FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत

NPTEL

SWAYAM & NPTEL COURSE ON

Mathematics for Machine Learning

by

Prof. Debjani Chakraborty

DEPARTMENT OF MATHEMATICS
IIT Kharagpur

Module 8.1

Lecture 36: MLE for Discrete Probability Distribution

Concepts Covered

Maximum Likelihood Estimation for some discrete probability distributions

MLE for Binomial distribution parameter

$$Bin(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Suppose that the random variables $X = \{X_1, X_2, \ldots, X_n\}$ form a *i.i.d.* random sample from a population of Binomial distribution $Bin(k|n,\theta)$. We would like to find $\boldsymbol{\theta_{MLE}}$ for θ.

if X is discrete random variable, $f(x|\theta)$ is point mass function, then for every observed random sample $x_1, x_2, \cdots, x_n$, we define joint probability or the **Likelihood Function**

$$f(x_1, x_2, \cdots, x_n|\theta) = f(x_1|\theta)\cdots f(x_n|\theta)$$

The likelihood function is

$$\boldsymbol{L(\theta) = f(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n}\binom{n}{x_i}\theta^{x_i}(1-\theta)^{(n-x_i)}}$$

MLE for Binomial distribution parameter

Suppose that the random variables $X = \{X_1, X_2, \ldots, X_n\}$ form a *i.i.d.* random sample from a population of Binomial distribution $Bin(k|n, \theta)$. We would like to find $\boldsymbol{\theta_{MLE}}$ for θ.

The likelihood function is

$$\boldsymbol{L(\theta) = f(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} \binom{n}{x_i} \theta^{x_i}(1-\theta)^{(n-x_i)}}$$

The log-likelihood function is

$$\boldsymbol{log\ L(\theta) = log \prod_{i=1}^{n} \binom{n}{x_i} \theta^{x_i}(1-\theta)^{(n-x_i)}}$$

$$\boldsymbol{= log \sum_{i=1}^{n} \left(log \binom{n}{x_i} + x_i\ log\ \theta + (1-x_i) log\ (1-\theta) \right)}$$

MLE for Binomial distribution parameter

Suppose that the random variables $X = \{X_1, X_2, \ldots, X_n\}$ form a *i.i.d.* random sample from a population of Binomial distribution $Bin(k|n, \theta)$. We would like to find $\boldsymbol{\theta_{MLE}}$ for θ.

The log-likelihood function is:

$$\boldsymbol{log \sum_{i=1}^{n} \left(log \binom{n}{x_i} + x_i \, log \, \theta + (1 - x_i) log \, (1 - \theta) \right)}$$

Taking derivative of log-likelihood function and equating to zero –

$$\boldsymbol{\frac{d \, L(\theta)}{d\theta} = \sum_{i=1}^{n} \left(\frac{x_i}{\theta} - \frac{1 - x_i}{1 - \theta} \right) = 0}$$

$$\boldsymbol{\theta_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

MLE for Poisson distribution parameter

$$\text{Poisson}(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda > 0$$

Suppose that the random variables $X = \{X_1, X_2, \ldots, X_n\}$ form a *i.i.d.* random sample from a population of Poisson distribution $\text{Poisson}(k|\lambda)$. We would like to find $\boldsymbol{\lambda_{MLE}}$ for λ.

if X is discrete random variable, $f(x|\lambda)$ is point mass function, then for every observed random sample $x_1, x_2, \cdots, x_n$, we define joint probability or the **Likelihood Function**

$$f(x_1, x_2, \cdots, x_n|\lambda) = f(x_1|\lambda) \cdots f(x_n|\lambda)$$

The likelihood function is

$$\boldsymbol{L(\lambda) = f(x_1, x_2, \cdots, x_n|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}}$$

MLE for Poisson distribution parameter

Suppose that the random variables $X = \{X_1, X_2, \ldots, X_n\}$ form a *i.i.d.* random sample from a population of Binomial distribution $Bin(k|n, \theta)$. We would like to find $\boldsymbol{\theta_{MLE}}$ for θ.

The likelihood function is

$$\boldsymbol{L(\lambda) = f(x_1, x_2, \cdots, x_n|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}}$$

The log-likelihood function is

$$\boldsymbol{log\ L(\lambda) = log \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}}$$

$$\boldsymbol{= -n\lambda + log\ \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} log\ x_i!}$$

MLE for Binomial distribution parameter

Suppose that the random variables $X = \{X_1, X_2, \ldots, X_n\}$ form a *i.i.d.* random sample from a population of Binomial distribution $Bin(k|n, \theta)$. We would like to find $\boldsymbol{\theta_{MLE}}$ for θ.

The log-likelihood function is:

$$-n\lambda + log\ \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} log\ x_i!$$

Taking derivative of log-likelihood function and equating to zero –

$$\frac{d\ L(\lambda)}{d\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

$$\lambda_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example: Let a population has been considered with the following probability distribution

$$f(x|\theta) = \begin{cases} \frac{1}{4} & if\ x = 0 \\ \frac{1+2\theta}{4} & if\ x = 1 \\ \frac{1}{4} & if\ x = 2 \\ \frac{1-2\theta}{4} & if\ x = 3 \end{cases}$$

A sample with 10 observations has been considered, which is as follows:

$$(3, 0, 2, 1, 3, 2, 1, 0, 2, 1\ )$$

Find $\boldsymbol{\theta_{MLE}}$ for $\boldsymbol{\theta}$.

Example: Let a population has been considered with the following probability distribution

$$f(x|\theta) = \begin{cases} \frac{1}{4} & if\ x = 0 \\ \frac{1+2\theta}{4} & if\ x = 1 \\ \frac{1}{4} & if\ x = 2 \\ \frac{1-2\theta}{4} & if\ x = 3 \end{cases}$$

Sample: (3, 0, 2, 1, 3, 2, 1, 0, 2, 1)

The likelihood function is

$$L(\theta) = P(x=3)\ P(x=0)P(x=2)P(x=1)P(x=3)P(x=2)P(x=1)P(x=0)$$
$$P(x=2)P(x=1)$$

$$= \left(\frac{1}{4}\right)^2 \left(\frac{1+2\theta}{4}\right)^3 \left(\frac{1}{4}\right)^3 \left(\frac{1-2\theta}{4}\right)^2$$

The log-likelihood function is:

$$Log\ L(\theta) = \{2\ log\ 1 - 2log\ 4\} + \{3\ log\ (1+2\theta) - 3log\ 4\} + \{3\ log\ 1 - 3log\ 4\} +$$
$$\{2\ log\ (1-2\theta) - 2log\ 4\}$$

Example: Let a population has been considered with the following probability distribution

$$f(x|\theta) = \begin{cases} \frac{1}{4} & if\ x = 0 \\ \frac{1+2\theta}{4} & if\ x = 1 \\ \frac{1}{4} & if\ x = 2 \\ \frac{1-2\theta}{4} & if\ x = 3 \end{cases}$$

Sample: (3, 0, 2, 1, 3, 2, 1, 0, 2, 1)

The log-likelihood function is:

$$\boldsymbol{Log\ L(\theta)}$$
$$= \{2\ log\ 1 - 2log\ 4\} + \{3\ log\ (1+2\theta) - 3log\ 4\} + \{3\ log\ 1 - 3log\ 4\} + \{2\ log\ (1-2\theta) - 2log\ 4\}$$

Taking derivative of log-likelihood function and equating to zero –

$$\frac{\boldsymbol{d\ L(\theta)}}{\boldsymbol{d\theta}} = \frac{\boldsymbol{6}}{\boldsymbol{1+2\theta}} - \frac{\boldsymbol{4}}{\boldsymbol{1-2\theta}} = \boldsymbol{0}$$

$$\boldsymbol{\theta_{MLE} = .1}$$

Example: Let a population has been considered with the following probability distribution

$$f(x|\theta) = \begin{cases} \frac{1+2\theta}{4} & if\ x = 0 \\ \frac{1-\theta}{4} & if\ x = 1 \\ \frac{1+\theta}{4} & if\ x = 2 \\ \frac{1-2\theta}{4} & if\ x = 3 \end{cases}$$

A sample with 10 observations has been considered, which is as follows:

$$(3, 0, 2, 1, 3, 2, 1, 0, 2, 1\ )$$

Find $\boldsymbol{\theta_{MLE}}$ for $\boldsymbol{\theta}$.

Example: Let's assume our observed data points are: $x = [1,3,2,4,5,2,1,3,2,4]$ from a Poisson distribution.

The likelihood function for the Poisson distribution is:

$$L(\lambda) = \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

Taking the logarithm of the likelihood function, we get the log-likelihood function:

$$\log L(\lambda) = -N\lambda + \sum_{i=1}^{N} x_i \log \lambda - \sum_{i=1}^{N} \log(x_i!)$$

To find the maximum likelihood estimate for λ, we differentiate the log-likelihood function with respect to λ, set it equal to zero, and solve for λ:

$$\frac{d}{d\lambda} \log L(\lambda) = -N + \frac{1}{\lambda} \sum_{i=1}^{N} x_i = 0$$

$$\Rightarrow \quad \lambda = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Now, let's calculate the MLE of using our sample data:

$$\boldsymbol{\lambda_{MLE}} = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{1}{10}(1+3+2+4+5+2+1+3+2+4) = \frac{27}{10} = 2.7$$

- ✓ **Many practical problems we examine appropriate conditions and select the probability distribution**

- ✓ **A range of discrete probability distributions are considered here**

Different continuous probability distributions are in the next…..

References

- ✓ Linear Algebra and Learning from Data (2019), Gilbert Strang, Wellesley Cambridge Press
- ✓ Machine Learning: A Probabilistic Perspective, Kevin P. Murphy (MIT Press), 2021 edition
- ✓ Deisenroth MP, Faisal AA, Ong CS. Mathematics for machine learning. Cambridge University Press; 2020 Apr 23.


Thank You...
INDIAN INSTITUTE OF TECHNOLOGY
NPTEL
FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
1951
योगः कर्मसु कौशलम्
NPTEL ONLINE CERTIFICATION COURSES
IIT KHARGPUR

SWAYAM & NPTEL COURSE ON

Mathematics for Machine Learning

by

Prof. Debjani Chakraborty

DEPARTMENT OF MATHEMATICS
IIT Kharagpur

Module 8.2

Lecture 37: MLE for Continuous Probability Distribution

Concepts Covered

Maximum Likelihood Estimation for some continuous probability distributions

MLE for Exponential distribution parameter

Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n exponential random variables, we would like to find $\boldsymbol{\lambda_{MLE}}$ for λ.

$$\mathbf{f}(\boldsymbol{x};\boldsymbol{\lambda}) = \begin{cases} \boldsymbol{\lambda e^{-\lambda x}} & \boldsymbol{x \geq 0} \\ \mathbf{0} & \boldsymbol{otherwise} \end{cases}$$

if X is continuous random variable, $f(x|\lambda)$ is probability density function, then for every observed random sample $x_1, x_2, \cdots, x_n$, we define joint probability or the **Likelihood Function**

$$f(x_1, x_2, \cdots, x_n|\lambda) = f(x_1|\lambda) \cdots f(x_n|\lambda)$$

$\lambda_{MLE} = \underset{\lambda}{\text{argmax}}\ L(\lambda)$

$= \underset{\lambda}{\text{argmax}}\ \log L(\lambda)$

The likelihood function is

$$\boldsymbol{L(\lambda) = f(x_1, x_2, \cdots, x_n|\hat{\theta}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}}$$

MLE for Exponential distribution parameter

Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n exponential random variables, we would like to find $\boldsymbol{\lambda_{MLE}}$ for λ.

$$\mathbf{f}(x;\boldsymbol{\lambda}) = \begin{cases} \boldsymbol{\lambda e^{-\lambda x}} & \boldsymbol{x \geq 0} \\ \mathbf{0} & \boldsymbol{otherwise} \end{cases}$$

The likelihood function is

$$\boldsymbol{L(\lambda) = f(x_1, x_2, \cdots, x_n|\lambda)} = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

The log-likelihood function is

$$\boldsymbol{\log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i}$$

MLE for Exponential distribution parameter

Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n exponential random variables, we would like to find λ_{MLE} for λ.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & otherwise \end{cases}$$

The log-likelihood function is

$$log\ L(\lambda) = n\ log\ \lambda - \lambda \sum_{i=1}^{n} x_i$$

Taking derivative of log-likelihood function w.r.t. μ and equating to zero –

$$\frac{dlog\ L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \qquad \frac{n}{\lambda} = \sum_{i=1}^{n} x_i$$

$$\lambda = \frac{1}{\bar{x}} \quad \bar{x} = \text{sample}$$

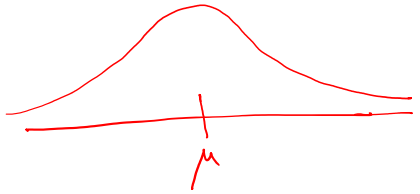
$$\lambda_{MLE} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

MLE for Normal distribution parameters

Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables We would like to find $\boldsymbol{\mu_{MLE}}$ for μ and $\boldsymbol{\sigma_{MLE}}$ for σ.

$$\boldsymbol{f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}}$$

where $-\infty < x, \mu < \infty, \sigma > 0$



if X is continuous random variable, $f(x|\mu,\sigma)$ is probability density function, then for every observed random sample $x_1, x_2, \cdots, x_n$, we define joint probability or the **Likelihood Function**

$$f(x_1, x_2, \cdots, x_n|\mu,\sigma) = f(x_1|\mu,\sigma) \cdots f(x_n|\mu,\sigma)$$

$\mu_{MLE} = \underset{\mu}{\operatorname{argmax}}\, L(\mu,\sigma)$

σ^2_{MLE}

The likelihood function is

$$\boldsymbol{L(\theta) = f(x_1, x_2, \cdots, x_n|\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}}$$

MLE for Normal distribution parameters

Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables We would like to find $\boldsymbol{\mu_{MLE}}$ for μ and $\boldsymbol{\sigma_{MLE}}$ for σ.

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

where $-\infty < x, \mu < \infty, \sigma > 0$

The likelihood function is

$$L(\mu,\sigma) = f(x_1, x_2, \cdots, x_n|\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

The log-likelihood function is

$$log\ L(\mu,\sigma) = -\frac{n}{2} log\ 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

MLE for Normal distribution parameters

Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables We would like to find $\boldsymbol{\mu_{MLE}}$ for μ and $\boldsymbol{\sigma_{MLE}}$ for σ.

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

where $-\infty < x, \mu < \infty, \sigma > 0$

The log-likelihood function is

$$log\ L(\mu,\sigma) = -\frac{n}{2} log\ 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Taking derivative of log-likelihood function w.r.t. μ and equating to zero –

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

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\sum x_i - n\mu = 0$$

$$\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

MLE for Normal distribution parameters

Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables We would like to find $\boldsymbol{\mu_{MLE}}$ for μ and $\boldsymbol{\sigma_{MLE}}$ for σ.

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

where $-\infty < x, \mu < \infty, \sigma > 0$

The log-likelihood function is

$$log\ L(\mu,\sigma) = -\frac{n}{2} log\ 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

partial

Taking derivative of log-likelihood function w.r.t. σ^2 and equating to zero –

$$\frac{\delta log\ L(\mu,\sigma)}{\delta\sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{n}{(\sigma^2)^2}\left(-\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right) = 0$$

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

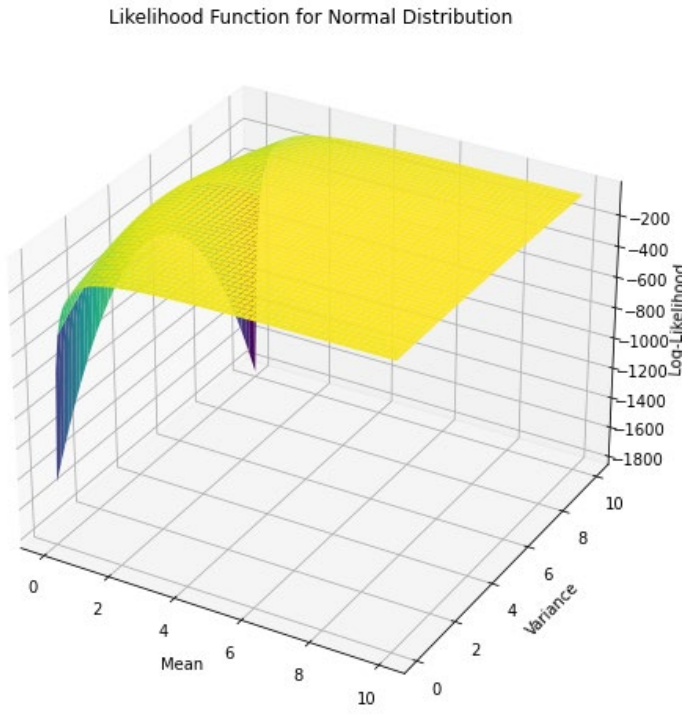
Example: Now, let's take a numerical example with some sample data. Suppose we have the following 10 observations drawn from a Normal distribution: X={2.5, 3.1, 4.0, 2.8, 3.7,8.9, 4.3,10.3,1.1,5.6}. Find MLE of μ and σ^2.

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The log-likelihood function is

$$log\ L(\mu, \sigma) = -\frac{10}{2} log\ 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$



μ_{MLE} = 4.63.

σ^2_{MLE} = 7.54.

Example: Now, let's take a numerical example with some sample data. Suppose we have the following 10 observations drawn from the following distribution: X={2.5, 3.1, 4.0, 2.8, 3.7,8.9, 4.3,10.3,1.1,5.6}. Find MLE of θ. Find Mean and standard deviation also.

$$f(x|\theta) = (1-2\theta) + 4\theta x, \qquad 0 < x < 1, \qquad -1 < \theta < 1$$

$$E(x) = \int_0^1 x f(x)\, dx.$$

$$= \int_0^1 \{(1-2\theta)x + 4\theta x^2\}\, dx.$$

$$= \frac{(1-2\theta)x^2}{2}\Big|_0^1 + \frac{4\theta x^3}{3}\Big|_0^1$$

$$= \frac{1}{2} - \theta + \frac{4}{3}\theta = \left(\frac{1}{2} + \frac{\theta}{3}\right)$$

$$\int_0^1 f(x)\, dx = \int_0^1 \{(1-2\theta)x + 4\theta x\}\, dx$$

$$= (1-2\theta)x\Big|_0^1 + 2\theta x^2\Big|_0^1$$

$$= 1 - 2\theta + 2\theta = 1$$

$$Var(x) = \int_0^1 \{(x - E(x)\}^2 f(x)\, dx.$$

Example: Now, let's take a numerical example with some sample data. Suppose we have the following 10 observations drawn from the following distribution: X={2.5, 3.1, 4.0, 2.8, 3.7,8.9, 4.3,10.3,1.1,5.6}. Find MLE of θ. Find Mean and standard deviation also.

$$f(x|\theta) = (1-2\theta) + 4\theta x, \qquad 0 < x < 1, \qquad -1 < \theta < 1$$

$$Var(X) = \int_0^1 \left\{x - \left(\frac{1}{2}+\frac{\theta}{3}\right)\right\}^2 \left\{(1-2\theta) + 4\theta x\right\} dx.$$

$$= \int_0^1 \left[x^2 + \left(\frac{1}{2}+\frac{\theta}{3}\right)^2 - 2x\left(\frac{1}{2}+\frac{\theta}{3}\right)\right]\left[(1-2\theta) + 4\theta x\right] dx$$

$$\log L(\theta) = \log \prod_{i=1}^{10} f(x_i|\theta) = \log \prod_{i=1}^{10} \left\{(1-2\theta) + 4\theta x_i\right\}$$

$$= \log g(\theta)$$

$$\frac{d \log L(\theta)}{d\theta} = 0, \qquad \theta_{MLE} = ?$$

Example: Now, let's take a numerical example with some sample data. Suppose we have the following 10 observations drawn from the following distribution: X={2.5, 3.1, 4.0, 2.8, 3.7,8.9, 4.3,10.3,1.1,5.6}. Find MLE of θ.

$$f(x|\theta) = \frac{\theta}{x^{\theta+1}},$$

$$L(\theta) = \prod_{i=1}^{10} \frac{\theta}{x_i^{\theta+1}}, \quad \log L(\theta) = 10 \log \theta - (\theta+1) \sum_{i=1}^{10} \log x_i$$

$x_i \in X$

$$\frac{d \log L(\theta)}{d\theta} = \frac{10}{\theta} - \sum \log x_i = 0$$

$$\theta_{MLE} = \theta_{MLE} = \frac{10}{\sum_{i=1}^{10} \log x_i}$$

Example: Now, let's take a numerical example with some sample data. Suppose we have the following 10 observations drawn from the following distribution: X={2.5, 3.1, 4.0, 2.8, 3.7,8.9, 4.3,10.3,1.1,5.6}. Find MLE of θ.

$$f(x|\theta) = \frac{\theta}{x^{\theta+1}},$$

```python
import math

def MLE_theta(sample_points):
    # This code computes the MLE estimate of alpha
    log_sum = 0
    for x in sample_points:
        log_sum += math.log(x)
    n = len(sample_points)
    return n / log_sum

def main():
    sample_points = [2.5, 3.1, 4.0, 2.8, 3.7,8.9, 4.3,10.3,1.1,5.6]
    theta = MLE_theta(sample_points)
    print(theta)

if __name__ == '__main__':
    main()
```

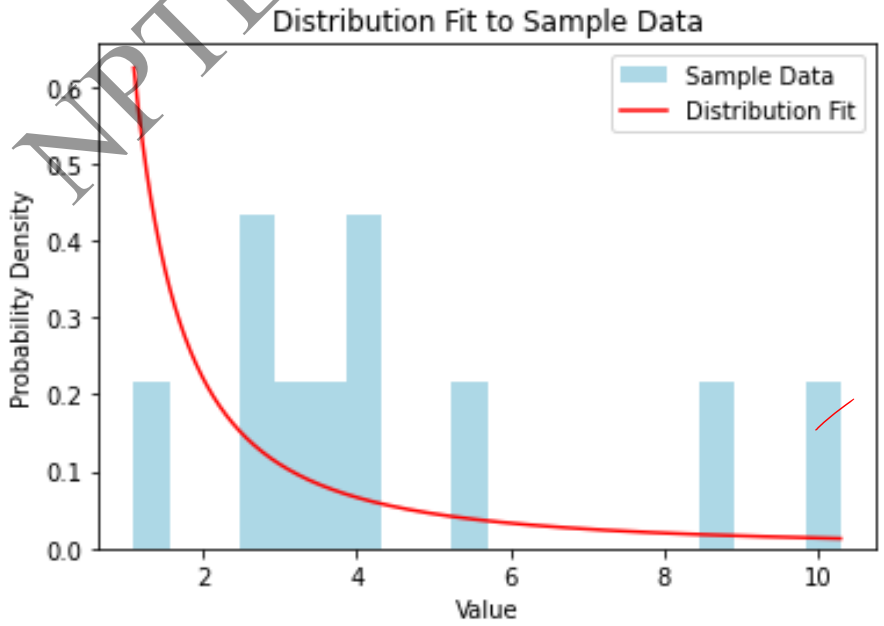
$\sum_{i=1}^{n} \log x_i$

$$\theta_{MLE} = 0.7370920578760686$$

$$\theta_{MLE} = \frac{10}{\sum_{i=1}^{n} \log x_i}$$

```python
import math
import numpy as np
import matplotlib.pyplot as plt

def MLE_theta(sample_points):
    # This code computes the MLE estimate of theta
    log_sum = 0
    for x in sample_points:
        log_sum += math.log(x)
    n = len(sample_points)
    return n / log_sum

def show_pdf(x, theta):
    # distribution PDF
    return theta  / x**(theta + 1)

def main():
    sample_points = [2.5, 3.1, 4.0, 2.8, 3.7, 8.9, 4.3, 10.3, 1.1, 5.6]
    theta = MLE_theta(sample_points)
    print(theta)


    # Plot the histogram of the sample data
    plt.hist(sample_points, bins=20, density=True,
             color='lightblue', label='Sample Data')

    # Generate points for the fitted distribution
    x_values = np.linspace(min(sample_points), max(sample_points), 1000)
    y_values = [show_pdf(x, theta) for x in x_values]

    # Plot the fitted distribution
    plt.plot(x_values, y_values, color='red', label='Distribution Fit')

    plt.xlabel('Value')
    plt.ylabel('Probability Density')
    plt.title('Distribution Fit to Sample Data')
    plt.legend()
    plt.show()

if __name__ == '__main__':
    main()
```

```python
import math
import numpy as np
import matplotlib.pyplot as plt

def MLE_theta(sample_points):
    # This code computes the MLE estimate of theta
    log_sum = 0
    for x in sample_points:
        log_sum += math.log(x)
    n = len(sample_points)
    return n / log_sum

def show_pdf(x, theta):
    # distribution PDF
    return theta  / x**(theta + 1)

def main():
    sample_points = [2.5, 3.1, 4.0, 2.8, 3.7, 8.9, 4.3, 10.3, 1.1, 5.6]
    theta = MLE_theta(sample_points)
    print(theta)


    # Plot the histogram of the sample data
    plt.hist(sample_points, bins=20, density=True,
             color='lightblue', label='Sample Data')

    # Generate points for the fitted distribution
    x_values = np.linspace(min(sample_points), max(sample_points), 1000)
    y_values = [show_pdf(x, theta) for x in x_values]

    # Plot the fitted distribution
    plt.plot(x_values, y_values, color='red', label='Distribution Fit')

    plt.xlabel('Value')
    plt.ylabel('Probability Density')
    plt.title('Distribution Fit to Sample Data')
    plt.legend()
    plt.show()

if __name__ == '__main__':
    main()
```



EXAMPLE:

Let us consider a random variable X$\sim N(\mu, \sigma^2)$. If we consider a i.i.d. sample of size 3 and consider these two estimators for μ

Estimator 1: $\hat{\theta}_1 = x_1 + x_2 - x_3$

Estimator 2: $\hat{\theta}_2 = x_1 + x_2 - 2x_3$

$\mu_{MLE} = \bar{x} = \frac{x_1 + x_2 + x_3}{3}$

Estimator 1: $E(\hat{\theta}_1) = E[x_1 + x_2 - x_3] = \mu$

Estimator 2: $E(\hat{\theta}_2) = E(x_1 + x_2 - 2x_3) = \mu$

But,

Estimator 1: $var(\hat{\theta}_1) = var[x_1 + x_2 - x_3] = 3\sigma^2$,

Estimator 2: $var(\hat{\theta}_2) = var(x_1 + x_2 - 2x_3) = 6\sigma^2$

What about MSE?

Conclusion

- ✓ **MLE for some continuous probability distributions**

To do: Properties of MLE

References

- ✓ Linear Algebra and Learning from Data (2019), Gilbert Strang, Wellesley Cambridge Press
- ✓ Machine Learning: A Probabilistic Perspective, Kevin P. Murphy (MIT Press), 2021 edition
- ✓ Deisenroth MP, Faisal AA, Ong CS. Mathematics for machine learning. Cambridge University Press; 2020 Apr 23.


Thank You...
INDIAN INSTITUTE OF TECHNOLOGY
NPTEL
FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
1951
योगः कर्मसु कौशलम्
NPTEL ONLINE CERTIFICATION COURSES
IIT KHARGPUR

FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत

NPTEL

SWAYAM & NPTEL COURSE ON

Mathematics for Machine Learning

by

Prof. Debjani Chakraborty

DEPARTMENT OF MATHEMATICS
IIT Kharagpur

Module 8.3

Lecture 38: Properties of Estimators

Concepts Covered

1. Unbiased Estimator
2. Consistent Estimators
3. Minimum Variance Estimators
4. Bias vs Variance

Estimators: $\widehat{\theta} = f(X_1, X_2 \cdots X_n)$

Estimate: Value of the Estimator

Bias: Expected difference between the value of the Estimator and True value of Estimator

$$Bias = E\left(\widehat{\theta} - \theta\right) = E_{X_1, X_2 \cdots X_n}[argmax_\theta\, f(X_1, X_2 \cdots X_n) - \theta]$$

- Even the sample is large, due to bias estimator does not converge to true estimator
- That is, it is not expressive enough to approximate the true value arbitrarily well
- If Bias is zero it is called unbiased Estimator
- To converge, how many sample points we need!!!!!

Estimators: $\widehat{\theta} = f(X_1, X_2 \cdots X_n)$

Estimate: Value of the Estimator

Variance: Expected difference between the value of the Estimator and True value of Estimator

$$\boldsymbol{Variance} = E\left(\left(\widehat{\theta} - E(\widehat{\theta})\right)^2\right)$$

$$= E_{X_1, X_2 \cdots X_n}\left((argmax_\theta f(X_1, X_2 \cdots X_n) - E(argmax_\theta f(X_1, X_2 \cdots X_n)))^2\right)$$

- If more training examples are taken then variance decreases
- If variance of an estimator is minimum compared to any other estimator of the same parameter and it is unbiased then it is called **Minimum Variance Unbiased Estimator (MVUE)**

Sample Mean is unbiased estimator of population Mean for $N(\mu, \sigma^2)$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} \cdot n\mu = \mu$$

$\bar{X}$

$$\text{Bias}(\bar{X}) = E(\bar{X} - \mu) = 0$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n^2} \cdot \sum_{i=1}^{n} \text{Var}(x_i)$$

$$= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

$$variance(\overline{X}) = \frac{\sigma^2}{n}$$

Sample Variance is not unbiased estimator of population Variance for $N(\mu, \sigma^2)$

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2$$

$$= \sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(\bar{x}-\mu)^2 + 2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu)$$

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

$$E\left(S^1_{Var}\right) = \frac{1}{n}\sum_{i=1}^{n}E(x_i-\mu)^2 - \frac{1}{n}\sum_{i=1}^{n}E(\bar{x}-\mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}Var(x_i) - Var(\bar{x})$$

$$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

$$\times S^1_{Var} = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

$$\checkmark S^2_{Var} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Sample Variance is not unbiased estimator of population Variance for $N(\mu, \sigma^2)$

$$E\left(S'_{var}\right) = \frac{n-1}{n}\sigma^2$$

$$S^2_{var} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$$

$$Bias\left(S'_{var}\right) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

$$= -var(\bar{x})$$

$$E\left(\frac{n}{n-1}S'_{var}\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2$$

S^2_{var}

$$E\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \sigma^2$$

Consistent Estimators

If $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ is estimator of θ for $f(\boldsymbol{X_1}, \boldsymbol{X_2} \cdots \boldsymbol{X_n}|\boldsymbol{\theta})$ based on a sample of size n, if $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} \rightarrow \boldsymbol{\theta}$ in probability when $\boldsymbol{n} \rightarrow \infty$ then $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ is **consistent** estimator of $\boldsymbol{\theta}$.

$$P\left\{|\hat{\theta}_n - \theta| \rightarrow 0\right\} \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

Efficient Estimators

If $\hat{\boldsymbol{\theta}}_{\boldsymbol{1}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{2}}$ are two unbiased estimators of $\boldsymbol{\theta}$ for $f(\boldsymbol{X_1}, \boldsymbol{X_2} \cdots \boldsymbol{X_n}|\boldsymbol{\theta})$ based on same sample of size n, then $\hat{\boldsymbol{\theta}}_{\boldsymbol{1}}$ is more efficient than $\hat{\boldsymbol{\theta}}_{\boldsymbol{2}}$ when $\boldsymbol{var}(\hat{\boldsymbol{\theta}}_{\boldsymbol{1}}) < \boldsymbol{var}(\hat{\boldsymbol{\theta}}_{\boldsymbol{2}})$.

Let $X_1, X_2 \cdots X_n$ be a random sample from $X \sim Exponential(\lambda)$. Consider the following estimators $\widehat{\theta_1} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\widehat{\theta_2} = \frac{1}{n+1}\sum_{i=1}^{n} X_i$ of the population mean.

1. Are these unbiased estimators?
2. Find the Mean Square Errors.
3. Which estimator is better and why?

$E(\hat{\theta}_1) = \frac{1}{n}\sum E(X_i)$ $\quad \theta = \frac{1}{\lambda}$

$= \frac{\theta}{n} \cdot n = \theta$

$E(\hat{\theta}_1) - \theta = 0$ ✓

$E(\hat{\theta}_2) - \theta = \frac{1}{n+1}\sum_{i=1}^{n} E(X_i) - \theta = \frac{\theta \cdot n}{n+1} - \theta = -\frac{\theta}{n+1}$ ✓

$Var(\hat{\theta}_1) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot \frac{n}{\lambda^2} = \frac{\theta^2}{n}$

$\widehat{\theta_1} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\widehat{\theta_2} = \frac{1}{n+1}\sum_{i=1}^{n} X_i$

1. Are these unbiased estimators?
2. Find the Mean Square Errors.
3. Which estimator is better and why?

$$Var(\hat{\theta}_2) = \frac{1}{(n+1)^2}\sum_{i=1}^{n} Var(X_i) = \frac{n}{(n+1)^2}\cdot\frac{1}{\lambda^2} = \frac{n\theta}{(n+1)^2}$$

$$Var(\hat{\theta}_1) = \frac{\theta^2}{n}$$

$$MSE(\hat{\theta}_1) = Bias^2 + Variance = 0 + \frac{\theta^2}{n} = \frac{\theta^2}{n}$$

$$MSE(\hat{\theta}_2) = \left(\frac{\theta}{n+1}\right)^2 + \frac{n\theta}{(n+1)^2} = \frac{\theta^2 + n\theta}{(n+1)^2} = \frac{\theta}{n+1}$$

$$Bias(\hat{\theta}_1) < Bias(\hat{\theta}_2)$$

$$MSE(\hat{\theta}_1) > MSE(\hat{\theta}_2)$$

$$\widehat{\theta_1} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \widehat{\theta_2} = \frac{1}{n+1}\sum_{i=1}^{n} X_i$$

1. Are these unbiased estimators?
2. Find the Mean Square Errors.
3. Which estimator is better and why?

Conclusion

- ✓ **Criteria to check efficiency of different estimators**

References

- ✓ Linear Algebra and Learning from Data (2019), Gilbert Strang, Wellesley Cambridge Press
- ✓ Machine Learning: A Probabilistic Perspective, Kevin P. Murphy (MIT Press), 2021 edition
- ✓ Deisenroth MP, Faisal AA, Ong CS. Mathematics for machine learning. Cambridge University Press; 2020 Apr 23.


Thank You...
INDIAN INSTITUTE OF TECHNOLOGY
NPTEL
FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
1951
योगः कर्मसु कौशलम्
NPTEL ONLINE CERTIFICATION COURSES
IIT KHARGPUR

FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत

NPTEL

SWAYAM & NPTEL COURSE ON

Mathematics for Machine Learning

by

Prof. Debjani Chakraborty

DEPARTMENT OF MATHEMATICS
IIT Kharagpur

Module 8.4

Lecture 39: MLE for Linear and Logistic Regression

Concepts covered

- ✓ **MLE for Linear Regression**
- ✓ **MLE for logistic Regression**

The Method of Maximum Likelihood Estimation

If n random samples $X_1, X_2, .., X_n$ are drawn from a population with probability function $f(X|\theta)$, then joint probability function L of $X_1, X_2, .., X_n$ is,

$$\boldsymbol{L(\theta) = f(x_1|\theta) f(x_2|\theta) \cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)}$$

***L*: Likelihood function.**

Since, $X_1, X_2, .., X_n$ are independently and identically distributed.

Objective: To find θ, such that Likelihood function is maximum

i.e. $\hat{\theta} = \underset{\theta}{argmax}\ L(\theta)$

Linear Regression (Parameter Fitting with Maximum Likelihood)

- $X =$ arbitrary distribution
- If $X =$ x then $Y = w_0 + w_1 x + \beta$, for some constants w_0, w_1, and some random noise variable β.
- $\beta \sim N(0, \sigma^2)$ which is independent of X.
- Here, we are assuming noise variable β has particular Gaussian distribution.

Conditional probability density function of Y, $p(y|X = x; w_0, w_1, \sigma^2)$ for each x.

$\beta = Y - w_0 - w_1 x \sim N(0, \sigma^2)$

Training Data

x_i	y_i	Prob.
x_1	y_1	p_1
x_2	y_2	p_2
...	...	...
x_n	y_n	p_n

The probability density under the model for the given data:

$$\prod_{i=1}^{n} p(y_i|x_i; w_0, w_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_1-(w_0+w_1x_i))^2/2\sigma^2}$$

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

Linear Regression (Parameter Fitting with Maximum Likelihood)

When we visualize data, we don't known the true parameters, consider $(\theta_0, \theta_1, s^2)$ gives the probability density:

$$\prod_{i=1}^{n} p(y_i|x_i;\theta_0,\theta_1,s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} e^{-\left(y_1-(\theta_0+\theta_1 x_i)\right)^2 / 2s^2}$$

The above equation is the likelihood, a function of the parameter values.

It is easier to work with **log-likelihood**,

$$L(\theta_0,\theta_1,s^2) = \log \prod_{i=1}^{n} p(y_i|x_i;\theta_0,\theta_1,s^2)$$

$$= \sum_{i=1}^{n} \log p(y_i|x_i;\theta_0,\theta_1,s^2) = -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}^{n}\left(y_i-(\theta_0+\theta_1 x_i)\right)^2$$

Linear Regression (Parameter Fitting with Maximum Likelihood)

$$L(\theta_0, \theta_1, s^2) = -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}^{n}\left(y_i - (\theta_0 + \theta_1 x_i)\right)^2 \quad \text{---------(1)}$$

To finding the parameters θ_0 and θ_1, first we will partial differentiate equation (1) with respect to θ_0,

$$\frac{\partial L}{\partial \theta_0} = 0$$

$$\frac{1}{2s^2} \times 2\sum_{i=1}^{n}\{y_i - (\theta_0 + \theta_1 x_i)\} = 0 \Rightarrow \sum_{i=1}^{n}\{y_i - (\theta_0 + \theta_1 x_i)\} = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \theta_0 - \sum_{i=1}^{n} \theta_1 x_i = 0$$

Linear Regression (Parameter Fitting with Maximum Likelihood)

$$L(\theta_0, \theta_1, s^2) = -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}^{n}\left(y_i - (\theta_0 + \theta_1 x_i)\right)^2 \quad \text{---------(1)}$$

To finding the parameters θ_0 and θ_1, first we will partial differentiate equation (1) with respect to θ_0,

$$\frac{\partial L}{\partial \theta_0} = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \theta_0 - \sum_{i=1}^{n} \theta_1 x_i = 0$$

$$n\bar{y} - n\theta_0 - n\theta_1 \times \frac{1}{n}\sum_{i=1}^{n} x_i = 0$$

$$n\bar{y} - n\theta_0 - n\theta_1\bar{x} = 0 \implies \bar{y} - \theta_0 - \theta_1\bar{x} = 0$$

$$\boldsymbol{\theta_0^{MLE} = \bar{y} - \theta_1^{MLE}\bar{x}} \quad \text{----------(2)}$$

Linear Regression (Parameter Fitting with Maximum Likelihood)

$$L(\theta_0, \theta_1, s^2) = -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_i))^2 \quad \text{---------(1)}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x} \quad \text{----------(2)}$$

partial differentiate equation (1) with respect to θ_1, $\frac{\partial L}{\partial \theta_1} = 0$

$$\frac{1}{2s^2} \times 2\sum_{i=1}^{n}\{y_i - (\theta_0 + \theta_1 x_i)\}x_i = 0 \Rightarrow \sum_{i=1}^{n}\{y_i - (\theta_0 + \theta_1 x_i)\}x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\theta_0 x_i - \sum_{i=1}^{n}\theta_1 x_i^2 = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i - n \times \theta_0 \times \frac{1}{n}\sum_{i=1}^{n} x_i - \theta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\theta_0 \bar{x} - \theta_1 \sum_{i=1}^{n} x_i^2 = 0 \quad \text{----------(3)}$$

Linear Regression (Parameter Fitting with Maximum Likelihood)

$$L(\theta_0, \theta_1, s^2) = -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_i))^2 \quad \text{---------(1)}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x} \quad \text{----------(2)}$$

$\frac{\partial L}{\partial \theta_0} = 0$

$$\sum_{i=1}^{n} x_i y_i - n\theta_0 \bar{x} - \theta_1 \sum_{i=1}^{n} x_i^2 = 0 \quad \text{----------(3)}$$

$\frac{\partial L}{\partial \theta_1} = 0$

Using equation (2) in equation (3),

$$\sum_{i=1}^{n} x_i y_i - \theta_1 \sum_{i=1}^{n} x_i^2 - n(\bar{y} - \theta_1 \bar{x})\bar{x} = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i - \theta_1 \sum_{i=1}^{n} x_i^2 - n\bar{x}\bar{y} + n\theta_1 \bar{x}^2 = 0$$

$$\theta_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\Rightarrow \theta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Linear Regression (Parameter Fitting with Maximum Likelihood)

$$\theta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\theta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n\bar{x}^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}}{\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2}$$

$$\theta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2\,\bar{x} x_i + \sum_{i=1}^{n} \bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x} + \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}$$

$$\theta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^{n} \bar{x}(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + n\bar{x}\bar{y} - n\bar{x}\bar{y}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\boldsymbol{\theta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Linear Regression (Parameter Fitting with Maximum Likelihood)

In the maximum likelihood method, the following estimators we will obtain:

$$w_1^{MLE} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$w_0^{MLE} = \bar{y} - w_1^{MLE}\bar{x}$$

$$\sigma_{MLE}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(w_0^{MLE} + w_1^{MLE}x_i\right)\right)^2$$

- The above estimators for the slope and intercept is exactly similar to the least square estimators.
- This is the crucial property of considering independent Gaussian noise.

Linear Regression (Parameter Fitting with Maximum Likelihood)

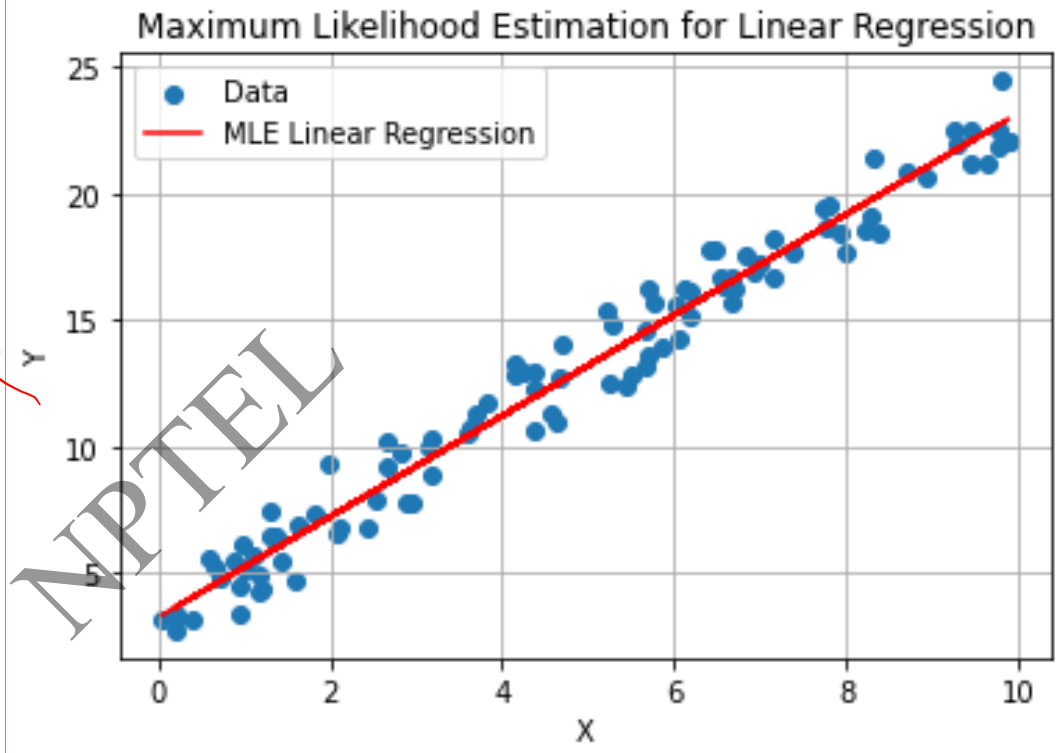
In the maximum likelihood method, the following estimators we will obtain:

$$w_1^{MLE} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$w_0^{MLE} = \bar{y} - w_1^{MLE}\bar{x}$$

$$\sigma_{MLE}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(w_0^{MLE} + w_1^{MLE}x_i\right)\right)^2$$

- The above estimators for the slope and intercept is exactly similar to the least square estimators.
- This is the crucial property of considering independent Gaussian noise.

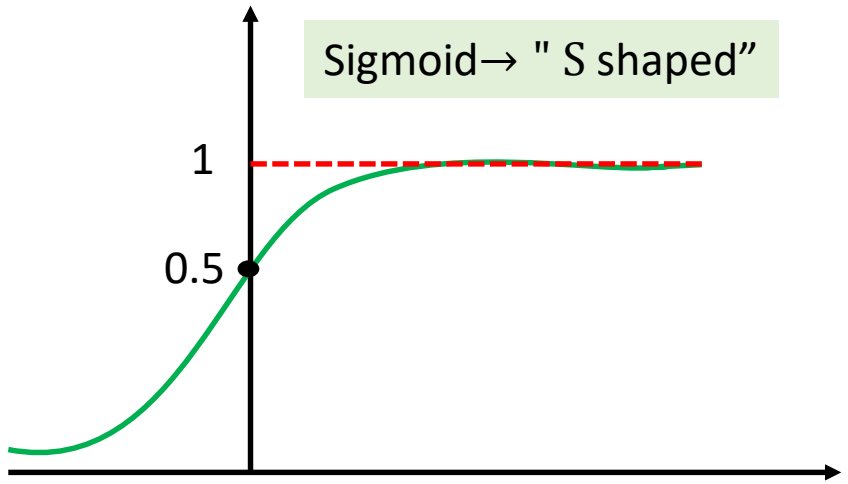


Logistic Regression (Parameter Fitting with Maximum Likelihood)

$$\sigma(x) = h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x)}}$$

Where $0 \leq \sigma(x) \leq 1$ is the sigmoid function used in logistic regression.

$\sigma(x)$= Gives the probability



$$P(y=1|x=x_i) = p(x_i) = \sigma(x_i) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_i)}}$$

Rearrange: $z = \theta_0 + \theta_1 x_i = \ln\left[\frac{p(x_i)}{1-p(x_i)}\right]$

$$1 - p(x_i) = 1 - \frac{1}{1+e^{-(\theta_0+\theta_1 x_i)}}$$

$$= \frac{e^{-(\theta_0+\theta_1 x_i)}}{1+e^{-(\theta_0+\theta_1 x_i)}}$$

Logistic Regression (Parameter Fitting with Maximum Likelihood)

Probability for being in category 1,

$$p(y = 1|x = x_i) = p(x_i) = \sigma(x_i) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_i)}} \quad \text{---------- (1)}$$

Probability for being in category 0,

$$p(y = 0|x = x_i) = 1 - p(x_i) \quad \text{---------- (2)}$$

By equation (1) and (2),

$$P(y = y_i|x = x_i) = p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i} \text{ where } y_i \in \{0,1\}$$

when $y_i = 1,\ p(y = y_i|x = x_i) = p(x_i)$ ---------- (3)

when $y_i = 0,\ p(y = y_i|x = x_i) = 1 - p(x_i)$ ---------- (4)

Fitting using maximum likelihood,

$$p(x_i) = \frac{1}{1 + e^{-(\theta_0+\theta_1 x_i)}}$$

$$P(y = y_i|x = x_i) = p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}$$

$y_i = 0$

$y_i = 1$

Logistic Regression (Parameter Fitting with Maximum Likelihood)

For n independent training example, we can write

$$= P_1 \times P_2 \times \cdots \times P_n$$

$$= \prod_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i}$$

For n data points the likelihood function.

Find θ_0 and $\theta_i, i = 1,2 \ldots n$ that maximize L or $\log(L)$.

$$\sigma(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n)}}$$

$$\log\left[\frac{p(x_i)}{1-p(x_i)}\right] = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

$$\log(L) = \sum_{i=1}^{n} y^{(i)} \log p(x_i) + \left(1 - y^{(i)}\right) \log\left(1 - p(x_i)\right)$$

It does not have a closed- form solution so we cannot write down a simple expression of the estimator

Training Data

x_i	y_i (1 or 0)	Prob.
x_1	y_1	P_1
x_2	y_2	P_2
...	...	...
x_n	y_n	P_n

Logistic Regression (Parameter Fitting with Maximum Likelihood)

To find the parameters maximize $\log(L)$ equivalently we can minimize $\log(likelihood)$ (NLL) to maintain uniformity with other algorithm:

$$J(w) = -\log(L) = -\sum_{i=1}^{n} \{y^{(i)} \log p(x_i) + (1 - y^{(i)}) \log(1 - p(x_i)\}$$

$$= -\sum_{i=1}^{n} \{y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}\}$$

References

- ✓ Linear Algebra and Learning from Data (2019), Gilbert Strang, Wellesley Cambridge Press
- ✓ Machine Learning: A Probabilistic Perspective, Kevin P. Murphy (MIT Press), 2021 edition
- ✓ Deisenroth MP, Faisal AA, Ong CS. Mathematics for machine learning. Cambridge University Press; 2020 Apr 23.

SWAYAM & NPTEL COURSE ON

Mathematics for Machine Learning

by

Prof. Debjani Chakraborty

DEPARTMENT OF MATHEMATICS
IIT Kharagpur

Module 8.5

Lecture 40: MLE for Naïve Bayes Model

Concepts covered

- ✓ **MLE for Naïve Bayes Model**

The Method of Maximum Likelihood Estimation

If k random samples $X_1, X_2, .., X_k$ are drawn from a population with probability function $f(X|\theta)$, then joint probability function L of $X_1, X_2, .., X_k$ is,

$$\boldsymbol{L(\theta) = f(x_1|\theta) f(x_2|\theta) \cdots f(x_k|\theta) = \prod_{i=1}^{k} f(x_I|\theta)}$$

***L*: Likelihood function.**

Since, $X_1, X_2, .., X_k$ are independently and identically distributed.

Objective: To find θ, such that Likelihood function is maximum

i.e. $\hat{\theta} = \underset{\theta}{argmax}\ L(\theta)$

Assume,

1. we have some prior knowledge $p(x)$
2. a random variable y, assumes value from random sample
3. We can update the knowledge $p(x)$ after observing y, $P(x/y)$

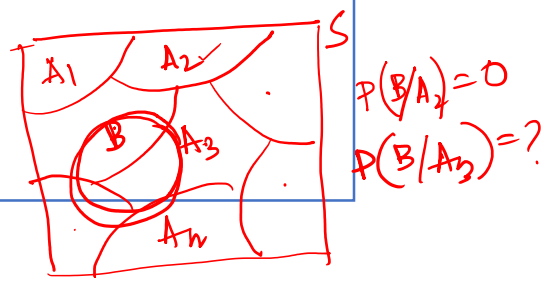
$$\underbrace{P(x/y)}_{\text{posterior}} = \frac{\overbrace{P(y/x)}^{\text{likelihood}}\overbrace{P(x)}^{\text{prior}}}{\underbrace{P(y)}_{\text{evidence}}}$$

Bayes Rule :

Hypothesis: (Equal division of ignorance): If nothing is known about the prior probabilities $P(A_1) \cdots P(A_n)$ then they are all equal.

Let $A_1, A_2 \cdots A_n$ denote a disjoint partition of a outcome set S and let B be any event. Let $P(A_i) \neq 0, i = 1,2 \cdots n$ and $P(B) \neq 0$, then for $i = 1,2 \cdots n$

$$P(A_i/B) = \frac{P(B/A_i)P(B)}{\sum_{i=1}^{n} P(B/A_i)P(B)}$$



Result: Let $A_1, A_2 \cdots A_n$ denote a disjoint partition of a outcome set S and let B be any event. Let $P(A_i) \neq 0, i = 1,2 \cdots n$ and $P(B) \neq 0$, then for $i = 1,2 \cdots n$

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B/A_i)$$

Naïve Bayes Model

Example: Classifying documents using bag of words

Document classification is the problem of classifying text documents into different categories. One simple approach is to represent each document as a binary vector,

each word is present or not, so $x_{ij} = 1$ iff word j occurs in document i, otherwise $x_{ij} = 0$

Email Subject	Label
Offer: Double your income	Spam
Free gift	spam
You have been selected for Award	spam
Open for a free webinar	Not Spam
Meeting scheduled for tomorrow	Not Spam
Reminder: Review meeting agenda	Not spam

In Naive Bayes, we make a "naive" assumption that all features are conditionally independent given the class label.

Example: Classifying documents using bag of words

Email Subject	Label
Offer: Double your income	Spam
Free gift	spam
You have been selected for Award	spam
Open for a free webinar	Not Spam
Meeting scheduled for tomorrow	Not Spam
Reminder: Review meeting agenda	Not spam

$$P(Spam \mid \text{"offer"}, \text{"free"})$$

$$= \frac{P(\text{Spam}) \cdot P(\text{"offer"}|\text{Spam}) \cdot P(\text{"free"}|\text{Spam})}{P(\text{"offer"}, \text{"free"})}$$

$$P(spam) = .5, P(not\ spam) = .5$$

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$$

$$P(\text{“offer”}|\text{”}spam\text{”})$$

$$= \frac{\text{number of occurrences of feature} + \alpha}{\text{total number of instances of the given class} + \alpha * Dimension\ of\ the\ data}$$

$$= \frac{3+1}{3+2} = \frac{4}{5} = .8$$

$$P(\text{“offer”}|\text{”Not }spam\text{”}) = .2$$

Naïve Bayes Model

Classifying documents using bag of words

Email Subject	Label
Double your income	Spam
Offer: Free gift	spam
You have been selected for Award	spam
Open for a free webinar	Not Spam
Meeting scheduled for tomorrow	Not Spam
Reminder: Review meeting agenda	Not spam

$$P(Spam \mid \text{"offer"}, \text{"free"})$$

$$= \frac{P(\text{Spam}) \cdot P(\text{"offer"} \mid \text{Spam}) \cdot P(\text{"free"} \mid \text{Spam})}{P(\text{"offer"}, \text{"free"})}$$

$$P(spam) = .5, P(not\ spam) = .5$$

$$P(\text{“free”} \mid \text{”}spam\text{”}) = \frac{\text{number of occurrences of feature} + \alpha}{\text{total number of instances of the given class} + \alpha * Dimension\ of\ the\ data}$$

$$= \frac{1+1}{3+2} = \frac{2}{5} = .4$$

$$P(\text{“free”} \mid \text{”Not } spam\text{”}) = .6$$

Naïve Bayes Model

Classifying documents using bag of words

Email Subject	Label
Offer: Double your income	Spam
Free gift	spam
You have been selected for Award	spam
Open for a free webinar	Not Spam
Meeting scheduled for tomorrow	Not Spam
Reminder: Review meeting agenda	Not spam

$$P(Spam \mid \text{"offer"}, \text{"free"})$$

$$= \frac{P(\text{Spam}) \cdot P(\text{"offer"} \mid \text{Spam}) \cdot P(\text{"free"} \mid \text{Spam})}{P(\text{"offer"}, \text{"free"})}$$

$$P(spam) = .5, P(not\ spam) = .5$$

$$\frac{P(\text{Spam}) \cdot P(\text{"offer"} \mid \text{Spam}) \cdot P(\text{"free"} \mid \text{Spam})}{P(\text{"offer"}, \text{"free"})}\ equivalent\ to\ .5..8..4 = .16$$

Naïve Bayes Model

Document classification is the problem of classifying text documents into different categories.
One simple approach is to represent each document as a binary vector,

each word is present or not, so $x_{ij} = 1$ iff word j occurs in document i, otherwise $x_{ij} = 0$

We can then use the following class conditional density:

$$P(Spam \mid \text{"offer"}, \text{"free"})$$

$$= \frac{P(\text{Spam}) \cdot P(\text{"offer"} \mid \text{Spam}) \cdot P(\text{"free"} \mid \text{Spam})}{P(\text{"offer"}, \text{"free"})}$$

$$P(\text{class}_i / \text{feature}_{n \in [1..N]})$$

$$= P(\text{class}_i) \prod_{n=1}^{N} P(\text{feature}_n / c_i)$$

Fitting a naive Bayes model with maximum likelihood

$$P(\text{class}_i/\text{feature}_{n\in[1..n]}) = P(\text{class}_i)\prod_{n=1}^{N} P(\text{feature}_n/c_i)$$

We use the maximum likelihood method method in finding parameters that maximize the likelihood of the observed w_n data set of size D, $\{w_n, c_i\}_{i=1..C,n=1..N}$ in class c_i

Then likelihood under a certain model with parameters $f(X|\theta_i)$ of class c_i

$$L\left(D\middle|\theta_{i\in[1..C]}, f(X|\theta_i)\right) = \prod_{d=1}^{D}\left(P(c_i)\prod_{n=1}^{N} P(w_n/c_i)\right)$$

Taking log-likelihood

$$\log L = \sum_{d=1}^{D}\sum_{i=1}^{C}\log P(c_i) + \sum_{d=1}^{D}\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{v=1}^{V}\log P(w_n|c_i),$$

Fitting a naive Bayes model with maximum likelihood

$$P(\text{class}_i/\text{feature}_{n\in[1..n]}) = P(\text{class}_i)\prod_{n=1}^{N} P(\text{feature}_n/c_i)$$

Taking log-likelihood

$$\log L = \sum_{d=1}^{D}\sum_{i=1}^{C} \log P(c_i) + \sum_{d=1}^{D}\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{v=1}^{V} \log P(w_n|c_i),$$

e.g. D: Number of emails C: Number of classes
V: Total number of words in the vocabulary
N: Number of words / features we are looking for

$$\sum_{i=1}^{C} P(c_i) = 1$$

$$\sum_{n=1}^{N}\sum_{i=1}^{C} P(w_n|c_i) = 1$$

$$P(c_i) = \frac{Total\ number\ of\ time\ we\ see\ the\ class\ c_i}{Number\ of\ data}$$

Fitting a naive Bayes model with maximum likelihood

$$P(\text{class}_i/\text{feature}_{n\in[1..n]}) = P(\text{class}_i)\prod_{n=1}^{N} P(\text{feature}_n/c_i)$$

Taking log-likelihood

$$\log L = \sum_{d=1}^{D}\sum_{i=1}^{C} \log P(c_i) + \sum_{d=1}^{D}\sum_{c=1}^{C}\sum_{n=1}^{N}\sum_{v=1}^{V} \log P(w_n|c_i)\,,$$

e.g. D: Number of emails C: Number of classes
V: Total number of words in the vocabulary
N: Number of words / features we are looking for

$$P(c_i) = \frac{Total\ number\ of\ time\ we\ see\ the\ class\ c_i}{Number\ of\ data}$$

Each c_i has its own probability of success θ_i in the probability distribution

$$\theta_i^{MLE} = \frac{\text{number of time we see the word in the class } c_i}{\text{Total number of times we see the word in all classes}}$$

Conclusion

Maximum likelihood estimation based on training data

References

- ✓ Linear Algebra and Learning from Data (2019), Gilbert Strang, Wellesley Cambridge Press
- ✓ Machine Learning: A Probabilistic Perspective, Kevin P. Murphy (MIT Press), 2021 edition
- ✓ Deisenroth MP, Faisal AA, Ong CS. Mathematics for machine learning. Cambridge University Press; 2020 Apr 23.


Thank You...
INDIAN INSTITUTE OF TECHNOLOGY
NPTEL
FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
1951
योगः कर्मसु कौशलम्
NPTEL ONLINE CERTIFICATION COURSES
IIT KHARGPUR