TOÁN CHO KHOA HỌC MÁY TÍNH

ƯỚC LƯỢNG HỢP LÝ CỰC ĐẠI

TS. Lương Ngọc Hoàng

Nội dung

1. Ước lượng hợp lý cực đại
2. Ước lượng hợp lý cực đại với các phân phối xác suất
3. Ước lượng hợp lý cực đại cho Hồi quy tuyến tính
4. Quá khớp (overfitting), Điều chuẩn (regularization), và Ước lượng hậu nghiệm cực đại

ƯỚC LƯỢNG HỢP LÝ CỰC ĐẠI

MAXIMUM LIKELIHOOD ESTIMATION

Ví dụ

- Trong một túi mù có 3 viên bi. Mỗi viên bi có thể là màu đỏ hoặc màu xanh.
- Gọi θ là số viên bi màu xanh, θ có thể là 0, 1, 2, hoặc 3.
- Ta lấy ngẫu nhiên một viên bi ra khỏi túi, quan sát màu của viên bi, và trả lại viên bi vào trong túi. Ta thực hiện như vậy 4 lần (chọn ngẫu nhiên có thay thế).
- Ta định nghĩa 4 biến ngẫu nhiên X_1, X_2, X_3, X_4 như sau:

$$X_i = \begin{cases} 1, & \text{nếu viên bi thứ } i \text{ màu xanh} \\ 0, & \text{nếu viên bi thứ } i \text{ màu đỏ} \end{cases}$$

- Giả sử, sau thực nghiệm trên, các giá trị X_i mà ta quan sát được là: $x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1$.
- Các biến ngẫu nhiên X_i là *i.i.d.* (independent and identically distributed – độc lập và có cùng phân phối) và có thể được mô hình hóa theo phân phối Bernoulli $X_i \sim Bernoulli(\frac{\theta}{3})$.
- **Câu hỏi:** Giá trị θ bằng bao nhiêu thì khả năng ta quan sát được kết quả thực nghiệm trên là lớn nhất?

Ví dụ

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

- Vì các X_i là độc lập, hàm khối xác suất đồng thời (joint probability mass function) là:

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

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

θ	$P_{X_1X_2X_3X_4}(1,0,1,1)$
0	0
1	$(1/3)^3 \cdot (2/3) \approx 0.0247$
2	$(2/3)^3 \cdot (1/3) \approx 0.0988$
3	0

- Kết quả thực nghiệm mà ta quan sát được khả năng xảy ra cao nhất khi $\theta = 2$.
- Ta chọn $\hat{\theta} = 2$ là ước lượng (estimate) về giá trị của θ.

Ước lượng hợp lý cực đại

- Một quá trình quan trọng trong máy học là ước lượng giá trị các tham số $\boldsymbol{\theta}$ của mô hình trên tập dữ liệu D, thường được gọi là huấn luyện mô hình (model training), hay khớp mô hình (model fitting).
- Có nhiều phương pháp để ước lượng $\boldsymbol{\theta}$, và ta thường cần giải quyết bài toán tối ưu hoá sau:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

với $L(\boldsymbol{\theta})$ là hàm mất mát (loss), hay còn gọi là hàm mục tiêu (objective function).

- Một cách ước lượng tham số (parameter estimation) phổ biến chính là chọn bộ giá trị tham số $\boldsymbol{\theta}$ sao cho khả năng xảy ra (likelihood) của tập dữ liệu D là lớn nhất. Phương pháp này gọi là Ước lượng hợp lý cực đại (maximum likelihood estimation – MLE).

$$\widehat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta})$$

Ước lượng hợp lý cực đại

$$\widehat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta}} p\,(D \mid \boldsymbol{\theta})$$

- Ta thường giả sử các mẫu dữ liệu trong D là độc lập với nhau và có cùng phần phối xác suất (giả thiết **iid**). Hàm khả năng (likelihood function) trở thành:

$$p(D \mid \boldsymbol{\theta}) = p(y_1, y_2, \ldots, y_N \mid x_1, x_2, \ldots, x_N; \boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i \mid x_i; \boldsymbol{\theta})$$

- Để thuận tiện trong việc xử lý và tính toán, ta thường sử dụng hàm log likelihood để phân rã thành tổng các vế, mỗi vế ứng với một mẫu dữ liệu.

$$LL(\boldsymbol{\theta}) = \log p(D \mid \boldsymbol{\theta}) = \log \prod_{i=1}^{N} p(y_i \mid x_i; \boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(y_i \mid x_i; \boldsymbol{\theta})$$

Ước lượng hợp lý cực đại

- Ước lượng hợp lý cực đại (MLE) được biến đổi thành:

$$\widehat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta}} p\,(D \mid \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p(y_i \mid x_i; \boldsymbol{\theta})$$

- Do các thuật toán tối ưu hóa trong các công cụ máy học thường được cài đặt để cực tiểu hóa hàm mất mát, hàm mục tiêu trên có thể được định nghĩa lại thành hàm negative log likelihood (NLL) như sau:

$$NLL(\boldsymbol{\theta}) = -\log p(D \mid \boldsymbol{\theta}) = -\sum_{i=1}^{N} \log p(y_i \mid x_i; \boldsymbol{\theta})$$

- Cực tiểu hóa hàm NLL tương đương với MLE:

$$\widehat{\boldsymbol{\theta}}_{\text{mle}} = \arg\min_{\boldsymbol{\theta}} NLL(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^{N} \log p(y_i \mid x_i; \boldsymbol{\theta})$$

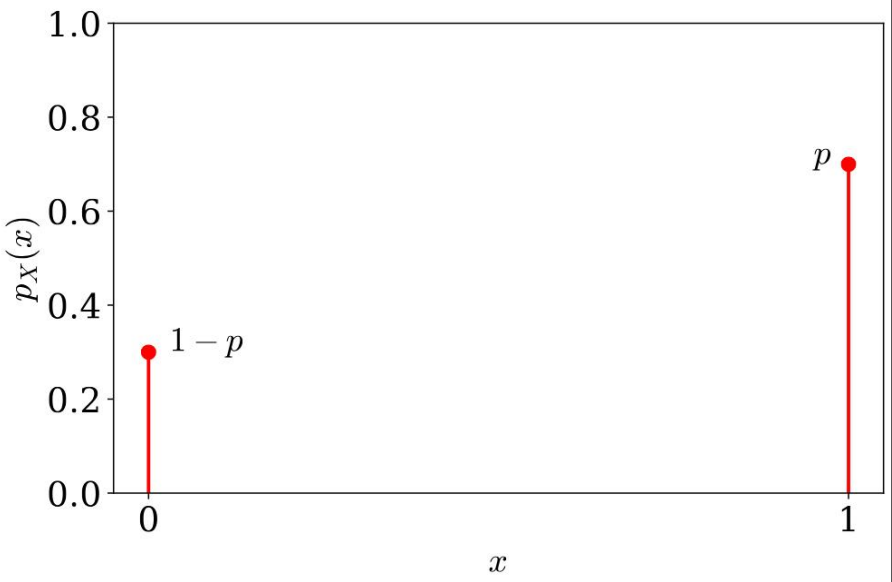
MLE VỚI CÁC PHÂN PHỐI XÁC SUẤT

MLE FOR PROBABILITY DISTRIBUTIONS

MLE cho phân phối Bernoulli

- Một biến ngẫu nhiên (random variable) Bernoulli nhận một trong hai giá trị, có thể ký hiệu là 0 và 1, mô hình hóa một thực nghiệm có yếu tố ngẫu nhiên với hai kết cục có thể xảy ra (thường được gọi là "*thành công*" hoặc "*thất bại*"). Ví dụ:
  - Tung một đồng xu. Kết cục có thể là mặt ngửa (heads) hoặc mặt xấp (tails).
  - Tham gia một kỳ thi. Kết cục có thể là đậu (pass) hoặc rớt (fail).
  - Phân lớp một bức hình. Một bức hình có thể được phân lớp thành lớp mèo (cat) hoặc lớp không chứa mèo (non-cat).
- Một biến ngẫu nhiên X là một biến ngẫu nhiên Bernoulli với tham số $p \in [0,1]$, biểu diễn với $X \sim Bernoulli(p)$, nếu hàm khối xác suất của nó có dạng:

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



MLE cho phân phối Bernoulli

- Giả sử Y là một biến ngẫu nhiên biểu diễn kết quả khi tung một đồng xu.
- Biến cố (event) $Y = 1$ tương ứng với trường hợp mặt ngửa (heads), và biến cố $Y = 0$ tương ứng với trường hợp mặt xấp (tails).
- Biến ngẫu nhiên này tuân theo phân phối Bernoulli. Hàm negative log likelihood (NLL) cho tham số θ của phân phối Bernoulli với kết quả từ N lần tung đồng xu là:

$$NLL(\theta) = -\log \prod_{i=1}^{N} p(y_i|\theta) = -\log \prod_{i=1}^{N} \theta^{\mathbb{I}(y_i=1)}(1-\theta)^{\mathbb{I}(y_i=0)}$$

$$= -\sum_{i=1}^{N} [\mathbb{I}(y_i = 1)\log\theta + \mathbb{I}(y_i = 0)\log(1-\theta)]$$

$$= -[N_1 \log\theta + N_0 \log(1-\theta)]$$

với $N_1 = \sum_{i=1}^{N} \mathbb{I}(y_i = 1)$ là số lần mặt ngửa (heads) xuất hiện và $N_0 = \sum_{i=1}^{N} \mathbb{I}(y_i = 0)$ là số lần mặt xấp (tails) xuất hiện.

- $N = N_0 + N_1$ là kích thước mẫu (sample size).

MLE cho phân phối Bernoulli

$$NLL(\theta) = -\left[N_1 \log\theta + N_0 \log(1-\theta)\right]$$

Đạo hàm của hàm NLL theo tham số θ là:

$$\frac{d}{d\theta} NLL(\theta) = -\frac{N_1}{\theta} + \frac{N_0}{1-\theta}$$

Ước lượng hợp lý cực đại (MLE) cho θ có thể được tính bằng cách giải $\frac{d}{d\theta} NLL(\theta) = 0$.

$$-\frac{N_1}{\theta} + \frac{N_0}{1-\theta} = 0$$
$$N_0\theta - N_1 + N_1\theta = 0$$
$$\theta(N_0 + N_1) = N_1$$

Ước lượng hợp lý cực đại (MLE) được tính bởi:

$$\hat{\theta}_{\text{MLE}} = \frac{N_1}{N_0 + N_1}$$

Chính là số lần mặt ngửa (heads) xuất hiện trong thực nghiệm tung đồng xu N lần.

MLE cho phân phối phân loại (categorical)

- Giả sử ta tung một xúc xắc có K mặt N lần.
- Gọi biến ngẫu nhiên $Y_i \in \{1, ..., K\}$ biểu diễn kết cục của lần tung thứ i. Các biến ngẫu nhiên $Y_i \sim \text{Categorical}(\boldsymbol{\theta})$ tuân theo phân phối phân loại (categorical distribution).
- Ta muốn ước lượng các tham số $\boldsymbol{\theta}$ của phân phối dựa vào tập dữ liệu $D = \{y_i \mid i = 1{:}\, N\}$ là kết cục của N lần tung.

- Hàm negative log likelihood (NLL) của $\boldsymbol{\theta}$ là:

$$NLL(\boldsymbol{\theta}) = -\sum_{\boldsymbol{k}} N_k \log \theta_k$$

với N_k là số lần mà biến cố $Y = k$ được quan sát (số lần mặt thứ k xuất hiện).

- Để tính giá trị của ước lượng hợp lý cực đại, ta cần cực tiểu hóa NLL với ràng buộc là

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

MLE cho phân phối phân loại (categorical)

- Ta sử dụng phương pháp nhân tử Lagrange (Lagrange multipliers) như sau:

$$L(\theta, \lambda) = -\sum_k N_k \log \theta_k - \lambda \left(1 - \sum_k \theta_k\right)$$

- Lấy đạo hàm của $L(\theta, \lambda)$ theo λ, và đặt đạo hàm này bằng 0 sẽ cho ta ràng buộc:

$$\frac{\partial L}{\partial \lambda} = 1 - \sum_k \theta_k = 0$$

- Lấy đạo hàm theo của $L(\theta, \lambda)$ theo các tham số θ_k, ta được:

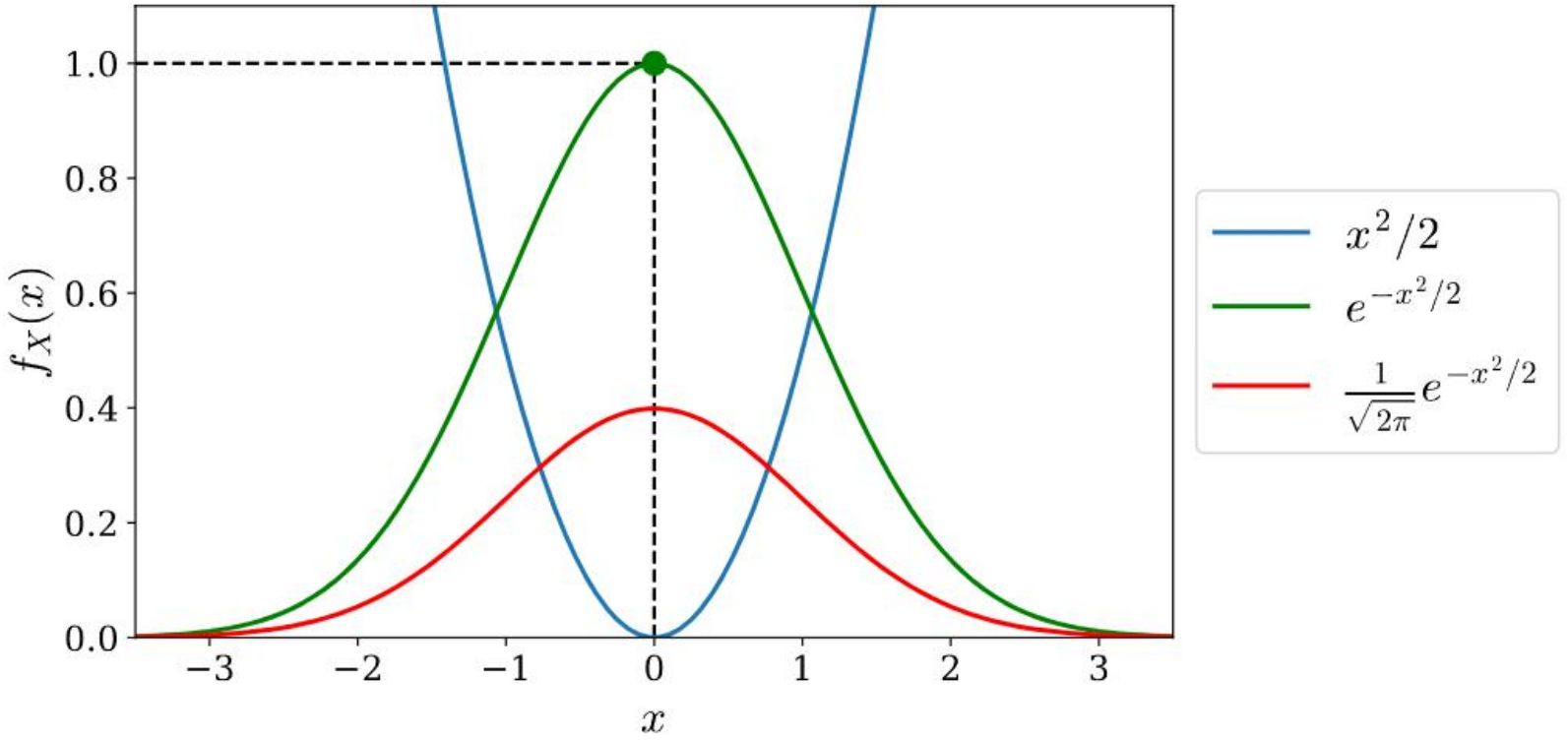
$$\frac{\partial L}{\partial \theta_k} = -\frac{N_k}{\theta_k} + \lambda = 0 \rightarrow N_k = \lambda \theta_k \rightarrow \theta_k = \frac{N_k}{\lambda}$$

- Ta có thể tính giá trị của λ dựa vào ràng buộc tổng các xác suất bằng 1:

$$N = \sum_k N_k = \sum_k \lambda \theta_k = \lambda \sum_k \theta_k = \lambda$$

- Do đó, ước lượng hợp lý cực đại (MLE) của các tham số θ_k là: $\hat{\theta}_k = \frac{N_k}{\lambda} = \frac{N_k}{N}$, chính là số lần mặt k xuất hiện trong N lần tung xúc xắc.

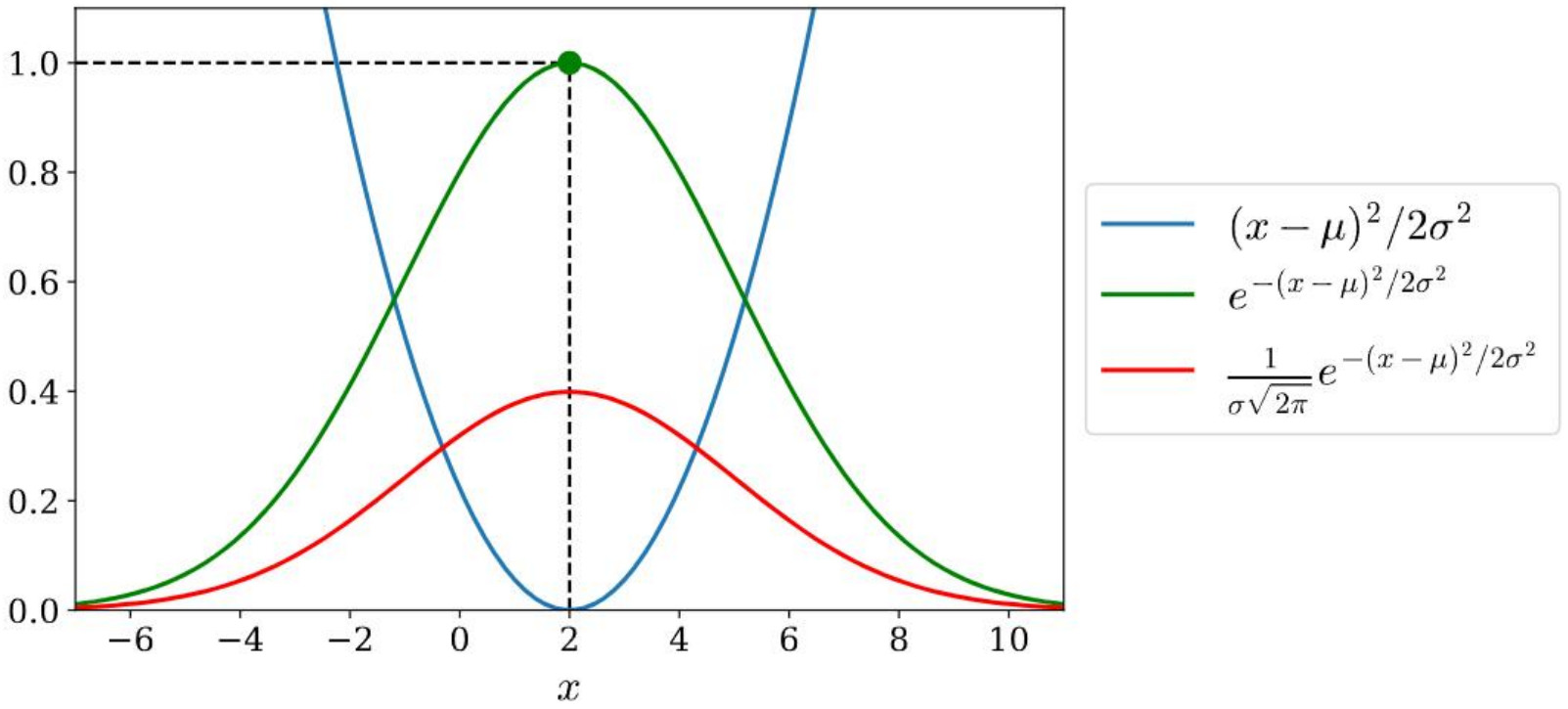
Phân phối chuẩn tắc (standard normal) $N(0,1)$



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

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

Phân phối chuẩn (normal distribution) $N(\mu, \sigma^2)$



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

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

$$E[X] = \mu, \quad Var(X) = \sigma^2$$

MLE cho phân phối chuẩn (Gaussian)

- Biến ngẫu nhiên chuẩn $Y \sim N(\mu, \sigma^2)$ và $D = \{y_i \mid i = 1{:}N\}$ là một tập dữ liệu với N mẫu dữ liệu độc lập và có cùng phân phối.

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

- Ta có thể ước lượng giá trị các tham số $\boldsymbol{\theta} = (\mu, \sigma^2)$ sử dụng MLE.
- Hàm negative log likelihood (NLL) của các tham số là:

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

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

- Lời giải cực tiểu của hàm NLL cần thỏa mãn điều kiện sau:

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

MLE cho phân phối chuẩn (Gaussian)

- Ước lượng hợp lý cực đại cho μ và σ^2 là:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{\mu}_{\text{MLE}})^2 = \frac{1}{N}\sum_{i=1}^{N} \left(y_i^2 + \hat{\mu}^2_{\text{MLE}} - 2y_i\hat{\mu}_{\text{MLE}}\right) = s^2 - \bar{y}^2$$

$$s^2 \triangleq \frac{1}{N}\sum_{i=1}^{N} y_i^2$$

- Ta có thể thấy phương sai (variance) σ^2 được ước lượng với một công thức khác:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \hat{\mu}_{\text{MLE}})^2$$

Công thức trên là một cách ước lượng khác, nhưng không phải là ước lượng hợp lý cực đại (MLE).

MLE cho phân phối chuẩn – Ví dụ

- Ta có $N = 3$ điểm dữ liệu $y_1 = 1, y_2 = 0.5, y_3 = 1.5$ độc lập với nhau và cùng được phát sinh từ một phân phối chuẩn. Ta biết phân phối này có phương sai (variance) là 1 nhưng chưa biết được kỳ vọng (mean) μ.

$$y_i \sim N(\mu, 1)$$

- Hàm khả năng (likelihood):

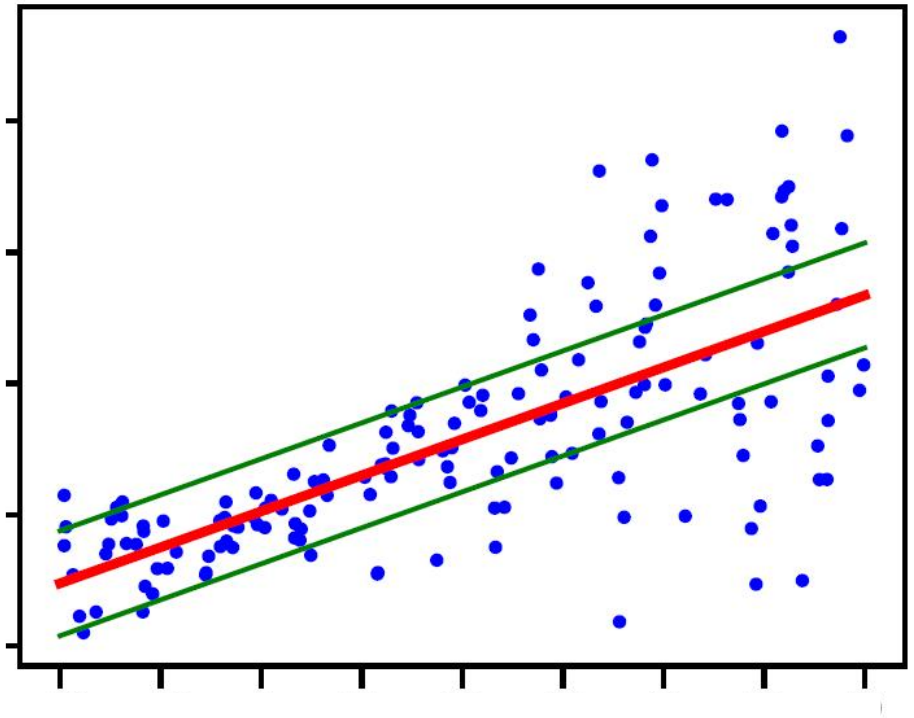
$$P(y_1, y_2, y_3 \mid \mu) = P(y_1 \mid \mu)P(y_2 \mid \mu)P(y_3 \mid \mu)$$

- Ta xét 2 trường hợp $\mu = 1.0$ và $\mu = 2.5$. Trường hợp nào có khả năng cao hơn?
- Ta cần tìm giá trị μ để cực đại hóa hàm khả năng: $P(y_1 \mid \mu)P(y_2 \mid \mu)P(y_3 \mid \mu)$ được cực đại hóa.

MLE CHO HỒI QUY TUYẾN TÍNH

MLE FOR LINEAR REGRESSION

Ví dụ



- Dự đoán điểm Toán cho KHMT của một sinh viên dựa vào điểm Đại số tuyến tính của sinh viên đó.
- Ta thu thập dữ liệu các sinh viên khóa trước. Mỗi sinh viên ứng với một điểm dữ liệu (màu xanh).
- ***Trục hoành***: Điểm môn Đại số tuyến tính. ***Trục tung***: Điểm môn Toán cho KHMT.

MLE cho hồi quy tuyến tính

- Ta có thể cài đặt các tham số của phân phối chuẩn thành dạng các hàm số với các đặc trưng đầu vào của dữ liệu.

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = N(y \mid f_\mu(\boldsymbol{x}; \boldsymbol{\theta}),\ f_\sigma(\boldsymbol{x}; \boldsymbol{\theta})^2)$$

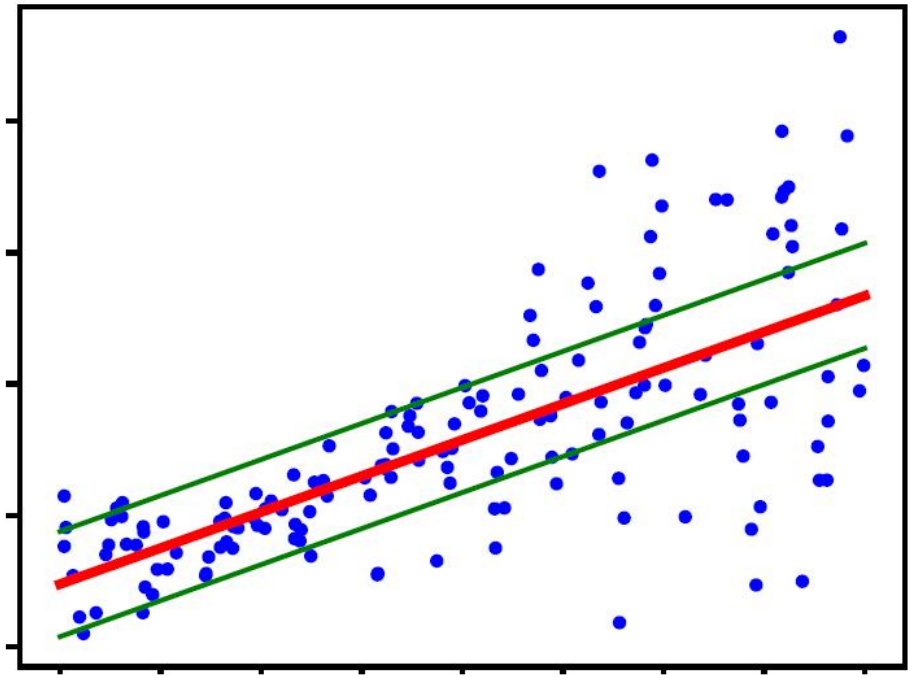
với $f_\mu(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathbb{R}$ dự đoán giá trị kỳ vọng (mean), và $f_\sigma(\boldsymbol{x}; \boldsymbol{\theta})^2 \in \mathbb{R}_+$ dự đoán phương sai (variance).

- Để đơn giản, ta có thể giả định phương sai (variance) có giá trị cố định (fixed) và độc lập với các đặc trưng đầu vào $\boldsymbol{x}$. Do đó, ta chỉ cần dự đoán giá trị kỳ vọng.
- Ta cũng có thể giả định kỳ vọng (mean) là một **hàm tuyến tính** của các đặc trưng đầu vào. Mô hình dự đoán này được gọi là hồi quy tuyến tính (linear regression):

$$p(y \mid \boldsymbol{x};\ \boldsymbol{\theta}) = N(y \mid \boldsymbol{w}^T\boldsymbol{x} + b,\ \sigma^2)$$

với $\boldsymbol{\theta} = (\boldsymbol{w}, b, \sigma^2)$.

MLE cho hồi quy tuyến tính



Hồi quy tuyến tính dự đoán giá trị đích tuân phân phối chuẩn với kỳ vọng (mean) $\mu(x) = w^T x + b$ và phương sai (variance) σ^2.

- Hình vẽ minh họa khoảng dự đoán $[\mu(x) - 2\sigma, \mu(x) + 2\sigma]$.
- Khoảng giá trị này thể hiện sự không chắc chắn (uncertainty) trong việc dự đoán quan sát y với giá trị đặc trưng đầu vào x, thể hiện sự biến thiên của các điểm dữ liệu (màu xanh).

MLE cho hồi quy tuyến tính

- Mô hình hồi quy tuyến tính

$$p(y \mid \boldsymbol{x};\ \boldsymbol{\theta}) = N(y \mid \boldsymbol{w}^T \boldsymbol{x}\ , \sigma^2)$$

với $\boldsymbol{\theta} = (\boldsymbol{w}, \sigma^2)$ và $\boldsymbol{w} = (b, w_1, w_2, ..., w_D), \boldsymbol{x} = (1, x_1, x_2, ..., x_D)$.

- Giả sử phương sai σ^2 là cố định, ta chỉ cần ước lượng giá trị các tham số $\boldsymbol{w}$. Hàm negative log likelihood (NLL) của $\boldsymbol{w}$ là:

$$NLL(\boldsymbol{w}) = -\sum_{i=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2 \right) \right]$$

- Sau khi đơn giản biểu thức trên và loại bỏ các hằng số không quan trọng, ta có công thức tương đương với $\mathrm{NLL}(\mathbf{w})$ chính là hàm tổng phần dư bình phương (residual sum of squares – RSS):

$$RSS(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2 = \sum_{i=1}^{N} r_i^2$$

với r_i là sai số phần dư của điểm dữ liệu thứ i.

MLE cho hồi quy tuyến tính

- Tổng phần dư bình phương (residual sum of squares – RSS):

$$RSS(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

- Sai số bình phương trung bình (mean squared error – MSE):

$$MSE(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

- Căn của sai số bình phương trung bình (root mean squared error – RMSE):

$$RMSE(\boldsymbol{w}) = \sqrt{MSE(\boldsymbol{w})} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2}$$

- Ta có thể tính ước lượng hợp lý cực đại (MLE) bằng cách cực tiểu hóa các hàm NLL, RSS, MSE, hoặc RMSE. Tất cả đều cho cùng lời giải.

MLE cho hồi quy tuyến tính

- Hàm tổng phần dư bình phương (RSS) có thể được viết thành dạng ma trận – vector:

$$RSS(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 = (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

- Gradient của RSS là:

$$\nabla_{\boldsymbol{w}} RSS(\boldsymbol{w}) = \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{y}$$

- Đặt gradient $\nabla_{\boldsymbol{w}} RSS(\boldsymbol{w}) = \boldsymbol{0}$ và giải, ta được:

$$\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T \boldsymbol{y}$$

- Đây được gọi là Normal Equations.
- Lời giải của ước lượng hợp lý cực đại (MLE) còn được gọi là lời giải bình phương tối thiểu thông thường (ordinary least squares):

MLE cho hồi quy tuyến tính

$$\hat{\boldsymbol{w}}_{\text{mle}} = \arg\min_{\boldsymbol{w}} RSS(\boldsymbol{w}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- Đại lượng $\boldsymbol{X}^\dagger = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ gọi là giả nghịch đảo (pseudo-inverse) của ma trận $\boldsymbol{X}$.
- Lời giải ước lượng hợp lý cực đại $\hat{\boldsymbol{w}}_{\text{mle}}$ có là duy nhất (unique)?
- Gradient của RSS là $\nabla_{\boldsymbol{w}} RSS(\boldsymbol{w}) = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}$ thì Hessian sẽ là:

$$H(w) = \frac{\partial^2}{\partial w^2} RSS(\boldsymbol{w}) = \boldsymbol{X}^T\boldsymbol{X}$$

- Nếu $\boldsymbol{X}$ là ma trận có hạng đầy đủ (full rank – các cột của X độc lập tuyến tính) thì H là ma trận xác định dương, vì với $\boldsymbol{v} \neq \boldsymbol{0}$ bất kỳ, ta có:

$$\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} = (\boldsymbol{X}\boldsymbol{v})^T(\boldsymbol{X}\boldsymbol{v}) = \|\boldsymbol{X}\boldsymbol{v}\|^2 > 0.$$

- Trường hợp $\boldsymbol{X}$ có hạng đầy đủ, $RSS(\boldsymbol{w})$ có một lời giải cực trị duy nhất.

ƯỚC LƯỢNG HẬU NGHIỆM CỰC ĐẠI

MAXIMUM A POSTERIORI ESTIMATION

Quá khớp (overfitting)

- MLE chọn các giá trị tham số cực tiểu hóa hàm mất mát trên tập dữ liệu huấn luyện. Tuy nhiên, điều này không đảm bảo mô hình sẽ hoạt động tốt trên dữ liệu tổng quát. Đây là vấn đề mô hình bị quá khớp vào dữ liệu huấn luyện.
- **Ví dụ:** ước lượng xác suất đồng xu ra mặt ngửa (heads) khi tung một đồng xu.
- Ta tung đồng xu $N = 3$ lần, và cả 3 lần mặt ngửa đều xuất hiện. Ước lượng hợp lý cực đại (MLE) sẽ là:

$$\hat{\theta}_{\text{MLE}} = \frac{N_1}{N_0 + N_1} = \frac{3}{0 + 3} = 1$$

- Nếu ta sử dụng phân phối $Bernoulli(y \mid \hat{\theta}_{\text{MLE}})$ này thì ta sẽ dự đoán đồng xu này sẽ luôn xuất hiện mặt ngửa trong tất cả các lần tung trong tương lai.
- Nếu mô hình (model) có đủ số tham số để khớp hoàn hảo (perfectly fit) vào dữ liệu huấn luyện quan sát được, nó sẽ trùng với phân phối thực nghiệm (empirical distribution).
- Tuy nhiên, phân phối thực nghiệm thường không giống hoàn toàn với phân phối thực (true distribution). Khi ta đặt toàn bộ phân phối vào tập dữ liệu huấn luyện gồm N điểm dữ liệu mà ta thu thập được, mô hình có thể không tổng quát hóa được với các điểm dữ liệu mới không có trong tập huấn luyện.

Ví dụ: MLE cho hồi quy tuyến tính

Ví dụ 1:

- Dữ liệu huấn luyện: $x_1 = \begin{bmatrix}1\\0\end{bmatrix},\ y_1 = 1$ và $x_2 = \begin{bmatrix}1\\\varepsilon\end{bmatrix},\ y_2 = 1$
- $\boldsymbol{X} = \begin{bmatrix}1 & 0\\1 & \varepsilon\end{bmatrix},\ \boldsymbol{y} = \begin{bmatrix}1\\1\end{bmatrix}$
- $\hat{\boldsymbol{w}}_{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- $\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix}1 & 1\\0 & \varepsilon\end{bmatrix}\begin{bmatrix}1 & 0\\1 & \varepsilon\end{bmatrix} = \begin{bmatrix}2 & \varepsilon\\\varepsilon & \varepsilon^2\end{bmatrix}$
- $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix}1 & -1/\varepsilon\\-1/\varepsilon & 2/\varepsilon^2\end{bmatrix}$
- $\hat{\boldsymbol{w}}_{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix}1 & -1/\varepsilon\\-1/\varepsilon & 2/\varepsilon^2\end{bmatrix}\begin{bmatrix}1 & 1\\0 & \varepsilon\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\0\end{bmatrix}$

Ví dụ: MLE cho hồi quy tuyến tính

Ví dụ 2:

- Dữ liệu huấn luyện: $x_1 = \begin{bmatrix}1\\0\end{bmatrix},\ y_1 = 1+\varepsilon$ và $x_2 = \begin{bmatrix}1\\\varepsilon\end{bmatrix},\ y_2 = 1$
- $\boldsymbol{X} = \begin{bmatrix}1 & 0\\1 & \varepsilon\end{bmatrix},\ \boldsymbol{y} = \begin{bmatrix}1+\varepsilon\\1\end{bmatrix}$
- $\hat{\boldsymbol{w}}_{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- $\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix}1 & 1\\0 & \varepsilon\end{bmatrix}\begin{bmatrix}1 & 0\\1 & \varepsilon\end{bmatrix} = \begin{bmatrix}2 & \varepsilon\\\varepsilon & \varepsilon^2\end{bmatrix}$
- $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix}1 & -1/\varepsilon\\-1/\varepsilon & 2/\varepsilon^2\end{bmatrix}$
- $\hat{\boldsymbol{w}}_{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix}1 & -1/\varepsilon\\-1/\varepsilon & 2/\varepsilon^2\end{bmatrix}\begin{bmatrix}1 & 1\\0 & \varepsilon\end{bmatrix}\begin{bmatrix}1+\varepsilon\\1\end{bmatrix} = \begin{bmatrix}1+\varepsilon\\-1\end{bmatrix}$

Điều chuẩn (regularization)

- Một giải pháp thường được sử dụng để giảm hiện tượng quá khớp (overfitting) là các kỹ thuật điều chuẩn (regularization).
- Ta thêm một vế phạt (penalty) vào hàm negative log likelihood (NLL):

$$L(\boldsymbol{\theta}, \lambda) = \left[\frac{1}{N}\sum_{i=1}^{N} l(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}))\right] + \lambda C(\boldsymbol{\theta})$$

với $\lambda \geq 0$ là tham số điều chuẩn (regularization parameter), và $C(\boldsymbol{\theta})$ là một hàm phạt độ phức tạp (complexity penalty) phù hợp.

- Một hàm phạt độ phức tạp có thể dùng là $C(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$ với $p(\boldsymbol{\theta})$ là phân phối xác suất tiên nghiệm (prior) của $\boldsymbol{\theta}$.
- Hàm mục tiêu được điều chuẩn là:

$$L(\boldsymbol{\theta}, \lambda) = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \lambda \log p(\boldsymbol{\theta})$$

Ước lượng hậu nghiệm cực đại (MAP)

$$L(\boldsymbol{\theta}, \lambda) = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \lambda \log p(\boldsymbol{\theta})$$

- Thay đổi giá trị λ sao cho phù hợp, ta có thể biểu diễn lại hàm mục tiêu trên thành:

$$L(\boldsymbol{\theta}, \lambda) = -\left[\sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right] = -\left[\log p(\boldsymbol{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right]$$

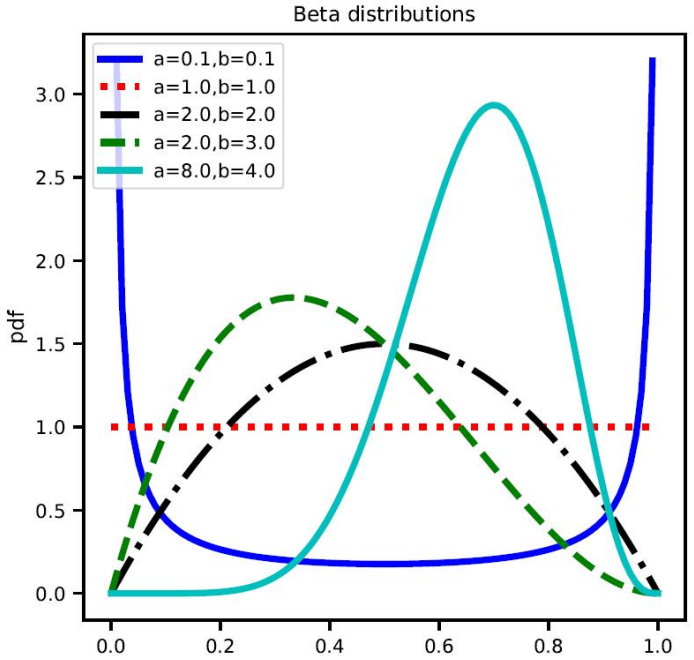
- Cực tiểu hóa hàm mục tiêu trên tương đương với cực đại hóa log phân phối xác suất hậu nghiệm (log posterior):

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{D}) = \arg\max_{\boldsymbol{\theta}} \log \frac{p(\boldsymbol{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{D})}$$
$$= \arg\max_{\boldsymbol{\theta}} \left[\log p(\boldsymbol{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \mathbf{const}\right]$$

- Công thức ước lượng trên được gọi là Ước lượng hậu nghiệm cực đại (maximum a posteriori estimation – MAP).

Ước lượng MAP cho phân phối Bernoulli

- Nếu ta tung đồng xu chỉ 1 lần duy nhất, và quan sát được mặt ngửa (heads). Thì ước lượng hợp lý cực đại (MLE) sẽ là $\widehat{\boldsymbol{\theta}}_{MLE} = 1$.
- Để tránh các giá trị cực đoan như $\theta = 0$ hoặc $\theta = 1$, ta có thể sử dụng các hàm phạt. Ta có thể sử dụng phân phối Beta như là phân phối tiên nghiệm (prior) $p(\theta) = \text{Beta}(\theta\,|a, b)$ với $a, b > 1$ sẽ khuyến khích các giá trị θ gần với $a/(a + b)$.



- Nếu $a = b = 1$, ta có phân phối đều (uniform distribution).
- Nếu a và b cùng nhỏ hơn 1, ta có phân phối đôi đỉnh (bimodal).
- Nếu a và b cùng lớn hơn 1, ta có phân phối đơn đỉnh (unimodal).

Ước lượng MAP cho phân phối Bernoulli

- Sử dụng phân phối Beta làm phân phối xác suất tiên nghiệm (prior) $p(\theta) = \text{Beta}(\theta \mid a, b)$, hàm log likelihood + log prior trở thành:

$$LL(\theta) = log\, p(D \mid \theta) + log\, p(\theta)$$
$$= [N_1 log\, \theta + N_0 log\, (1-\theta)] + [(a-1) log\, \theta + (b-1) log\, (1-\theta)]$$

- Ước lượng hậu nghiệm cực đại (MAP) là:

$$\hat{\theta}_{MAP} = \frac{N_1 + a - 1}{N_1 + N_0 + a + b - 2}$$

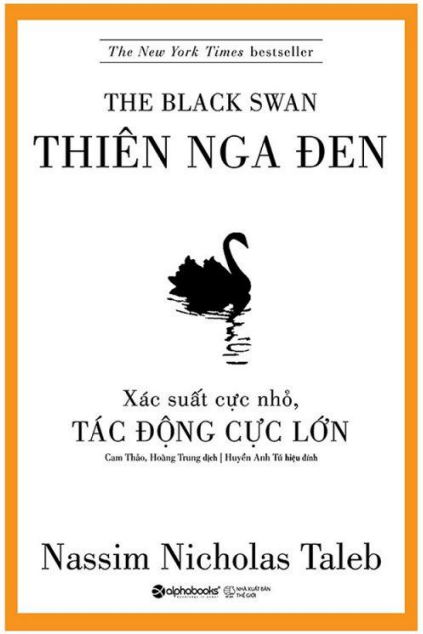
- Nếu ta đặt $a = b = 2$ để khuyến khích các giá trị θ gần với 0.5 thì ta có:

$$\hat{\theta}_{MAP} = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

- Đây gọi là kỹ thuật add-one smoothing để giải quyết vấn đề zero count.

Nghịch lý thiên nga đen (black swan paradox)

- Vấn đề **zero count**, và **quá khớp** (overfitting), tương tự như **nghịch lý thiên nga đen**.
- **Phép quy nạp (induction)**: làm cách nào để rút ra các kết luận tổng quát về tương lai dựa trên các quan sát cụ thể trong quá khứ?
- Để có thể dự đoán một cách hợp lý, ta cần kết hợp thông tin rút ra được từ **dữ liệu thực nghiệm** (empirical data) và **tri thức tiên nghiệm** (prior knowledge).



Giảm giá trị trọng số (weight decay)

- Hồi quy đa thức với quá nhiều đặc trưng bậc cao sẽ rất linh hoạt (flexible) để khớp tốt vào dữ liệu huấn luyện, có thể dẫn đến hiện tượng quá khớp (overfitting). Một giải pháp chính là phạt độ lớn (magnitude) các giá trị trọng số (regression coefficients).
- Ta sử dụng một phân phối chuẩn có kỳ vọng 0 làm phân phối xác suất tiên nghiệm (prior) $p(\boldsymbol{w})$. Hàm phạt $C(\boldsymbol{w}) = -\log p(\boldsymbol{w})$. Ước lượng hậu nghiệm cực đại sẽ là:

$$\hat{\boldsymbol{w}}_{MAP} = \arg\min_{\boldsymbol{w}} NLL(\boldsymbol{w}) + \boldsymbol{\lambda}\|\boldsymbol{w}\|_2^2$$

với $\|\boldsymbol{w}\|_2^2 = \sum_{d=1}^{D} w_d^2$. Ta thường không phạt giá trị bias b (hoặc w_0).

- Biểu thức trên được gọi là L_2 regularization hoặc **weight decay**.
- Giá trị của λ càng lớn, các tham số càng bị phạt nặng hơn khi giá trị tham số tăng lên (chính là khi càng rời xa phân phối tiên nghiệm có kỳ vọng 0). Do đó, mô hình sẽ trở nên ít linh hoạt hơn (less flexible).

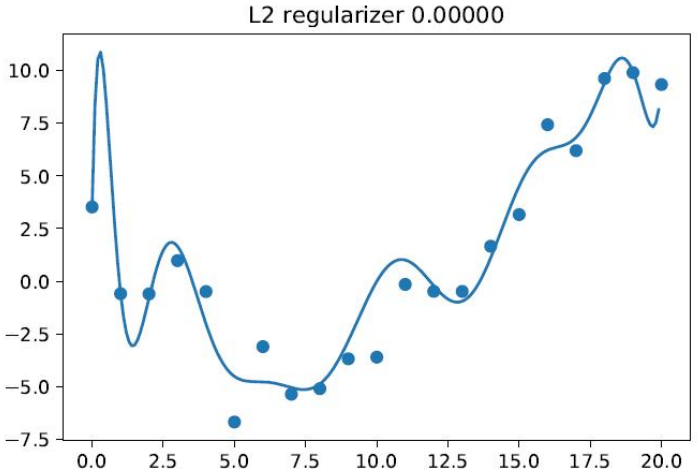
Hồi quy Ridge

- Phương pháp **weight decay** áp dụng trong trường hợp hồi quy tuyến tính (linear regression) còn được gọi là hồi quy ridge (ridge regression).
- Xét mô hình hồi quy đa thức (polynomial regression) như sau:

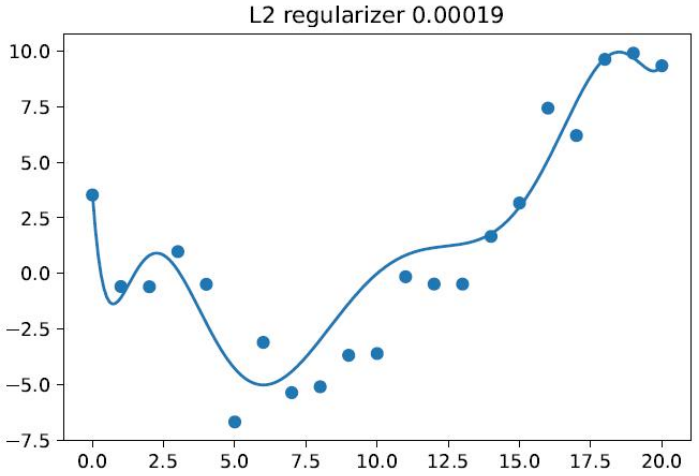
$$f(x; \boldsymbol{w}) = \sum_{d=0}^{D} w_d x^d = \boldsymbol{w}^T [1, x, x^2, ..., x^D]$$

- Giả sử ta sử dụng các đặc trưng bậc cao, ví dụ $D = 14$, mặc dù tập dữ liệu của chúng ta tương đối nhỏ, ví dụ $N = 21$ điểm dữ liệu.
- Áp dụng ước lượng hợp lý cực đại (MLE), chúng ta sẽ thu được một bộ các giá trị tham số mô hình khớp rất tốt vào tập dữ liệu, và mô hình này sẽ rất phức tạp, dẫn đến hiện tượng quá khớp (overfitting).
- Sử dụng **weight decay** giúp ta giảm thiểu nguy cơ mô hình bị quá khớp.

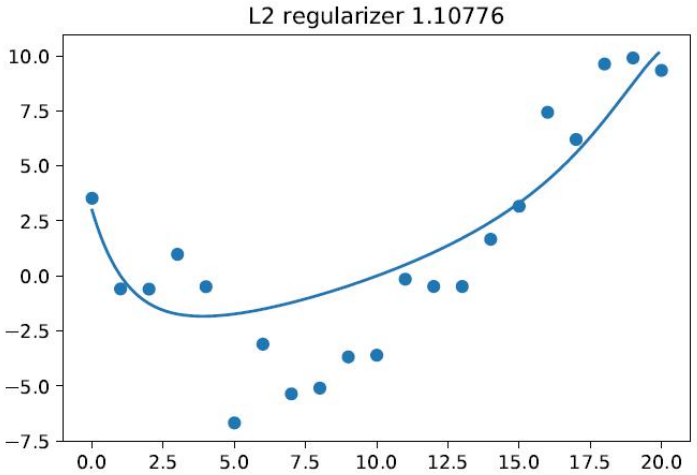
Hồi quy Ridge



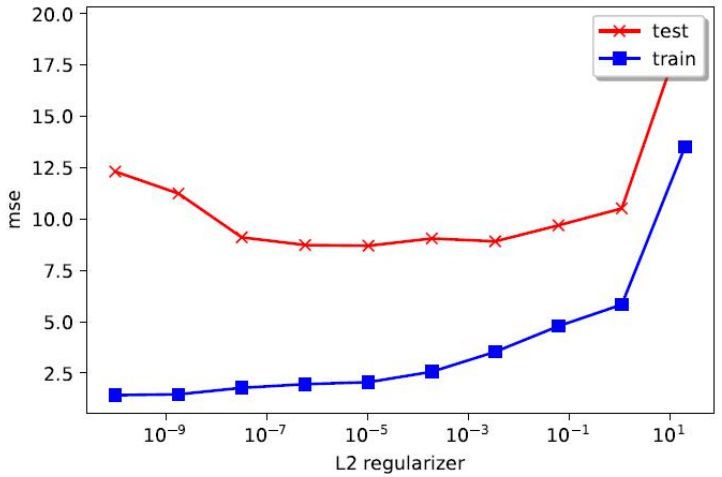
(a)



(b)



(c)



(d)

Hồi quy Ridge

- Ước lượng hậu nghiệm cực đại (MAP) cho hồi quy tuyến tính tương ứng với việc cực tiểu hóa hàm mục tiêu có thành phần phạt sau đây:

$$J(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2$$

với λ thể hiện mức độ điều chuẩn (regularization strength).

- Đạo hàm của hàm mục tiêu trên là:

$$\nabla_w J(\boldsymbol{w}) = 2(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y} + \lambda\boldsymbol{w})$$

- Do đó, ta có ước lượng hậu nghiệm cực đại (MAP) là:

$$\widehat{\boldsymbol{w}}_{MAP} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$= \left(\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T + \lambda\mathbf{I}\right)^{-1}\left(\sum_i y_i\boldsymbol{x}_i\right)$$

MLE và MAP cho hồi quy tuyến tính – Ví dụ

➢ **Ước lượng hợp lý cực đại (MLE).** Giả sử $\varepsilon = 0.1$

- **Ví dụ 1:** $\widehat{\boldsymbol{w}}_{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} 1 & -1/\varepsilon \\ -1/\varepsilon & 2/\varepsilon^2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & \varepsilon \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

- **Ví dụ 2:** $\widehat{\boldsymbol{w}}_{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} 1 & -1/\varepsilon \\ -1/\varepsilon & 2/\varepsilon^2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & \varepsilon \end{bmatrix}\begin{bmatrix} 1+\varepsilon \\ 1 \end{bmatrix} = \begin{bmatrix} 1+\varepsilon \\ -1 \end{bmatrix} = \begin{bmatrix} 1.1 \\ -1 \end{bmatrix}$

➢ **Ước lượng hậu nghiệm cực đại (MAP)** với $\lambda = 0.05$.

- $\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I} = \begin{bmatrix} 2+\lambda & \varepsilon \\ \varepsilon & \varepsilon^2+\lambda \end{bmatrix} = \begin{bmatrix} 2.05 & 0.1 \\ 0.1 & 0.06 \end{bmatrix}$
- $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I})^{-1} = \begin{bmatrix} 0.531 & -0.885 \\ -0.885 & 18.1416 \end{bmatrix}$

- **Ví dụ 1**: $\widehat{\boldsymbol{w}}_{MAP} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} 0.531 & -0.885 \\ -0.885 & 18.1416 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & \varepsilon \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.9735 \\ 0.0442 \end{bmatrix}$

- **Ví dụ 2**: $\widehat{\boldsymbol{w}}_{MAP} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} 0.531 & -0.885 \\ -0.885 & 18.1416 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & \varepsilon \end{bmatrix}\begin{bmatrix} 1+\varepsilon \\ 1 \end{bmatrix} = \begin{bmatrix} 1.0265 \\ -0.0442 \end{bmatrix}$