Xi'an Jiaotong-Liverpool University

# Pattern Recognition

## Lecture 07. Parametric methods: Maximum Likelihood Estimation (MLE)

Dr. Shanshan ZHAO & Dr. Yuxuan ZHAO

School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University

Academic Year 2023-2024

# Table of Contents

# Notations

- $X$ : The dataset observed
- $x$ : the random variable, i.e., the feature vector
- $x$ : the univariant , or a random variable in the feature vector
- $\theta$ : the parameters unknown in $p(x)$
- $N$ : Number of samples
- $p(\theta|X)$ or $p(X|\theta)$: we consider $\theta$ and $X$ as two random variables, this is to denote dependence between two variables
- $p(x_k;\theta)$: The semicolon means that it is the pdf with respect to $x_k$, the parameter of it is $\theta$ .

# Introduction

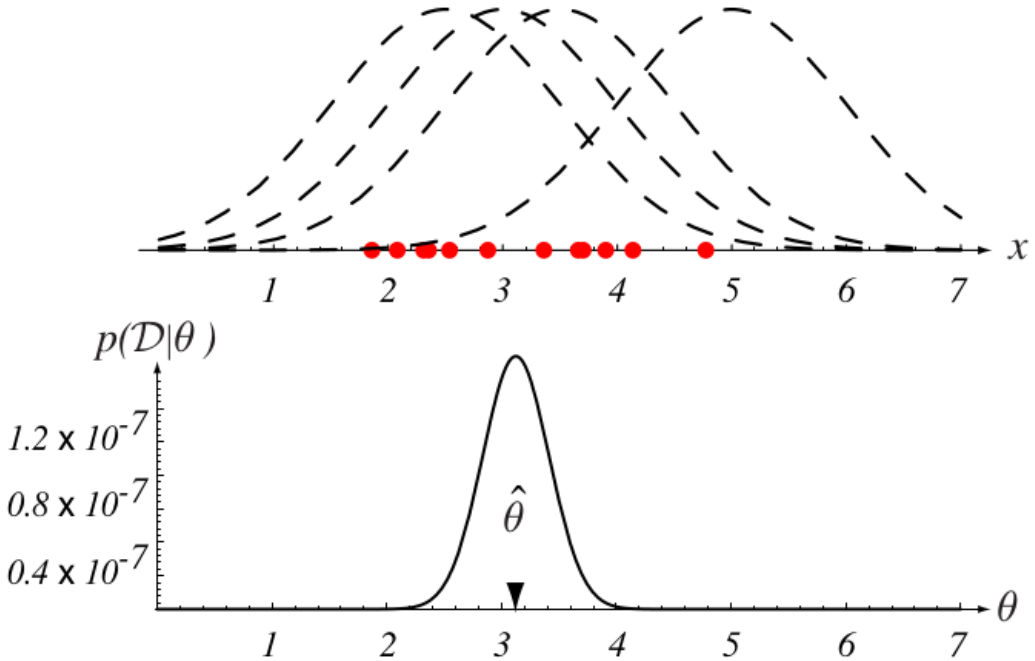
- So far, we have assumed that the **probability density functions are known**.(Which implies that the parameters of pdf are known).
- That is NOT the common case.
- The underlying pdf has to be estimated from the available data.

# Maximum Likelihood Estimation

- a technique used for estimating the parameters of a given distribution, using some observed data.
- e.g. A population is known to follow a normal distribution but the mean and variance are unknown.
- MLE can be used to estimate them using a limited sample of the population, by finding particular values of the mean and variance so that the observation is the **most likely** result to have occurred.

# Maximum Likelihood Intuition

- Assumption: Gaussian distribution.
- top figure: several training points in one dimension drawn from Gaussian with known variance and unknown mean.
- bottom figure: the likelihood as a function of the mean.



# Likelihood

- Observe some data $X = \{x_1, \ldots, x_N\}$
- Assume that the data is drawn from $p(X; \mu, \sigma^2) = \prod_{i=1}^{N} p(x_i; \mu, \sigma^2)$
- To find parameters is to maximize $p(X; \mu, \sigma^2)$ (maximize likelihood that data was generated by model)

**Example:** measure the weight of your cats.

# Probability vs Likelihood

- Probabilities are the areas under a fixed distribution $p(data|distribution)$
- Likelihoods are the y-axis values for fixed data points with distributions that can be moved $L(distribution|data)$

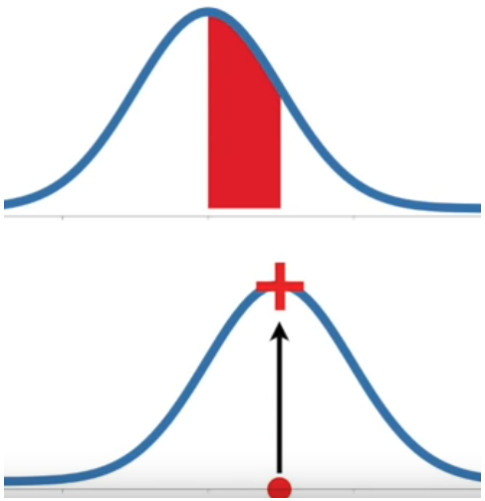


Figure 1: $x$ axis: weight of cats; $y$ axis of top figure : $pdf$; $y$ axis of bottom figure: likelihood

# Preliminaries

Let us consider an $M$-class problem with feature vectors distributed according to $p(x|\omega_i), i = 1, 2, ..., M$, We assume that these likelihood functions are given in a *parametric* form and that the corresponding parameters from the vectors $\theta_i$ which are unknown.

- Let $x_1, x_2, ..., x_N$ be random **samples** drawn from pdf $p(x; \theta)$
- Assuming *statistical independence* between the different samples (i.i.d. — independent and identically distributed)

$$p(X; \theta) \equiv p(x_1, x_2, ..., x_N; \theta) = \prod_{k=1}^{N} p(x_k; \theta) \tag{1}$$

## Preliminaries

$$p(X;\theta) \equiv p(x_1, x_2, ..., x_N;\theta) = \prod_{k=1}^{N} p(x_k;\theta)$$

This is a function of $\theta$, and it is also known as the likelihood function of $\theta$ with respect to $X$.

*The maximum likelihood(ML) method estimates $\theta$ so that the likelihood function takes its maximum value*, that is,

$$\hat{\theta}_{ML} = arg \max_{\theta} \prod_{k=1}^{N} p(x_k;\theta) \tag{2}$$

# Maximum Likelihood Estimation

A necessary condition that $\hat{\theta}_{ML}$ must satisfy in order to be a maximum is the gradient of the likelihood function with respect to $\theta$ to be zero, that is

$$\frac{\partial \prod_{k=1}^{N} p(x_k;\theta)}{\partial \theta} = 0 \tag{3}$$

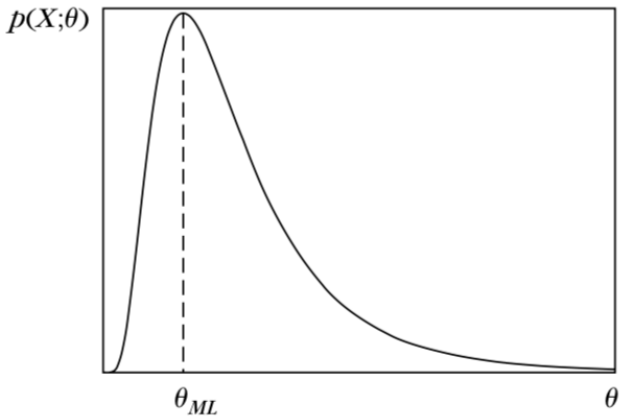
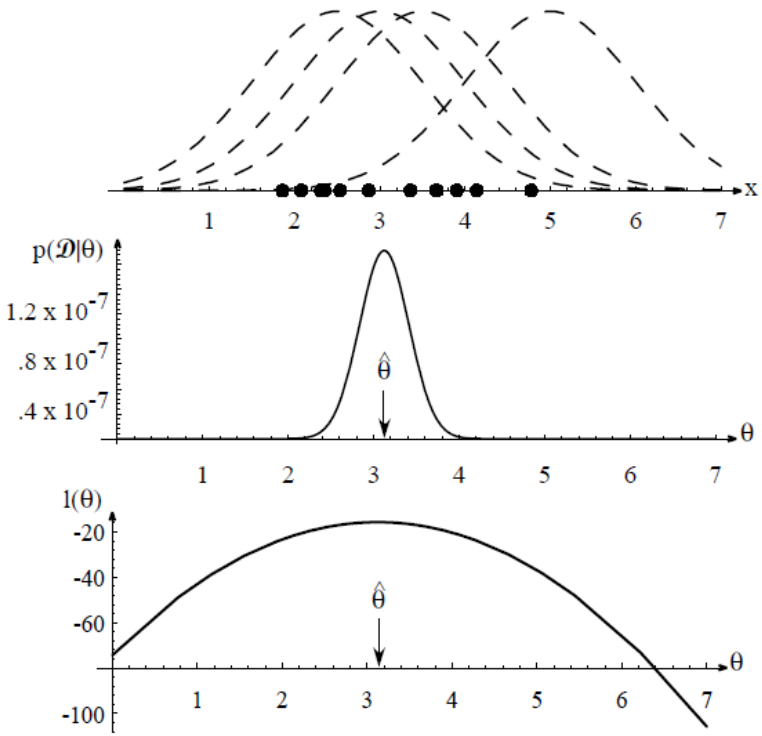



Figure 2: The maximum likelihood estimator $\theta_{ML}$corresponds to the peak of $p(X;\theta)$

# Maximum Likelihood Estimation




(1)top: several training points in one dimension drawn from Gaussian with known variance and unknown mean. (2)middle: the likelihood as a function of the mean. (3) logarithm of the likelihood.

# Maximum Likelihood Estimation

- Logarithmic funciton is monotonical

We define the *log-likelihood function* as

$$L(\theta) = \ln \prod_{k=1}^{N} p(x_k; \theta) \tag{4}$$

Gradient of likelihood function with respect to $\theta$ is equivalent to,

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{\partial \ln p(x_k; \theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{1}{p(x_k; \theta)} \frac{\partial p(x_k; \theta)}{\partial \theta} = 0 \tag{5}$$

# MLE *example 1.*

**Gaussian with unknown mean and known covariance**

Let $x_1, x_2, ..., x_N$ be vectors stemmed from a normal distribution with known convariance matrix and unknown mean, that is

$$p(x_k; \mu) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x_k - \mu)^T \Sigma^{-1}(x_k - \mu)) \quad (6)$$

where $l$ is the dimension of vector $x_k$ $(k = 1, ..., N)$.
Obtain the ML estimate of the unknown mean vector.
For $N$ available samples we have

$$L(\mu) \equiv \ln \prod_{k=1}^{N} p(x_k; \mu) = -\frac{N}{2} \ln((2\pi)^l |\Sigma|) - \frac{1}{2} \sum_{k=1}^{N} (x_k - \mu)^T \Sigma^{-1} (x_k - \mu) \quad (7)$$

## MLE *example 1.*

Taking the gradient with respect to $\mu$, we obtain

$$\frac{\partial L(\mu)}{\partial \mu} \equiv \begin{bmatrix} \frac{\partial L}{\partial \mu_1} \\ \frac{\partial L}{\partial \mu_2} \\ \vdots \\ \frac{\partial L}{\partial \mu_l} \end{bmatrix} = \sum_{k=1}^{N} \Sigma^{-1}(x_k - \mu) = 0 \tag{8}$$

then

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{k=1}^{N} x_k \tag{9}$$

That is, the ML estimate of the mean, for Gaussian densities, is the sample mean. However, this very "natural approximation" is not necessary ML optimal for non-Gaussian density functions.

# MLE *example 2.*

Assume that $N$ data points $x_1, x_2, ..., x_N$ have been generated by a one dimensional Gaussian pdf of known mean $\mu$, of of unknown variance. Derive the ML estimate of the variance.
The log-likelihood function for this case is given by

$$L(\sigma^2) = \ln \prod_{k=1}^{N} p(x_k; \sigma^2) = \ln \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} exp(-\frac{(x_k - \mu)^2}{2\sigma^2}) \tag{10}$$

## MLE *example 2.*

then

$$L(\sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k - \mu)^2 \tag{11}$$

Taking derivative of the above with respect to $\sigma^2$ and equating to zero, we obtain

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{k=1}^{N}(x_k - \mu)^2 = 0 \tag{12}$$

## MLE *example 2.*

finally the ML estimate of $\sigma^2$ results as the solution of:

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2 \tag{13}$$

Observe that, for the finite $N$, in Eq.13 is a biased estimate of the variance. In fact,

$$E[\hat{\sigma}^2_{ML}] = \frac{1}{N}\sum_{k=1}^{N} E[(x_k - \mu)^2] = \frac{N-1}{N}\sigma^2 \tag{14}$$

where $\sigma^2$ is the true variance of the Gaussian pdf. However, for large values of $N$, we have

$$E[\hat{\sigma}^2_{ML}] = (1 - \frac{1}{N})\sigma^2 \approx \sigma^2 \tag{15}$$

# MLE *example 3.*

## Possion MLE

### Exercise

How to calculate the MLE for the parameter $\lambda$ of a Poisson distribution.

1. Write the pdf

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

2. Write the likelihood function
3. Write the natural log likelihood function
4. Calculate the derivative of the natural log likelihood function with respect to $\lambda$
5. Set the derivative equal to zero and solve for $\lambda$

# MLE *example 3. solution*

1. the probability density function :

$$f(x_i;\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

2. the likelihood function :

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

3. the natural log likelihood function :

$$l(\lambda) = \ln L(\lambda) = \sum_{i=1}^{N} \ln[\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}] = \sum_{i=1}^{N}[-\lambda + x_i \ln\lambda - \ln(x_i!)]$$

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

# MLE *example 3. solution*(cont.)

4. Calculate the derivative of the natural log likelihood function with respect to $\lambda$

$$l'(\lambda) = -N + \frac{\sum_{i=1}^{N} x_i}{\lambda}$$

5. Set the derivative equal to zero and solve for $\lambda$

$$l'(\lambda) = 0 \Rightarrow \lambda = \frac{\sum_{i=1}^{N} x_i}{N}$$

6. $$l''(\lambda_0) < 0 \Rightarrow \lambda_0 \text{ is a maximum extreme point}$$

# MLE *example 4.*

## Binomial MLE

Assumptions:

- you are playing basketball
- 6 free throws and 5 of them make it in
- $X = \#$ of success, here $X = 5$, and $N = 6$
- assume X$\sim$ $Bin(n, p)$

## Exercise

Find the maximum likelihood estimate of p

# MLE *example 4. solution*

1. The probability density function of a Binomial distribution is :

$$f(x_i; p) = \binom{n}{x} p^x (1-p)^{n-x}$$

2. The likelihood function:

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

3. the natural log likelihood function :

$$l(p) = \ln L(p) = \ln\left(\binom{n}{x}\right) + x \ln p + (n-x) \ln(1-p)$$

# MLE *example 4. solution*

4. Calculate the derivative of the natural log likelihood function with respect to $\lambda$

$$l'(p) = \frac{x}{p} - \frac{n-x}{1-p}$$

5. Set the derivative equal to zero and solve for $\lambda$

$$l'(p) = 0 \Rightarrow \frac{x}{p} = \frac{n-x}{1-p} \Rightarrow \hat{p} = \frac{x}{n} = \frac{5}{6} \text{ (ML Estimate)}$$

6. 

$$l''(\hat{p}) < 0 \Rightarrow \hat{p} \text{ is a maximum extreme point}$$

# Thank You !

$\mathcal{Q}$ & $\mathcal{A}$