Linear Regression via Maximum Likelihood Estimation

Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

April 6, 2023

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?

Bernoulli Random Variables

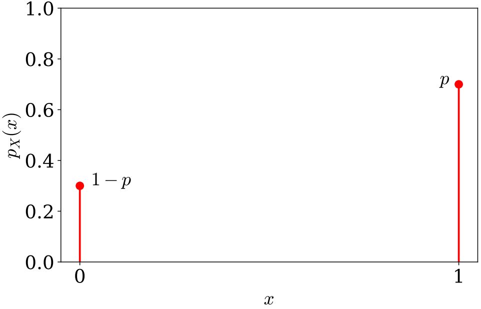
- A Bernoulli random variable X takes two possible values, usually 0 and 1, modeling random experiments that have two possible outcomes (e.g., "success" and "failure").
  - e.g., tossing a coin. The outcome is either Head or Tail.
  - e.g., taking an exam. The result is either Pass or Fail.
  - e.g.,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

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

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



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?

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.

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

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

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**.

$$\hat{\mathbf{b}}_{\texttt{mle}} = \operatorname*{argmax}_{\mathbf{b}} p(\mathcal{D} \mid \mathbf{b})$$

- 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.

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

Maximum Likelihood Estimation

- The MLE is given by

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

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

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

- Minimizing this will give the MLE.

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

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

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

  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 for the Bernoulli distribution

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

- The derivative of the NLL is

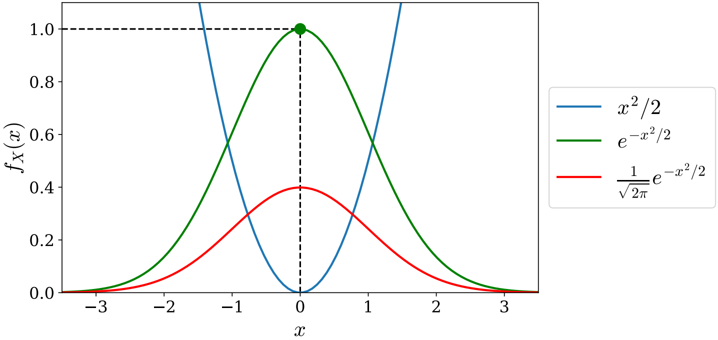
$$\frac{d}{db}\texttt{NLL}(b) = \frac{-N_1}{b} + \frac{N_0}{1-b}$$

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

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

which is the empirical fraction of heads.

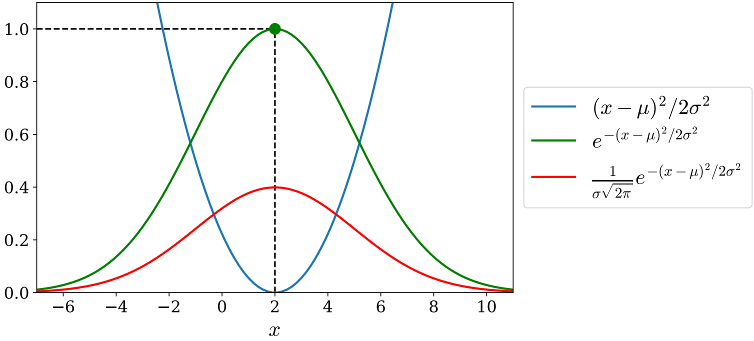
Standard Normal (Gaussian) Random Variable $N(0,1)$



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

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

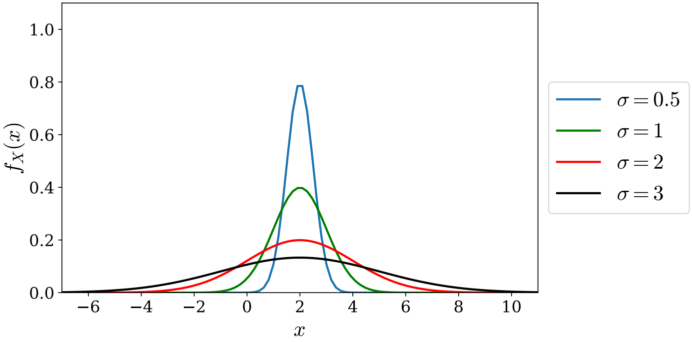
General Normal (Gaussian) Random Variable $N(\mu, \sigma^2)$



$$
\begin{aligned}
f_X(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{1}{2\sigma^2}(y-\mu)^2 \right)
\end{aligned}
$$

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

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



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

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 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

$$\begin{aligned} \texttt{NLL}(\mu, \sigma^2) &= - \sum_{n=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(- \frac{1}{2\sigma^2} (y_n - \mu)^2 \right) \right] \\ &= \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mu)^2 + \frac{N}{2} \log(2\pi\sigma^2) \end{aligned}$$

- The minimum of this function must satisfy the following conditions

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

MLE for the univariate Gaussian

- The solution is given by

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

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

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

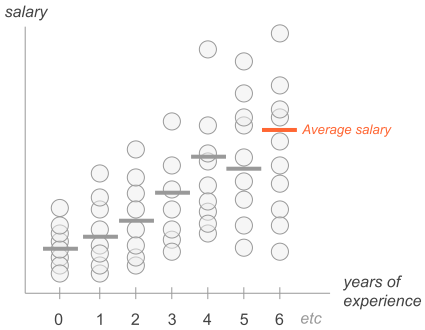
- The quantities $\bar{y}$ and s^2 are called the **sufficient statistics** of the data because they are sufficient to compute the MLE.
- Sometimes, we might se the estimate for the variance as

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

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

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:

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

Linear Regression Example

- The prediction is a conditional mean:

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

- 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 $\texttt{avg}(y_i | \mathbf{x}_i = \mathbf{x}_0)$, which can be formalized as:

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

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.

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 for linear regression

- The joint distribution of $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ 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 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 $\mathbf{b}$.