Linear Regression via Maximum Likelihood Estimation

Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

April 6, 2023

Hồi quy tuyến tính thông qua ước tính khả năng tối đa

Lương Ngọc Hoàng

Trường Đại học Công nghệ Thông tin (UIT), ĐHQG-HCM

Ngày 6 tháng 4 năm 2023

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

Maximum Likelihood Estimation (MLE) - Example

- A bag contains 3 balls, each ball is either red or blue.
- The number of blue balls can be $0, 1, 2, 3$.
- Choose 4 balls randomly **with replacement**.
- The following balls are observed: blue, red, blue, blue.
- How many **blue balls** should there be in the bag so that the probability of the observed sample (blue, red, blue, blue) is the largest?

Ước tính khả năng xảy ra tối đa (MLE) - Ví dụ

• Một hộp đựng 3 quả bóng, mỗi quả bóng có màu đỏ hoặc xanh. • Số bi xanh có thể là 0, 1, 2, 3. • Chọn ngẫu nhiên 4 bi thay thế. • Quan sát thấy các quả bóng sau: xanh, đỏ, xanh, xanh. • Trong túi phải có bao nhiêu quả bóng màu xanh để xác suất của mẫu quan sát (xanh, đỏ, xanh, lam) là lớn nhất?

Bernoulli Random Variables

- A Bernoulli random variable X takes two possible values, usually 0 and 1, modeling random experiments that have two possible outcomes (e.g., “success” and “failure”).
  - e.g., tossing a coin. The outcome is either Head or Tail.
  - e.g., taking an exam. The result is either Pass or Fail.
  - e.g., classifying images. An image is either Cat or Non-cat.

Biến ngẫu nhiên Bernoulli

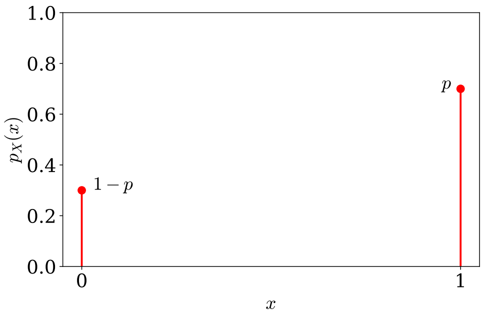
- Biến ngẫu nhiên Bernoulli X nhận hai giá trị có thể, thường là 0 và 1, mô hình hóa các thí nghiệm ngẫu nhiên có thể có hai kết quả (ví dụ: “thành công” và “thất bại”).
  - ví dụ như tung đồng xu. Kết quả là Đầu hoặc Đuôi.
  - ví dụ như làm bài kiểm tra. Kết quả là Đạt hoặc Không đạt.
  - ví dụ: phân loại hình ảnh. Một hình ảnh là Mèo hoặc Không phải mèo.

Bernoulli Random Variables

Definition

A random variable X is a Bernoulli random variable with parameter $p \in [0,1]$, written as $X \sim Bernoulli(p)$ if its PMF is given by

$$P_X(x) = \begin{cases} p, & \text{for } x = 1 \\ 1-p, & \text{for } x = 0. \end{cases}$$

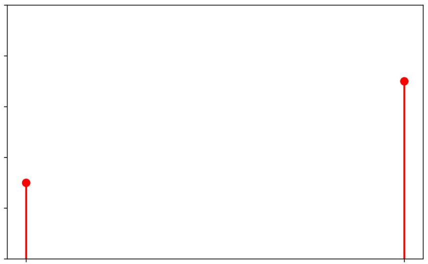


Biến ngẫu nhiên Bernoulli

Sự định nghĩ a

Biến ngẫu nhiên X là biến ngẫu nhiên Bernoulli với tham số p [0, 1], được viết là X Bernoulli(p) nếu PMF của nó được cho bởi

PX(x) = P, cho x = 1
1 p, với x = 0.



Example

- A bag contains 3 balls, each ball is either red or blue.
- The number of blue balls θ can be $0, 1, 2, 3$.
- Choose 4 balls randomly **with replacement**.
- Random variables X_1, X_2, X_3, X_4 are defined as

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th chosen ball is blue} \\ 0, & \text{if the } i\text{-th chosen ball is red} \end{cases}$$

- The following balls are observed: blue, red, blue, blue.
- Therefore, $x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1$.
- Note that X_i's are i.i.d. (independent and identically distributed) and $X_i \sim Bernoulli(\frac{\theta}{3})$. For which value of θ is the probability of the observed sample $(x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1)$ is the largest?

Ví dụ

- Một hộp đựng 3 quả bóng, mỗi quả bóng có màu đỏ hoặc xanh.
- Số bi xanh θ có thể là 0, 1, 2, 3.
- Chọn ngẫu nhiên 4 bi thay thế.
- Các biến ngẫu nhiên X1, X2, X3, X4 được định nghĩ a là

Xi = 1, nếu quả bóng được chọn thứ i có màu xanh lam
0, nếu quả cầu thứ i được chọn là màu đỏ

- Quan sát thấy các quả bóng sau: xanh, đỏ, xanh, xanh.
- Do đó, x1 = 1, x2 = 0, x3 = 1, x4 = 1.
- Lưu ý rằng Xi 's là iid (độc lập và phân phối đồng nhất) và Xi Bernoulli(mẫu $\frac{\theta}{3}$). Với giá trị nào của θ là xác suất của quan sát (x1 = 1, x2 = 0, x3 = 1, x4 = 1) lớn nhất?

Example

$$P_{X_i}(x) = \begin{cases} \frac{\theta}{3}, & \text{for } x = 1 \\ 1 - \frac{\theta}{3}, & \text{for } x = 0 \end{cases}$$

X_i's are independent, the joint PMF of X_1, X_2, X_3, X_4 can be written

$$P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)P_{X_4}(x_4)$$

$$P_{X_1X_2X_3X_4}(1, 0, 1, 1) = \frac{\theta}{3} \cdot \left(1 - \frac{\theta}{3}\right) \cdot \frac{\theta}{3} \cdot \frac{\theta}{3} = \left(\frac{\theta}{3}\right)^3 \left(1 - \frac{\theta}{3}\right)$$

θ	$P_{X_1X_2X_3X_4}(1, 0, 1, 1; \theta)$
0	0
1	0.0247
2	0.0988
3	0

The observed data is most likely to occur for $\theta = 2$.
We may choose $\hat{\theta} = 2$ as our estimate of θ.

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

Ví dụ

PXi (x) = θ 3 , cho x = 1
1 θ 3 , cho x = 0

Xi độc lập, PMF chung của X1, X2, X3, X4 có thể được viết

PX1X2X3X4 (x1, x2, x3, x4) = PX1 (x1)PX2 (x2)PX3 (x3)PX4 (x4)

PX1X2X3X4 (1, 0, 1, 1) = θ 3 (1 θ 3) θ 3 θ 3 = (θ 3)^3 (1 θ 3)

θ	PX1X2X3X4 (1, 0, 1, 1; θ)
0	0
1	0,0247
2	0,0988
3	0

Dữ liệu quan sát có nhiều khả năng xảy ra nhất đối với θ = 2.
Chúng ta có thể chọn ˆ θ = 2 làm ước lượng của θ.

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

Introduction

- The process of estimating the values of parameters $\mathbf{b}$ from some dataset $\mathcal{D}$ is called **model fitting**, or **training**, is at the heart of machine learning.
- There are many methods for estimating $\mathbf{b}$, and they involve an optimization problem of the form

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmin}}\, \mathcal{L}(\mathbf{b})$$

where $\mathcal{L}(\mathbf{b})$ is some kind of loss function or objective function.
- The process of quantifying uncertainty about an unknown quantity estimated from a finite sample of data is called **inference**.
- In deep learning, the term "inference" refers to "prediction", namely computing

$$p(y \mid \mathbf{x}, \hat{\mathbf{b}})$$

Giới thiệu

- Quá trình ước tính giá trị của tham số b từ một số tập dữ liệu D được gọi là điều chỉnh mô hình hoặc đào tạo, là trung tâm của học máy.
- Có nhiều phương pháp để ước tính b, và chúng liên quan đến một bài toán tối ưu dạng

b^ = argmin L(b)
b

trong đó L(b) là một số loại hàm mất mát hoặc hàm mục tiêu. •

Quá trình định lượng độ không đảm bảo về một đại lượng chưa biết được ước tính từ một mẫu dữ liệu hữu hạn được gọi là suy luận.

- Trong học sâu, thuật ngữ "suy luận" đề cập đến "dự đoán", cụ thể là tính toán

p(y x,b^)

Maximum Likelihood Estimation

- The most common approach to parameter estimation is to pick the parameters that assign the highest probability to the training data. This is called **maximum likelihood estimation** or **MLE**.

$$\hat{\mathbf{b}}_{\mathtt{mle}} = \underset{\mathbf{b}}{\operatorname{argmax}}\, p(\mathcal{D} \mid \mathbf{b})$$

- We usually assume the training examples are "independent and identically distributed", and are sampled from the same distribution (i.e., the **iid** assumption). The conditional likelihood becomes

$$p(\mathcal{D} \mid \mathbf{b}) = p(y_1, y_2, \ldots, y_n \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{b}) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \mathbf{b})$$

- We usually work with the **log likelihood**, which decomposes into a sum of terms, one per example.

$$\mathrm{LL}(\mathbf{b}) = \log p(\mathcal{D} \mid \mathbf{b}) = \log \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \mathbf{b}) = \sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathbf{b})$$

Ước lượng khả năng tối đa

- Cách tiếp cận phổ biến nhất để ước lượng tham số là chọn tham số gán xác suất cao nhất cho dữ liệu huấn luyện. Điều này được gọi là ước tính khả năng tối đa hoặc MLE.

b^ con lừa = argmax b p(D b)

- Chúng ta thường cho rằng các ví dụ huấn luyện là "độc lập và được phân phối giống hệt nhau" và được lấy mẫu từ cùng một phân phối (nghĩ a là giả định iid). Khả năng có điều kiện trở thành

p(D b) = p(y1, y2, . . . , yn x1,x2, ... , xn,b) = N p(yi xi ,b) tôi = 1

- Chúng tôi thường làm việc với khả năng xảy ra của nhật ký, phân tách thành tổng các số hạng, mỗi số hạng một ví dụ.

LL(b) = log p(D b) = log N p(yi xi ,b) = N log p(yi xi ,b) tôi = 1 tôi = 1

Maximum Likelihood Estimation

- The MLE is given by

$$\hat{\mathbf{b}}_{\mathtt{mle}} = \underset{\mathbf{b}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathbf{b})$$

- Because most optimization algorithms are designed to *minimize* cost functions, we redefine the objective function to be the conditional **negative log likelihood** or **NLL**:

$$\mathrm{NLL}(\mathbf{b}) = -\log p(\mathcal{D} \mid \mathbf{b}) = -\sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathbf{b})$$

- Minimizing this will give the MLE.

$$\hat{\mathbf{b}}_{\mathtt{mle}} = \underset{\mathbf{b}}{\operatorname{argmin}} -\sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathbf{b})$$

Ước lượng khả năng tối đa

- MLE được đưa ra bởi

b^ con lừa = argmax b N log p(yi xi ,b) tôi = 1

- Vì hầu hết các thuật toán tối ưu hóa được thiết kế để giảm thiểu hàm chi phí, nên chúng tôi xác định lại hàm mục tiêu là khả năng nhật ký âm có điều kiện hoặc NLL:

NLL(b) = log p(D b) = N log p(yi xi ,b) tôi = 1

- Giảm thiểu điều này sẽ cung cấp MLE.

b^ con lừa = argmin b N log p(yi xi ,b) tôi = 1

MLE for the Bernoulli distribution

- Suppose Y is a random variable representing a coin toss.
- The event $Y = 1$ corresponds to heads, $Y = 0$ corresponds to tails.
- The probability distribution for this rv is the Bernoulli. The NLL for the Bernoulli distribution is

$$\begin{aligned}
\text{NLL}(b) &= -\log \prod_{i=1}^{n} p(y_i \mid b) = -\log \prod_{i=1}^{n} b^{\mathbb{I}(y_i=1)}(1-b)^{\mathbb{I}(y_i=0)} \\
&= -\sum_{i=1}^{n} \mathbb{I}(y_i = 1)\log(b) + \mathbb{I}(y_i = 0)\log(1-b) \\
&= -[N_1 \log(b) + N_0 \log(1-b)]
\end{aligned}$$

where

- $N_1 = \sum_{i=1}^{n} \mathbb{I}(y_i = 1)$ is the number of heads
- $N_0 = \sum_{i=1}^{n} \mathbb{I}(y_i = 0)$ is the number of tails.

- $N = N_0 + N_1$ is the **sample size**.

MLE cho phân phối Bernoulli

• Giả sử Y là biến ngẫu nhiên đại diện cho việc tung đồng xu. • Biến cố Y = 1 tương ứng mặt ngửa, Y = 0 tương ứng mặt sấp. • Phân phối xác suất cho rv này là Bernoulli. NLL cho phân phối Bernoulli là

NLL(b) = log p(yi=0) b) = log b I(yi=1) (1 b)

= I(yi = 1) log(b) + I(yi = 0) log(1 b)

= [N1 log(b) + N0 log(1 b)]

Ở đâu

- N1 = I(yi = 1) là số mặt ngửa
- N0 = I(yi = 0) là số mặt sấp. •

N = N0 + N1 là cỡ mẫu.

MLE for the Bernoulli distribution

$$\mathtt{NLL}(b) = -[N_1 \log(b) + N_0 \log(1-b)]$$

- The derivative of the NLL is

$$\frac{d}{db}\mathtt{NLL}(b) = \frac{-N_1}{b} + \frac{N_0}{1-b}$$

- The MLE can be found by solving $\frac{d}{db}\mathtt{NLL}(b) = 0$.
- The MLE is given by

$$\hat{b}_{\mathtt{mle}} = \frac{N_1}{N_0 + N_1}$$

which is the empirical fraction of heads.

MLE cho phân phối Bernoulli

NLL(b) = [N1 log(b) + N0 log(1 b)]

- Đạo hàm của NLL là

đ / db NLL(b) = N1 / b + N0 / 1 b

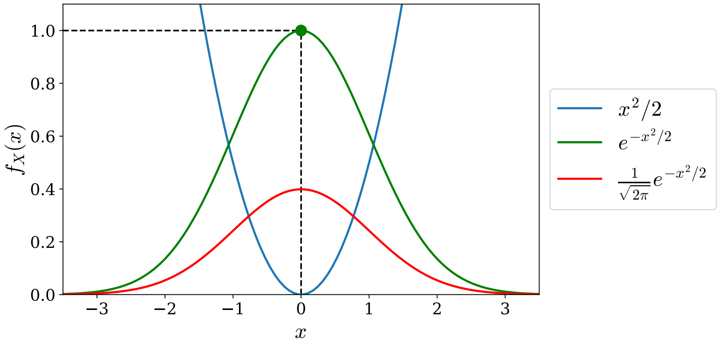
- Có thể tìm MLE bằng cách giải đ / dbNLL(b) = 0. •

MLE được cho bởi

ˆ bmle = N1 / N0 + N1

đó là phần thực nghiệm của người đứng đầu.

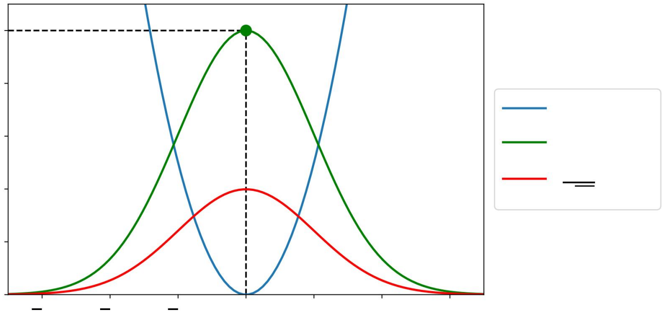
Standard Normal (Gaussian) Random Variable $N(0,1)$



$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

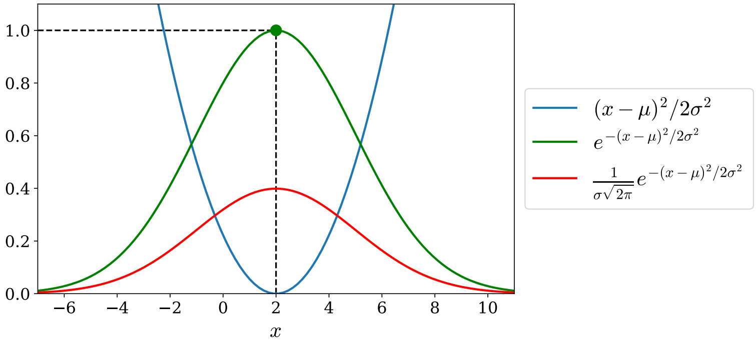
Tiêu chuẩn Bình thường (Gaussian) Biến ngẫu nhiên N(0, 1)



∞
e2 x /2 dx = √ 2π
∞

fX(x) = 1 / √ 2π e2 x /2

General Normal (Gaussian) Random Variable $N(\mu, \sigma^2)$

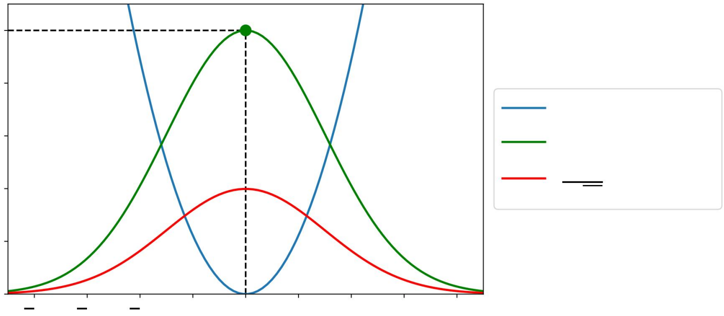


$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

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

$$[X] = \mu \qquad (X) = \sigma^2$$

Biến ngẫu nhiên bình thường chung (Gaussian) N(μ, σ2)



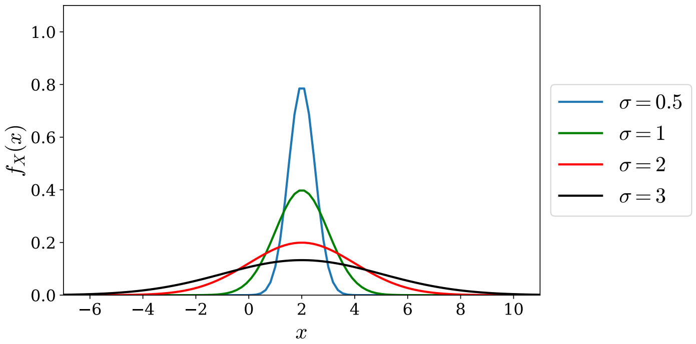
1 fX(x) = σ √ 2π e(x μ) 2 /2σ 2

= 1 √ 2πσ2 kinh nghiệm (- 1 2σ 2 (y μ) 2)

[X] = μ(X) = σ2

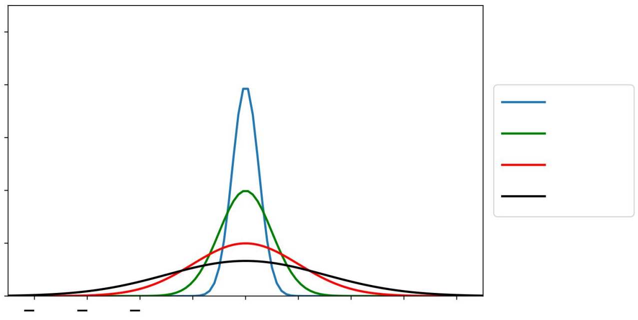
VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT
13 / 22

General Normal (Gaussian) Random Variable $N(\mu, \sigma^2)$



- Smaller σ, narrower PDF.
- Let $Y = aX + b$ $N \sim N(\mu, \sigma^2)$
- Then, $[Y] = aE[X] + b$ $(Y) = a^2\sigma^2$ (always true)
- But also, $Y \sim N(a\mu + b, a^2\sigma^2)$

Biến ngẫu nhiên bình thường chung (Gaussian) N(μ, σ2)



- σ càng nhỏ, PDF càng hẹp.
- Cho Y = aX + b N N(μ, σ2) •

Khi đó, [Y] = aE[X] + b (Y) = a 2σ 2 (luôn đúng) •

Ngoài ra, Y N(aμ + b , a2σ 2)

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT
14 / 22

MLE for Gaussian Example

- We have $N = 3$ data points $y_1 = 1,\ y_2 = 0.5,\ y_3 = 1.5$ which are independent and Gaussian with unknown mean μ and variance 1:

$$y_i \sim \mathcal{N}(\mu, 1)$$

- Likelihood $P(y_1 y_2 y_3|\mu) = P(y_1|\mu)P(y_2|\mu)P(y_3|\mu)$.
- Consider two guesses $\mu = 1.0$ and $\mu = 2.5$. Which has higher likelihood?
- Finding the μ that maximizes the likelihood is equivalent to moving the Gaussian until the product $P(y_1|\mu)P(y_2|\mu)P(y_3|\mu)$ is maximized.

MLE cho ví dụ Gaussian

- Ta có N = 3 điểm dữ liệu y1 = 1, y2 = 0,5, y3 = 1,5 là độc lập và Gaussian với giá trị trung bình μ chưa biết và phương sai 1:

yi N (μ, 1)

- Xác suất P(y1y2y3 μ) = P(y1 μ)P(y2 μ)P(y3 μ). • Xét

hai dự đoán μ = 1,0 và μ = 2,5. Cái nào cao hơn khả năng?

- Việc tìm μ tối đa hóa khả năng xảy ra tương đương với việc di chuyển Gaussian cho đến khi tích P(y1 μ)P(y2 μ)P(y3 μ) đạt cực đại.

MLE for the univariate Gaussian

- $Y \sim \mathcal{N}(\mu, \sigma^2)$ and $\mathcal{D} = \{y_n : n = 1 : N\}$ be an iid sample of size N.

$$p(y \mid \mathbf{b}) = \mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2 \right)$$

- We can estimate the parameters $\mathbf{b} = (\mu, \sigma^2)$ using MLE.
- We derive the NLL, which is given by

$$\mathrm{NLL}(\mu, \sigma^2) = -\sum_{n=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \mu)^2 \right) \right]$$

$$= \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mu)^2 + \frac{N}{2} \log(2\pi\sigma^2)$$

- The minimum of this function must satisfy the following conditions

$$\frac{\partial}{\partial \mu} \mathrm{NLL}(\mu, \sigma^2) = 0, \quad \frac{\partial}{\partial \sigma^2} \mathrm{NLL}(\mu, \sigma^2) = 0$$

MLE cho Gaussian đơn biến • Y N

(μ, σ2) và D = {yn n = 1 N} là một mẫu iid có kích thước N.

p(y b) = N (y μ, σ2) = 1 / √ 2πσ2 kinh nghiệm (- 1 / 2σ 2 (y μ) 2)

• Chúng ta có thể ước tính các tham số b = (μ, σ2) bằng MLE. • Chúng tôi rút ra NLL, được đưa ra bởi

NLL(μ, σ2) = N n=1 log [(1 / 2πσ2) 1 2 kinh nghiệm (- 1 / 2σ 2 (yn μ) 2)]

= 1 / 2σ 2 N n=1 (yn μ)² + Nhật ký N(2πσ 2) / 2

• Cực tiểu của hàm này phải thỏa mãn các điều kiện sau

μNLL(μ, σ2) = 0, NLL(μ, σ2) = 0 σ2

MLE for the univariate Gaussian

- The solution is given by

$$\hat{\mu}_{\texttt{mle}} = \frac{1}{N}\sum_{n=1}^{N} y_n = \bar{y}$$

$$\hat{\sigma}^2_{\texttt{mle}} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{\mu}_{\texttt{mle}})^2 = \frac{1}{N}\left[\sum_{n=1}^{N} y_n^2 + \hat{\mu}^2_{\texttt{mle}} - 2y_n\hat{\mu}_{\texttt{mle}}\right] = s^2 - \bar{y}^2$$

$$s^2 \triangleq \frac{1}{N}\sum_{n=1}^{N} y_n^2$$

- The quantities $\bar{y}$ and s^2 are called the **sufficient statistics** of the data because they are sufficient to compute the MLE.
- Sometimes, we might se the estimate for the variance as

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \hat{\mu}_{\texttt{mle}})^2$$

which is not the MLE, but is a different kind of estimate.

MLE cho Gaussian đơn biến

- Giải pháp được đưa ra bởi

$$\hat{\mu}_{\text{mle}} = \frac{1}{n}\sum_{n=1}^{N} y_n = \bar{y}$$

$$\hat{\sigma}^2_{\text{con lừa}} = \frac{1}{n}\sum_{n=1}^{N}(y_n \quad \hat{\mu}_{\text{mle}})^2 = \frac{1}{N}\Big[\sum_{n=1}^{N} {}_{\text{năm}}N^2 + \hat{\mu}^2_{\text{con lừa}} \quad 2y_n\hat{\mu}_{\text{mle}}\Big] = s^2 \quad \bar{y}^2$$

$$^2{}_{\text{giây}} \quad \frac{1}{n}\sum_{n=1}^{N} {}_{\text{năm}}N^2$$

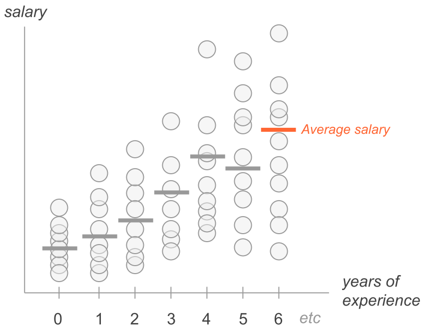
- Các đại lượng $\bar{y}$ và s^2 được gọi là thống kê đủ của dữ liệu vì chúng đủ để tính toán MLE. • Đôi khi, chúng ta có thể xem ước tính cho phương sai là

$$\hat{\sigma}^2 = \frac{1}{N \quad 1}\sum_{n=1}^{N}(y_n \quad \hat{\mu}_{\text{mle}})^2$$

đó không phải là MLE, mà là một loại ước tính khác.

Linear Regression Example

- We want to predict the salary of a new NBA player.
- If we know this new player has 6 years of experience, we look at the average salaries of players with the same experience.



- In all examples, the predicted salary is a conditional mean:

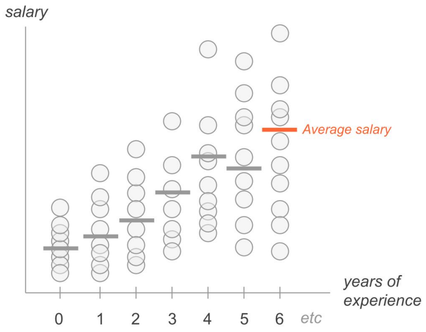
$$\hat{y}_0 = \text{avg}(y_i | \mathbf{x}_i = \mathbf{x}_0)$$

Ví dụ về hồi quy tuyến tính • Chúng tôi

muốn dự đoán mức lương của một cầu thủ NBA mới. • Nếu chúng tôi biết

người chơi mới này có 6 năm kinh nghiệm, chúng tôi sẽ xem xét mức lương trung bình

của những người chơi có cùng kinh nghiệm.



- Trong tất cả các ví dụ, mức lương dự đoán là giá trị trung bình có điều kiện:

y^ 0 = avg(yi xi = x0)

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT
18 / 22

Linear Regression Example

- The prediction is a conditional mean:

$$\hat{y}_0 = \mathtt{avg}(y_i|\mathbf{x}_i = \mathbf{x}_0)$$

- But this strategy only works if we have data points $\mathbf{x}_i$ match the query point $\mathbf{x}_0$.
- The core idea of regression: Obtaining prediction $\hat{y}_0$ using quantities of the form $\mathtt{avg}(y_i|\mathbf{x}_i = \mathbf{x}_0)$, which can be formalized as:

$$\mathbb{E}(y_i|x^*_{i1}, x^*_{i2}, \ldots, x^*_{ip}) \longrightarrow \hat{y}$$

where x^*_{ij} is the i-th measurement of the j-th variable.

- The **regression function**: a conditional expectation.
- In a **linear regression model**, we combine features X to say something about the response Y.
- In the **univariate case**, the regression function is a linear equation.

Ví dụ hồi quy tuyến tính

- Dự đoán là trung bình có điều kiện:

y^ 0 = avg(yi xi = x0)

- Nhưng chiến lược này chỉ hoạt động nếu chúng ta có điểm dữ liệu xi khớp với điểm truy vấn x0.
- Ý tưởng cốt lõi của hồi quy: Dự đoán y^ 0 bằng cách sử dụng các đại lượng có dạng avg(yi xi = x0), có thể được viết thành:

E(yi xi1 , Xi2 , . . . , x ip) y^

trong đó x là phép đo thứ i của biến thứ j. ij

- Hàm hồi quy: kỳ vọng có điều kiện. • Trong mô hình hồi quy tuyến tính, chúng ta kết hợp các đặc trưng X để nói điều gì đó về phản hồi Y . • Trong trường hợp đơn biến, hàm hồi quy là một phương trình tuyến tính.

MLE for linear regression

- Consider a linear regression model:

$$y_i = b_0 x_{i0} + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} + \epsilon_i = \mathbf{b}^\top \mathbf{x}_i + \epsilon_i$$

- Assume that the noise terms ϵ_i are independent and have a Gaussian distribution with mean 0 and and constant variance σ^2.

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Then we have:

$$y_i \sim \mathcal{N}(\mathbf{b}^\top \mathbf{x}_i, \sigma^2)$$

- Under this assumption, how can we obtain the parameters $\mathbf{b} = (b_0, b_1, \ldots, b_p)$ of the linear regression model?

MLE cho hồi quy tuyến tính

- Xét mô hình hồi quy tuyến tính:

yi = b0xi0 + b1xi1 + b2xi2 + · · · + bpxip + εi = b xi + εi

- Giả sử rằng các số hạng nhiễu εi là độc lập và có Phân phối Gaussian với giá trị trung bình 0 và phương sai không đổi σ 2 .

εi N (0, σ 2)

- Khi đó ta có:

yi N (b xi 2σ ,)

- Với giả định này, làm thế nào để có được các tham số b = (b0, b1, . . . , bp) của mô hình hồi quy tuyến tính?

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT
20 / 22

MLE for linear regression

- The joint distribution of $\mathbf{y} = (y_1, y_2, \dots, y_n)$ is:

$$
\begin{aligned}
P(\mathbf{y}|\mathbf{X}, \mathbf{b}, \sigma^2) &= \prod_{i=1}^{n} f(y_i; \mathbf{X}, \mathbf{b}, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mathbf{b}^\top \mathbf{x}_i)^2\right\} \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{b}^\top \mathbf{x}_i)^2\right\}
\end{aligned}
$$

- Taking logarithm, we have:

$$
\begin{aligned}
LL &= \log\left(P(\mathbf{y}|\mathbf{X}, \mathbf{b}, \sigma^2)\right) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{b}^\top \mathbf{x}_i)^2 \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{b})^\top(\mathbf{y} - \mathbf{X}\mathbf{b})
\end{aligned}
$$

MLE cho hồi quy tuyến tính •

Phân phối chung của y = (y1, y2, . . . , yn) là:

P(y X,b, σ 2) = N tôi = 1 f(yi ; X,b, σ 2)

= N tôi = 1 1 √ 2πσ2 hết hạn { 1 2σ 2 (yi b xi) 2 }

= (2πσ 2) N 2 điểm kinh nghiệm { 1 2σ 2 N tôi = 1 (yi b xi) 2}

- Lấy logarit, ta có:

LL = log (P(y X,b, σ 2))

= N 2 log(2πσ 2) 1 2σ 2 N tôi = 1 (yi b xi) 2

= N 2 log(2πσ 2)(y Xb) 1 2σ 2 (y Xb)

MLE for linear regression

$$
\begin{aligned}
LL &= \log\left(P(\mathbf{y}|\mathbf{X}, \mathbf{b}, \sigma^2)\right) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{b})^\top(\mathbf{y}-\mathbf{X}\mathbf{b}) \\
&= c - \frac{1}{2\sigma^2}(\mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b} - 2\mathbf{b}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y})
\end{aligned}
$$

- Taking derivative and set to 0, we have:

$$
\begin{aligned}
\frac{\partial LL}{\partial \mathbf{b}} &= -\frac{1}{2\sigma^2}(2\mathbf{X}^\top\mathbf{X}\mathbf{b} - 2\mathbf{X}^\top\mathbf{y}) \to 0 \\
\Rightarrow \mathbf{X}^\top\mathbf{X}\mathbf{b} &= \mathbf{X}^\top\mathbf{y} \\
\Rightarrow \hat{\mathbf{b}}_{MLE} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}
$$

- These are normal equations. If $\mathbf{X}^\top\mathbf{X}$ is invertible, the maximum likelihood estimator of $\mathbf{b}$ is exactly the same as the OLS of $\mathbf{b}$.

MLE cho hồi quy tuyến tính

LL = log (P(y X,b, σ 2))

= N/2 log(2πσ 2)(y 1 Xb)2σ 2 (y Xb)

= c 1/2σ 2 (b X Xb 2b X y + y y)

- Lấy đạo hàm đặt bằng 0, ta có:

LL/2 = 1/2σ (2X Xb 2X y) 0

b X Xb =

X y b^ MLE = (X X) 1X y

- Đây là những phương trình bình thường. Nếu X X là khả nghịch, ước lượng khả năng lớn nhất của b hoàn toàn giống với OLS của b.