\mu)^2$$

i) 확률밀도함수가 최대가 되는 모수값을 찾기 위해서는 각각의 모수로 미분한 값이 0이 되어야 한다.

$$\frac{\partial \log L}{\partial \mu} = \frac{\partial}{\partial \mu} \left\{ \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i-\mu)^2 \right\} = 0$$

$$\frac{\partial \log L}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2} \left\{ \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i-\mu)^2 \right\} = 0$$

위 두식을 풀면, 주어진 데이터에 대한 평균에 대해 모수의 가능도를 가장 크게 하는 모수의 값을 구할 수 있다. 먼저 시에 대한 미분을 정리하면 다음과 같다.

$$\frac{\partial \log L}{\partial \mu} = \frac{2}{2\sigma^2} \sum_{i=1}^N (x_i - \mu) = 0$$

$$N\mu = \sum_{i=1}^N x_i$$

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x}$$

다음으로 σ^2 에 대한 미분을 정리하면 다음과 같다.

$$\frac{\partial \log L}{\partial \sigma^2} = \frac{N}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} \sum_{i=1}^N (x_i - \mu)^2 = 0$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = S^2$$

결론:

최대가능도 추정법에 의한 정규분포의 기댓값은 표본평균과 같고,
분산은 (편향) 표본분산과 같다.

[다변수 정규분포의] 최대가능도 모수 추정

"다변수정규분포의 확률밀도함수"

$$x = M\text{차원 벡터} / \text{기댓값} = M\text{차원 벡터} / \text{공분산 행렬} = M \times M \text{ 행렬}$$

↑ 양의정부호 (positive definite)

라고 정의함.

따라서 정밀도 행렬 $\Sigma^{-1} = \Lambda$
가 존재할 수 있다.

$$p(x; \theta) = N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{M/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

그런데 N번의 반복 시행으로 표본데이터가 x_1, \dots, x_N 이 있는 경우에는 모두 독립적으로
전체 확률밀도함수는 각각의 확률밀도함수의 곱과 같다.

$$L(\mu; x_1, \dots, x_N) = \prod_{i=1}^N \frac{1}{(2\pi)^{M/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right)$$

미분을 쉽게 하기 위해 로그 변환을 한 로그 가능도를 구하면 다음과 같다. 여기에서
상수 부분은 모아서 (로 표기해라.)

(다음장으로) →

**** 다시 보는 행렬 미분법칙 ****

$$\textcircled{1} f(x) = w^T x$$

$$\nabla f = \frac{\partial w^T x}{\partial x} = \frac{\partial x^T w}{\partial x} = w$$

$$\textcircled{2} f(x) = w^T A x$$

$$\nabla f(x) = \frac{\partial x^T A x}{\partial x} = (A + A^T)x$$

$$\textcircled{3} f(x) = Ax$$

$$\nabla f(x) = \frac{\partial (Ax)}{\partial x} = A^T$$

$$\textcircled{4} f(x) = \text{tr}(Wx)$$

$$\nabla f(x) (W \in \mathbb{R}^{MN}) \quad x \in \mathbb{R}^{MN}$$

$$\textcircled{5} f(X) = \log |X| \quad (\log \text{of } X \text{ determinant})$$

$$\frac{\partial f}{\partial X} = \frac{\partial \log |X|}{\partial X} = (X^{-1})^T$$

→ 행렬식은 스칼라, 로그 행렬식도 스칼라, X 로 미분하여 원래 행렬의 역행렬
전치행렬.

두 정방행렬을 곱해서
만들어진 행렬의 대각성분은
스칼라!

$$\frac{\partial f}{\partial X} = \frac{\partial \text{tr}(Wx)}{\partial X} = W^T$$

$$\log L = \log p(x_1, \dots, x_n; \mu)$$

$$= \sum_{i=1}^N \left\{ -\log((2\pi)^{n/2} |\Sigma|^{1/2}) - \frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right\}$$

$$= C - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^N (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

여기에서 기호를 단순화하기 위해 정밀도행렬 Σ^{-1} 을 Λ 로 표시하자.

$$|\Lambda^{-1}| = |\Lambda|^T$$

$$\Lambda = \Sigma^{-1}$$

$$\log L = C + \frac{n}{2} \log |\Lambda| - \frac{1}{2} \sum_{i=1}^N (x_i - \mu)^T \Lambda (x_i - \mu)$$

이 학습률의 합성이 최대가 되는 모수값을 찾기 위해서는, 32개의 항수를 각각의 모수로 미분한 값이 0이 되어야 한다. 미분을 하기 전에 여기에서 사용될 트레이스공식과 행렬미분공식을 다시 정리하겠다.

$$\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$$

$$\frac{\partial w^T x}{\partial w} = \frac{\partial x^T w}{\partial w} = x \quad \dots \text{ 행렬미분법칙 } ①$$

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x \quad \dots \text{ 행렬미분법칙 } ②$$

$$\frac{\partial A x}{\partial x} = A^T \quad \dots \text{ 행렬미분법칙 } ③$$

$$\frac{\partial \text{tr}(w x)}{\partial x} = w^T \quad \dots \text{ 행렬미분법칙 } ④$$

$$\frac{\partial \log |x|}{\partial x} = (x^{-1})^T \quad \dots \text{ 행렬미분법칙 } ⑤$$

우선 32가능도함수를 기댓값벡터로 미분하면 다음과 같다.

$$\begin{aligned}
 \frac{\partial \log L}{\partial \mu} &= -\frac{\partial}{\partial \mu} \sum_{i=1}^N (\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu) \\
 &= -\sum_{i=1}^N \cancel{2\Delta} (\mathbf{x}_i - \mu) \quad \text{행렬미분법칙} \quad (2) \\
 &= -2\Lambda \sum_{i=1}^N (\mathbf{x}_i - \mu) \\
 &= 0
 \end{aligned}$$

Δ 간과 관계없이 이 식이 0이 되려면,

$$\sum_{i=1}^N (\mathbf{x}_i - \mu) = 0$$

$$\mu = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i = \bar{x}$$

32가능도함수를 정밀도행렬로 미분하면 다음과 같다.

$$\begin{aligned}
 \frac{\partial \log L}{\partial \Lambda} &= \boxed{\frac{\partial}{\partial \Lambda} \frac{N}{2} \log |\Lambda|} - \frac{\partial}{\partial \Lambda} \frac{1}{2} \sum_{i=1}^N \cancel{(\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu)} \quad \square \quad \text{행렬미분법칙} \quad (5) \\
 &= \frac{\partial}{\partial \Lambda} \frac{N}{2} \log |\Lambda| - \frac{\partial}{\partial \Lambda} \frac{1}{2} \sum_{i=1}^N \text{tr}((\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu)) \quad \downarrow \text{trace trick} \\
 &= \frac{\partial}{\partial \Lambda} \frac{N}{2} \log |\Lambda| - \frac{\partial}{\partial \Lambda} \frac{1}{2} \sum_{i=1}^N \text{tr}((\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T \Lambda) \\
 &= \boxed{\frac{N}{2} \Lambda^{-T}} - \frac{1}{2} \sum_{i=1}^N ((\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T)^T \quad \text{행렬미분법칙} \quad (4) \\
 &= 0
 \end{aligned}$$

$$\begin{aligned}
 &(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \Lambda \\
 &\quad \boxed{} \times \boxed{} \times \boxed{} \\
 &\quad \boxed{} \times \boxed{} \\
 &\quad \frac{\partial \text{tr}(Wx)}{\partial x} = W^T
 \end{aligned}$$

i) 식을 풀어 모두 합계를 구하면 다음과 같다.

$$\cancel{\Delta^{-1}} \quad \Delta^{-1} = \Sigma = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

(부) 공분산행렬 구하는 공식.

결과는 다음과 같다.

최대가능도 추정법에 의한 대변수점(분포) 기댓값은 표본평균벡터와 같고,
부산은 표본공분산행렬과 같다.