Maximum Likelihood Estimation (MLE)

소프트웨어 꼰대 강의

노기섭 교수

(kafa46@cju.ac.kr)

Possible Learning in Bayesian

- **Maximum Likelihood Estimation (MLE)**
  - Same as frequentist!
  - Dataset only!

목표: 오직 Likelihood만 최대화

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)} \approx P(D|\theta)$$

- **Maximum A Posterior (MAP)**
  - $P(D)$: 알고(given) 있다고 가정
  - $P(\theta)$: 정규분포라고 가정

Likelihood, prior를 동시에 최대화

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)} \approx P(D|\theta) \times P(\theta)$$

- **Bayesian Inference (Variation Inference)**
  - Likelihood, Posterior, Evidence 모두 고려
  - Computing $P(D)$ is intractable
  - Alternatively, using Variational Inference
  - $P(\theta|D)$ 계산이 어렵기 때문에 우리가 알고 있는 함수를 이용하여 잘 모사하도록 접근

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)} \approx Q(\theta|\theta')$$

목표: P를 잘 흉내내는 Q의 파라미터 θ'찾기

먼저 그림으로 감잡기!

■ **대한민국 성인의 평균 키의 분포를 구하는 방법?**

- 가장 단순한 방법

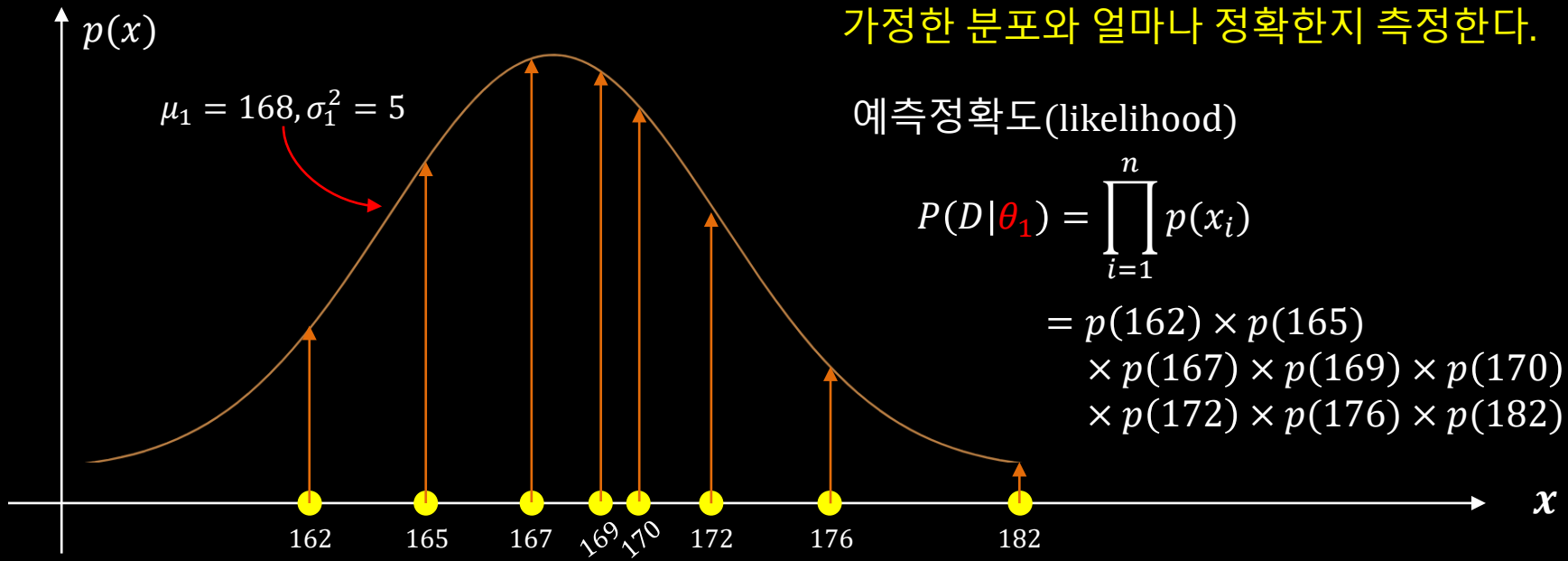
1. 사람 많이 다니는 곳으로 간다.
2. 지나가는 사람을 붙잡는다.
3. 키를 물어본다.
4. 충분한 데이터를 모을 때까지 2번~3번을 반복한다.
5. 평균과 표준편차를 구한다.

이게 맞는 방법인가???

Likelihood를 찾는 직관적 방법 (1/2)

■ **조금 더 있어 보이는 방법!**

- x, y 축을 그리고, 데이터 $D = \{x_i\}_{i=1}^{n}$ 를 표시한다.
- 정규분포를 따른다고 가정하고, 가능한 분포를 $\theta_1 = \{\mu_1 = 168, \sigma^2 = 5\}$ 생성해 본다.



가정한 분포와 얼마나 정확한지 측정한다.

예측정확도(likelihood)

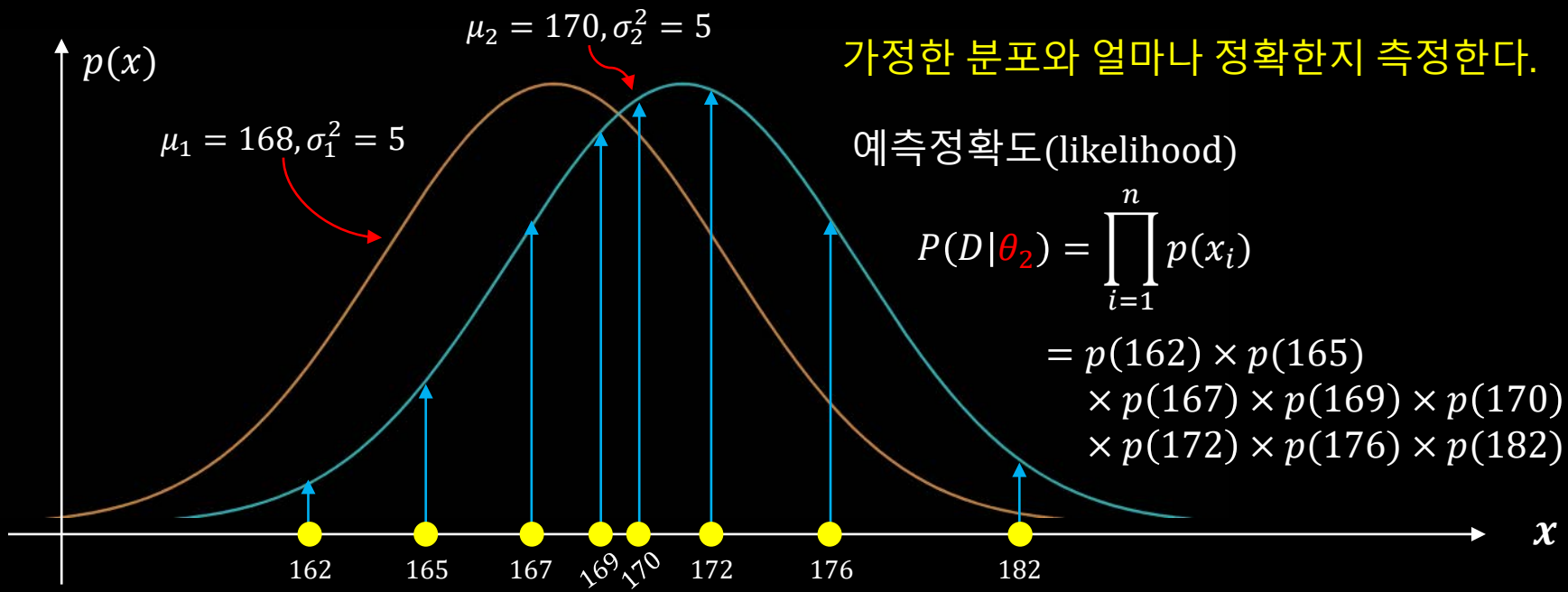
$$P(D|\theta_1) = \prod_{i=1}^{n} p(x_i)$$

$$= p(162) \times p(165) \times p(167) \times p(169) \times p(170) \times p(172) \times p(176) \times p(182)$$

Likelihood를 찾는 직관적 방법 (2/2)

■ **또 다른 파라미터 θ_2 으로 정규 분포를 가정해 본다.**

- 얼마나 정확한지 측정하고,
- 이전 가정보다 더 정확한지 확인한다.



가정한 분포와 얼마나 정확한지 측정한다.

예측정확도(likelihood)

$$P(D|\theta_2) = \prod_{i=1}^{n} p(x_i)$$

$$= p(162) \times p(165) \times p(167) \times p(169) \times p(170) \times p(172) \times p(176) \times p(182)$$

솔루션!

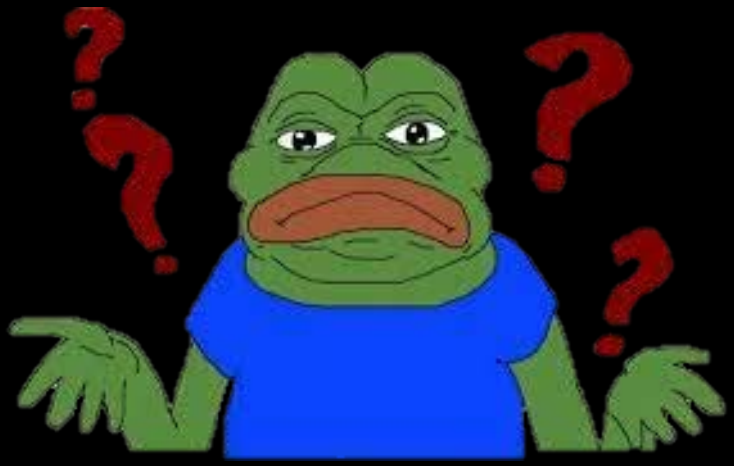
$P(D|\theta_1) \geq P(D|\theta_2)$ → 이전 분포(θ_1) 선택

$P(D|\theta_1) < P(D|\theta_2)$ → 현재 분포(θ_2) 선택

⇨ 가장 큰 값을 갖는 θ_i를 찾을 때까지 계속 반복!

Intuition Achieved, However, ……

- **하지만 이건 좀...**
- **너무 심하지 않나요?**
- **무한정 해 볼 수는 없는 거잖아요???**
  - 맞습니다.
    - 또 다른 간단 예제를 살펴보겠습니다.
    - 그리고 최종적으로 MLE를 사용하는 방법을 설명하겠습니다.



Recap: Basics of MLE

- **What is given & target?**

$$\boldsymbol{P(\theta|D)} = \frac{\boldsymbol{P(D|\theta)} \times \boldsymbol{P(\theta)}}{\boldsymbol{P(D)}} \approx \boldsymbol{P(D|\theta)}$$

Dataset D 는 주어진다.

$\boldsymbol{P(D|\theta)}$ 값을 최대화 하면 된다.

$\boldsymbol{P(D|\theta)}$ 값을 최대화 하는 θ 를 찾으면 된다.

가장 단순한 예제부터 시작해 볼까요? ^^

Toy Example with Simple Linear Function

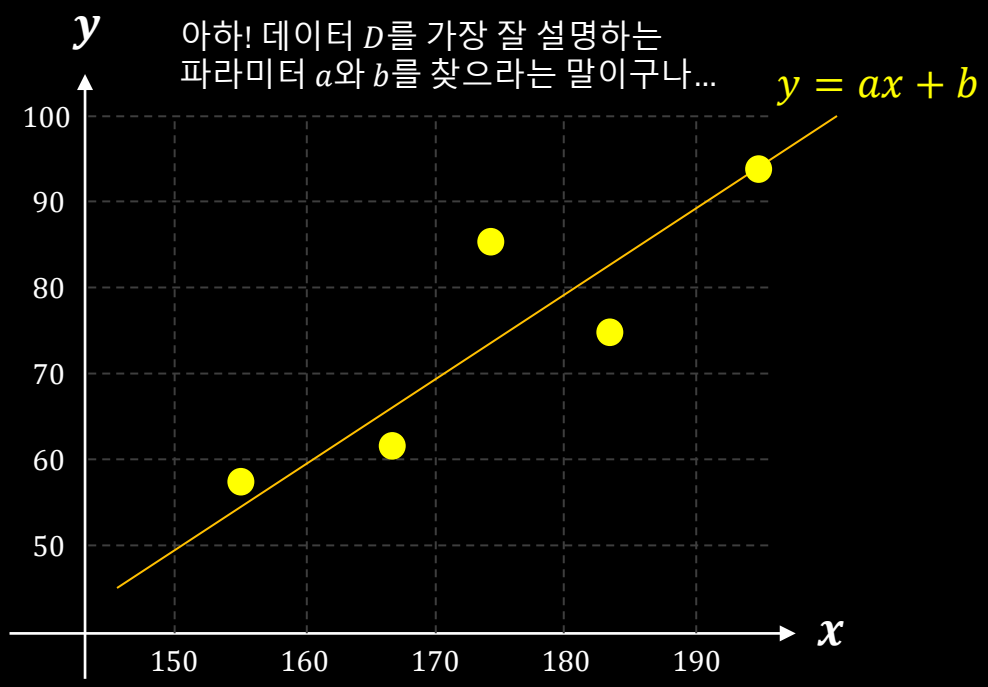
■ **Dataset**

$$D = \{(x_i,\ y_i)\}_{i=1}^{n}$$

Index (i)	Height (키, x)	Weight (몸무게, y)
1	183	75
2	167	61
3	155	57
4	174	85
5	196	92
⋮	⋮	⋮

1차 함수로 설계 했다면, 분명히 직선일 텐데...

아하! 데이터 D를 가장 잘 설명하는
파라미터 a와 b를 찾으라는 말이구나...



만약 $a = 1, b = -100$ 이라고 찾았다면...

$y = x - 100$ ← $y = f(x|\theta)$

만약 새로운 입력 176 Cm가 들어오면 어떻게 예측할까요?

$y = 176 - 100 = 76$ 76 Kg 이라고 예측!

그런데 이게 100%
정확한 예측일까요?

How to Guarantee this Solution is Correct?

■ **이게 정확한가요?**

만약 새로운 입력 176 Cm가 들어오면?

$y = 176 - 100 = 76$ 76 Kg 이라고 예측!

176 Cm인 사람이 74 Kg 일수도… 80 Kg 일수도 있잖아요 ^^.

결론

데이터를 관측해서 추론한 파라미터는 100% 맞을 수 없다!

가능한 접근 방법

1. 추측한 파라미터로 예측한 값은 오차가 있다.
2. 그 오차의 분포를 알면 되지 않을까?
3. 오차의 분포를 최소화 하도록 파라미터를 찾는다면, 그나마 비슷하게 맞출 것 같다.

One step more onto Solution

- **실제 값과 예측 값의 분포라는 게...**
- **교수님! 너무 막연해요 ㅠㅠ**

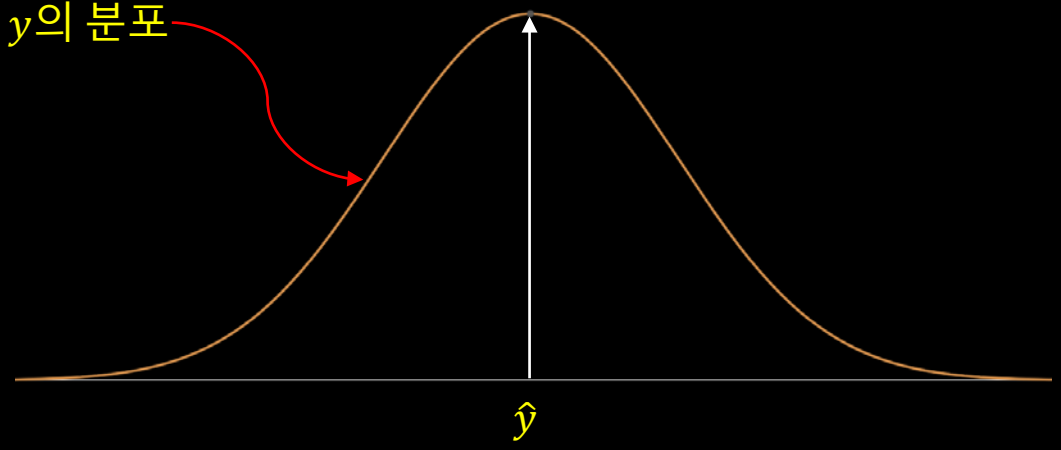
확률적으로 접근하는 방법이 가장 간단합니다. (사실은 복잡해요 ㅠㅠ)

내가 예측한 몸무게는 평균 μ, 표준편차 σ를 파라미터로 갖는 정규분포를 따른다고 가정합니다.

실제 몸무게 $\sim N$(예측한 몸무게, σ^2)

입력값(키)를 x,
예측한 몸무게를 $\hat{y}$, 실제 몸무게를 y라고 하면

$y \sim N(\hat{y}, \sigma^2)$



One step more onto Solution

실제 몸무게 ~ N(예측한 몸무게, σ^2)

$y \sim N(\hat{y}, \sigma^2)$

입력값(키)를 x,
예측한 몸무게를 $\hat{y}$, 실제 몸무게를 y라고 하면

$$y \sim N(f(x|\theta), \sigma^2), where\ \ \theta = \{a, b\}$$

정규분포 확률함수: $p(x\ ;\ \mu, \sigma^2)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

x 자리에 → $y|x$ 입력

μ자리에 → $\hat{y}$ 입력

$$p(y|x\ ;\ \hat{y}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\hat{y})^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-f(x|\theta))^2}{2\sigma^2}}$$

Find Objective Function

$$\boldsymbol{D} = \{(\boldsymbol{x_i},\ \boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{n}} = \{(\boldsymbol{x_1},\ \boldsymbol{y_1}), (\boldsymbol{x_2},\ \boldsymbol{y_2}), \cdots, (\boldsymbol{x_n},\ \boldsymbol{y_n})\}$$

$p(D)$ ⇨

x_1 주어졌을 때 y_1 나타날 확률

x_2 주어졌을 때 y_2 나타날 확률

⋮ ⋮ ⋮

x_n 주어졌을 때 y_n 나타날 확률

동시에 발생해야 함 → 결합확률 (각각의 확률 곱하기)

요게 likelihood!!!

$p(D)$는 θ 설정에 따라 바뀔 것임 ⇨ $p(D|\theta)$

우리가 최대화 해야 하는 값!!!

$$p(D|\theta) = p(y_1|x_1) \times p(y_2|x_2) \times \cdots \times p(y_n|x_n)$$

$$= \prod_{i=1}^{n} p(y_1|x_1) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}} = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-f(x_i|\theta))^2}{2\sigma^2}}$$

Finally, Log-likelihood!

■ **다시 간단히 정리하면,**

$$p(D|\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}}$$

양변에 로그를 취한다

- 최대값 구하는 것에는 영향 없음
- 곱하기를 더하기로 바꿀 수 있음

$$\ln p(D|\theta) = \ln \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}}$$

로그를 취한 likelihood를
다른 말로 "log-likelihood" 라고 부릅니다.

$$\ln p(D|\theta) = \sum_{i=1}^{n} \left(-\ln \sigma\sqrt{2\pi} - \frac{(y-\hat{y})^2}{2\sigma^2} \right) = -\sum_{i=1}^{n} \left(\ln \sigma\sqrt{2\pi} + \frac{(y_i-\hat{y}_i)^2}{2\sigma^2} \right)$$

마이너스 부호가 거슬림 → 마이너스 제거하고 maximize 문제를 minimize 문제로 변환

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\ln \sigma\sqrt{2\pi} + \frac{(y_i-\hat{y}_i)^2}{2\sigma^2} \right)$$

π, σ 는 상수이므로
max 과정에 영향 없음

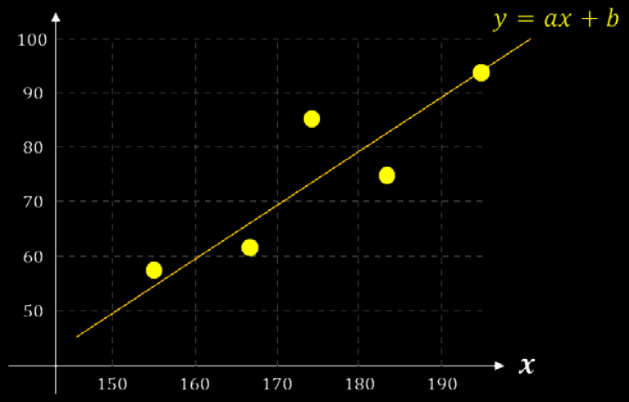
⇨

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i-\hat{y}_i)^2$$

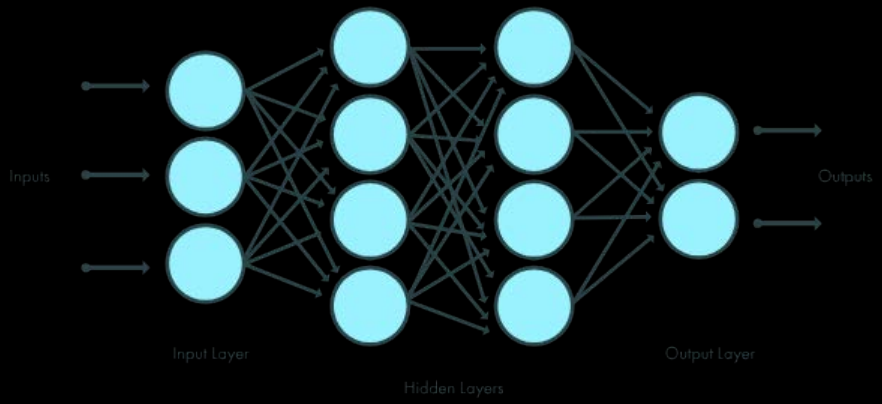
Expansion to Deep Learning

파라미터 구성에만 차이가 있을 뿐.

일차함수 파라미터 θ: 2개



딥러닝 파라미터 θ: 많다



$$= \prod_{i=1}^{n} p(y_1|x_1) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}} = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-f(x_i|\theta))^2}{2\sigma^2}}$$

파라미터 수만 다를 뿐
결론은 같다!

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

어차피 수식은 똑같음.
벡터 또는 행렬로 처리하는 것은 파라미터 개수와 무관

More Efficient Method to Find Optimal Parameters

■ 그러면 likelihood $p(D|\theta)$ 최대화 하는 θ 는 어떻게 찾지?

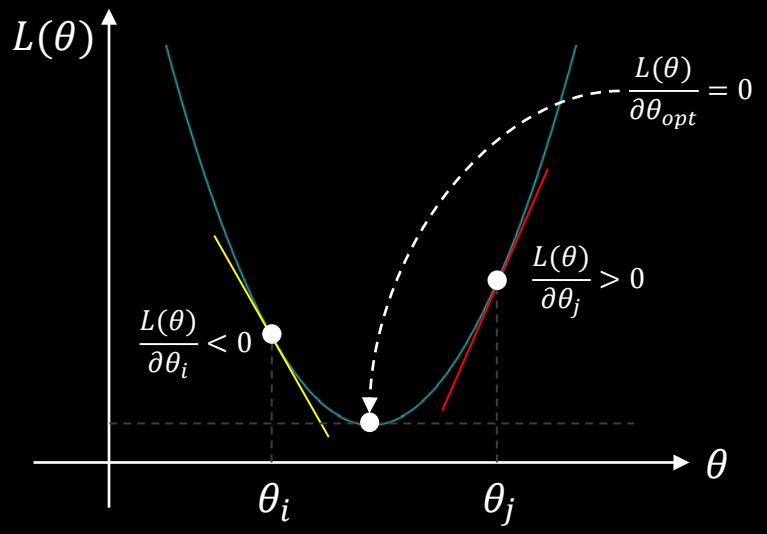
$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

요 놈을 어떻게 설정하냐에 따라 $(y_i - \hat{y}_i)$ 값이 달라질 것.

요럴 때 유용한 것이 미분!

$$\theta \leftarrow \theta - \eta \times \frac{L(\theta)}{\partial \theta}$$

요렇게 하는 것을 Gradient Descent 라고 부릅니다.



자세한 내용은 딥러닝 수학 "미분 시리즈"에서 자세히 다룹니다.

수고하셨습니다 ..^^..