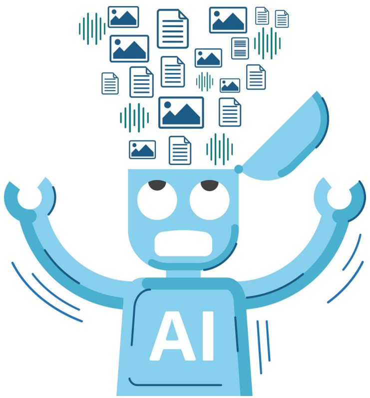
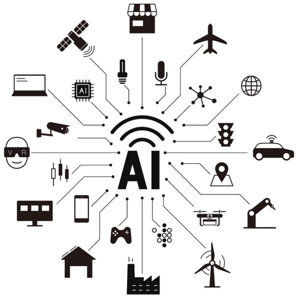
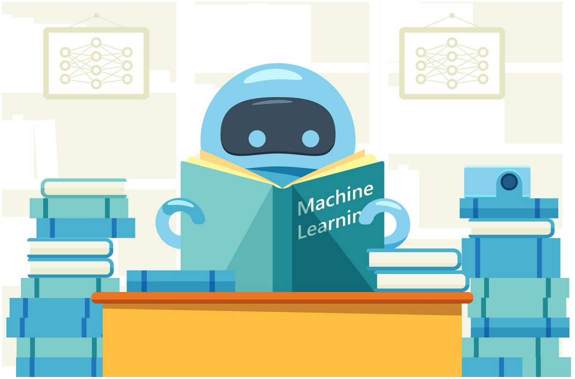
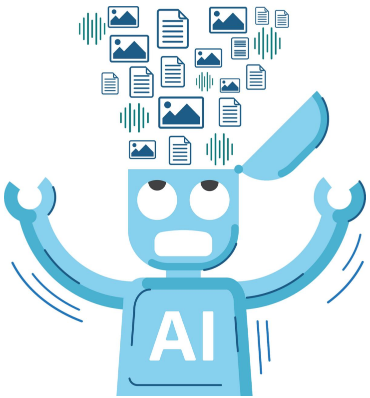


机器学习
Machine Learning

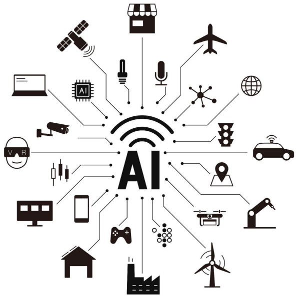


软 件 学 院　罗 昕
luoxin@sdu.edu.cn



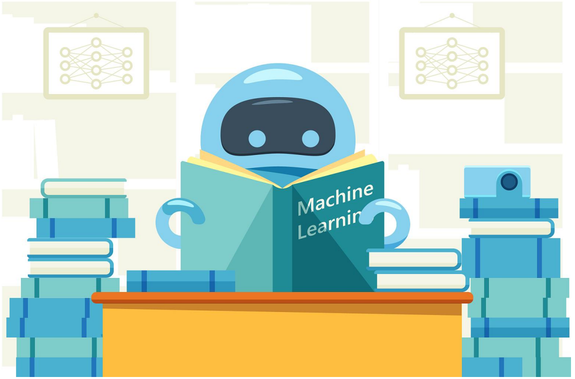


Chapter 3
Parameter Estimation



软 件 学 院 罗 昕

luoxin@sdu.edu.cn



Contents

MIMA

- Introduction
- Maximum-Likelihood Estimation
- Bayesian Estimation

Preliminaries and Notations

MIMA

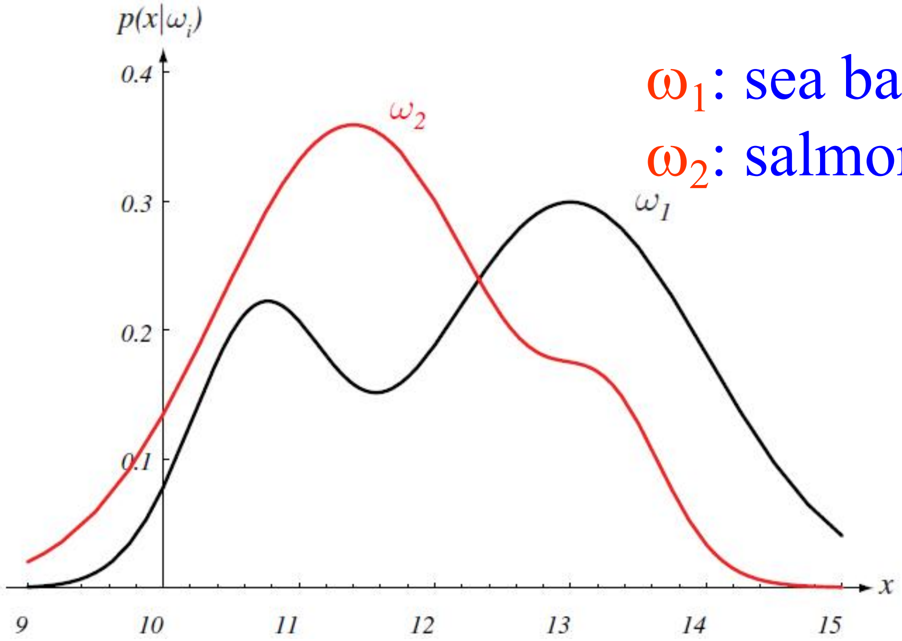
$\omega_i \in \{\omega_1, \omega_2, \ldots, \omega_c\}$:	a state of nature	
$P(\omega_i)$:	prior probability	先验概率
$\mathbf{x}$:	feature vector	
$p(\mathbf{x})$:	evidence probability	
$p(\mathbf{x} \mid \omega_i)$:	class-conditional density / likelihood	类条件概率密度/似然
$P(\omega_i \mid \mathbf{x})$:	posterior probability	后验概率

sea bass	salmon
鲈鱼	鲑鱼

ω_1: sea bass

ω_2: salmon

An example



class-conditional pdf for *lightness*

ω_1: sea bass

ω_2: salmon

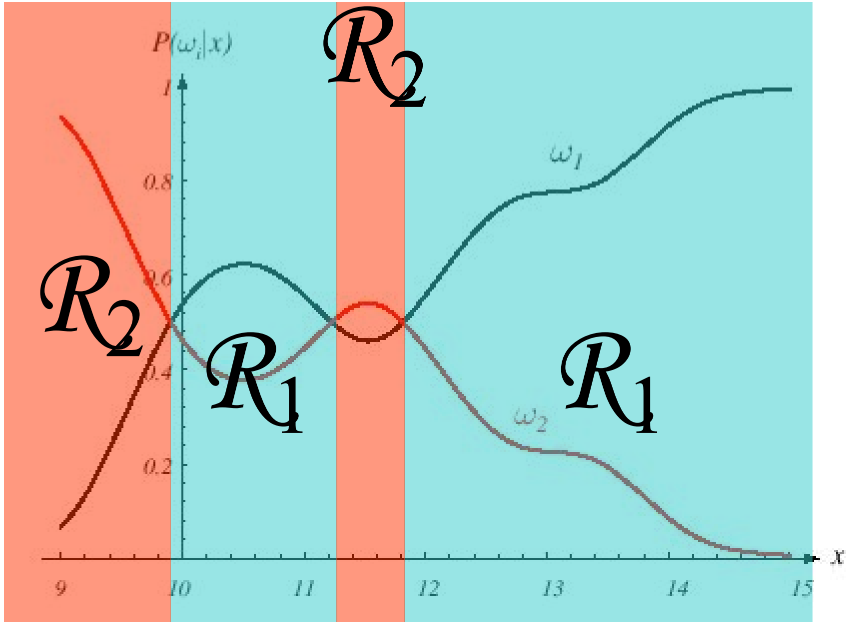
$P(\omega_1)=2/3$

$P(\omega_2)=1/3$

What will the posterior probability for either type of fish look like?

An example

MIMA



posterior probability for either type of fish

h-axis: lightness of fish scales
v-axis: posterior probability for each type of fish
Black curve: sea bass
Red curve: salmon

- For each value of x, the higher curve yields the output of Bayesian decision
- For each value of x, the posteriors of either curve sum to 1.0

Bayes Theorem

MIMA

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x} \mid \omega_i) P(\omega_i)$$

Thomas Bayes
(1702-1761)

Bayesian Theorem

MIMA

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} \mid \omega_i) P(\omega_i)$$

- To compute posterior probability $P(\omega_i \mid \mathbf{x})$, we need to know:

$$p(\mathbf{x} \mid \omega_i) \qquad P(\omega_i)$$

How can we get these values?

Feasibility of Bayes Formula

MIMA

- To compute posterior probability, we need to know prior probability and likelihood

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})} \quad \left(\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \right)$$

How do we know these probabilities?

➢ A simple solution: Counting Relative frequencies

Example - Counting

MIMA

- Collecting samples
  - Suppose we have randomly picked 1209 cars in the campus, got prices from their owners, and measured their heights.
- Compute $P(\omega_1)$ and $P(\omega_2)$

cars in ω_1 : 221

cars in ω_2 : 988

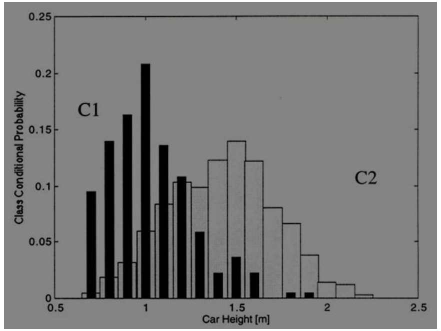
$$P(\omega_1) = \frac{221}{1209} = 0.183$$

$$P(\omega_2) = \frac{988}{1209} = 0.817$$

Example - Counting (Cont.)

MIMA

- Compute $P(x|\omega_1)$ $P(x|\omega_2)$
  - Discretize the height spectrum (say [0.5m, 2.5m]) into 20 intervals each with length 0.1m, and then count the number of cars falling into each interval for either class



- Suppose $x = 1.05$, which means

  that x falls into interval

  $I_x = [1.0m, 1.1m]$

For ω_1, # cars in I_x is 46,

For ω_2, # cars in I_x is 59,

$$P(x = 1.05 \mid \omega_1) = \frac{46}{221} = 0.2081$$

$$P(x = 1.05 \mid \omega_2) = \frac{59}{988} = 0.0597$$

Feasibility of Bayes Formula

MIMA

- To compute posterior probability, we need to know prior probability and likelihood

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})} \quad \left(\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \right)$$

How do we know these probabilities?

➢A simple solution: Counting Relative frequencies

➢**An advanced solution: Conduct Density estimation**

Contents

MIMA

- Introduction
- **Maximum-Likelihood Estimation**
- Bayesian Estimation

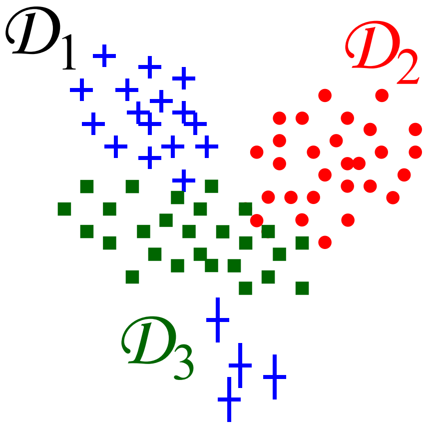
Samples

MIMA

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_c\}$$

The samples in D_j are drawn independently according to the probability law $p(x|\omega_j)$.

That is, examples in D_j are i.i.d. random variables, i.e., independent and identically distributed. 独立同分布



It is easy to compute the prior probability:

$$P(\omega_i) = \frac{|D_j|}{\sum_{i=1}^{c} |D_i|}$$

Samples

MIMA

- For class-conditional pdf:
  - Case I: $p(x|\omega_j)$ has certain parametric form
  - Case II: $p(x|\omega_j)$ doesn't have parametric form

Samples

- For class-conditional pdf:
  - Case I: $p(x|\omega_j)$ has certain parametric form
    - e.g.

$$p(\mathbf{x}\,|\,\omega_j) \sim N(\underbrace{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}_{\theta_j})$$

$$\theta_j \Rightarrow \boldsymbol{\theta}_j = (\theta_1, \theta_2, \ldots, \theta_m)^T$$

    - If $X \in R^d$ θ_j contains "$d+d(d+1)/2$" free parameters.

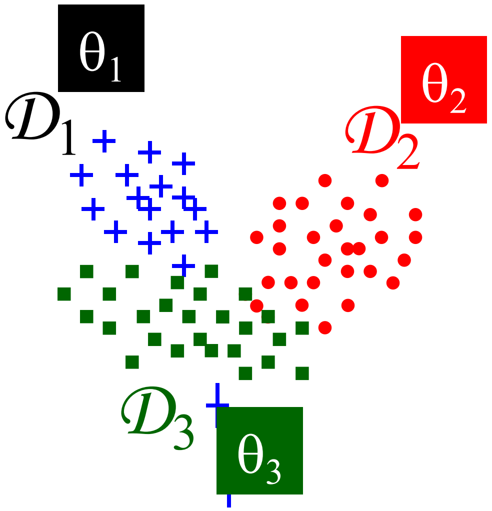
  - Case II: $p(x|\omega_j)$ doesn't have parametric form
    - Next chapter.

Goal

MIMA

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_c\}$$

$$\boxed{p(\mathbf{x} \mid \omega_j) \equiv p(\mathbf{x} \mid \boldsymbol{\theta}_j)}$$



Use $\mathcal{D}_j$ to estimate the unknown parameter vector θ_j

$$\boldsymbol{\theta}_j = (\theta_1, \theta_2, \ldots, \theta_m)^T$$

Estimation Under Parametric Form

MIMA

- Maximum-Likelihood Estimation

View parameters as quantities whose values are fixed but unknown → Estimate parameter values by maximizing the likelihood (probability) of observing the actual examples.

- Bayesian Estimation

View parameters as random variables having some known prior distribution → Observation of the actual training examples transforms parameters' prior into posterior distribution. (via Bayes rule)

概率与似然

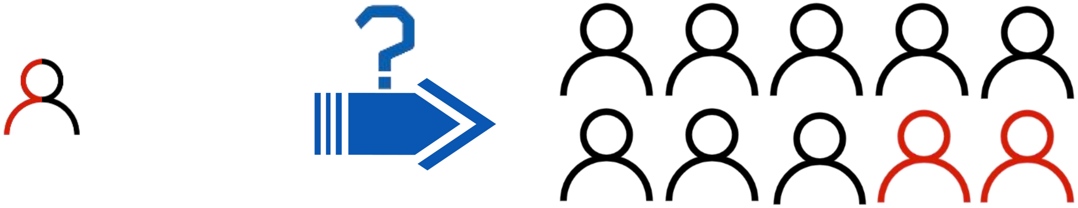
概 率

描述了参数已知时的随机变量的输出结果

似 然

用来描述已知随机变量输出结果时，**未知参数的可能取值**

概率与似然

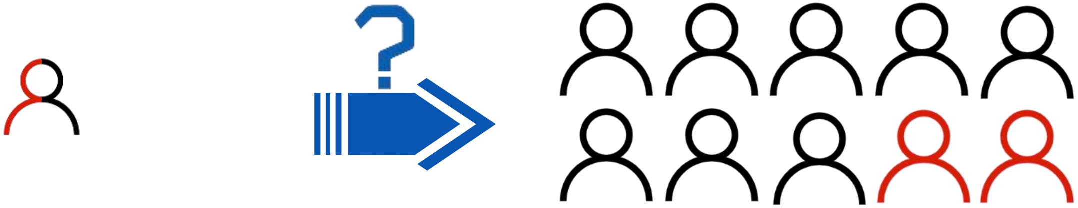


概　率

已知参数感染率θ

推测密切接触者感染
的各种情况的可能性

概率与似然



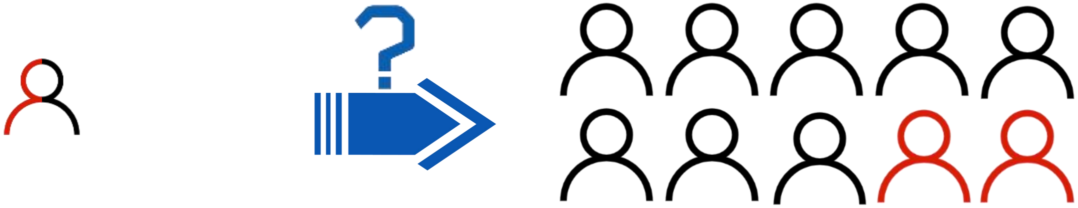
概 率

参数感染率θ为0.1

密切接触者感染的可能性为0.1

可以推测，在10个密切接触者中，出现2例确诊病例的概率为：

概率与似然



概 率

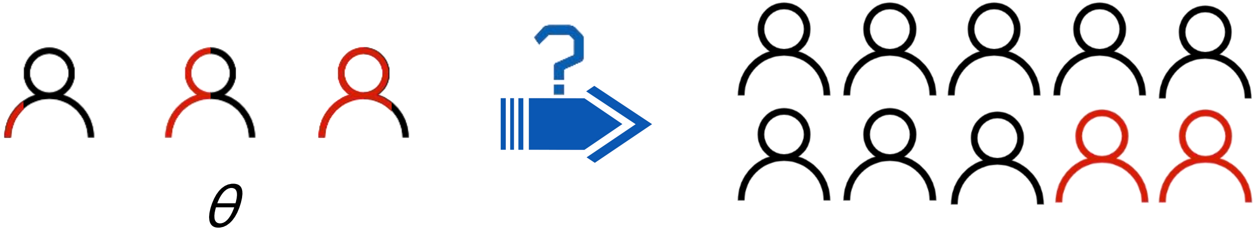
参数感染率θ为0.1

密切接触者感染的可能性为0.1

可以推测，在10个密切接触者中，出现2例确诊病例的概率为：

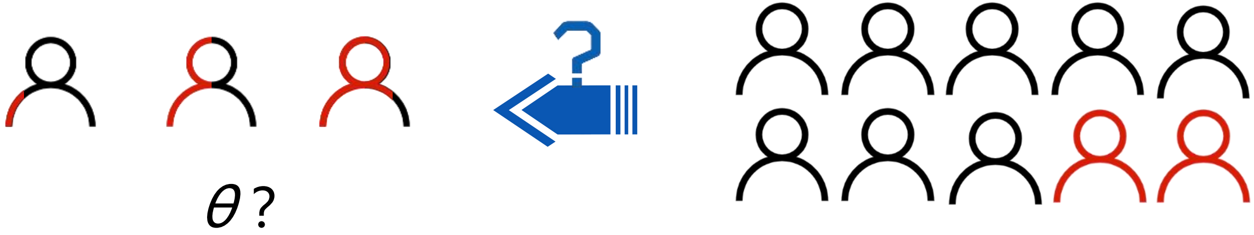
$$\binom{10}{2}0.1^2(1-0.1)^8 \approx 0.19$$

概率与似然



概 率

当我们对参数并不清楚，要通过采样的情况去推测参数



似 然

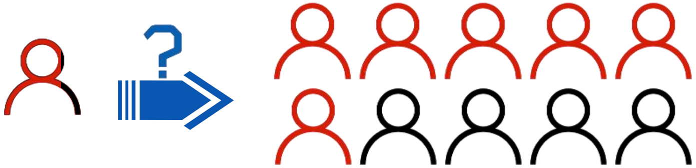
概率与似然

似然：通过证据，对参数θ进行推断。

最大似然估计：得到最可能的参数的过程。

极大似然估计

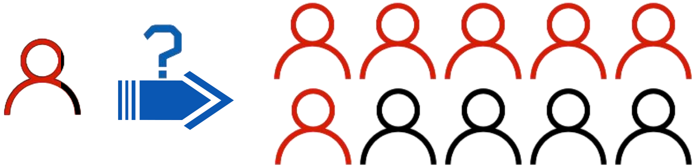
- 某地一天内增长无症状感染者2例，密切接触者10人并采取了相应的隔离举措，发现6人为阳性。



极大似然估计

MIMA

- 某地一天内增长无症状感染者2例，密切接触者10人并采取了相应的隔离举措，发现6人为阳性。



如果密切接触者感染的感染率为0.5，出现这个结果的可能性是：

$$\binom{10}{6}0.5^6(1-0.5)^4 \approx 0.21$$

极大似然估计

MIMA

- 某地一天内增长无症状感染者2例，密切接触者10人并采取了相应的隔离举措，发现6人为阳性。

如果密切接触者感染的感染率为0.5，出现这个结果的可能性是：

$$\binom{10}{6} 0.5^6 (1-0.5)^4 \approx 0.21$$

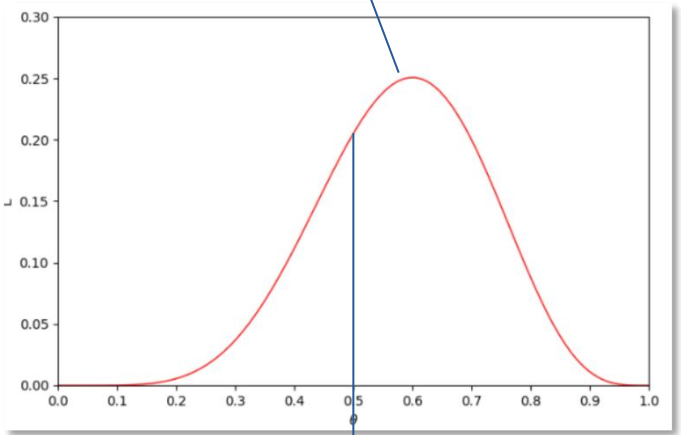
如果密切接触者感染的感染率为0.6，出现这个结果的可能性是：

$$\binom{10}{6} 0.6^6 (1-0.6)^4 \approx 0.25$$

θ=0.6作为参数的可能性是θ=0.5作为参数的可能性的1.19倍

极大似然估计

参数θ为0.6时，概率较大

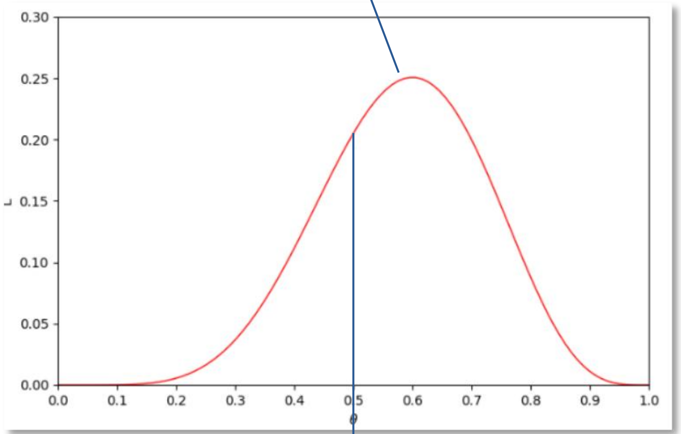


θ=0.5也是有可能的，
虽然可能性小一点

极大似然估计

参数θ为0.6时，概率较大

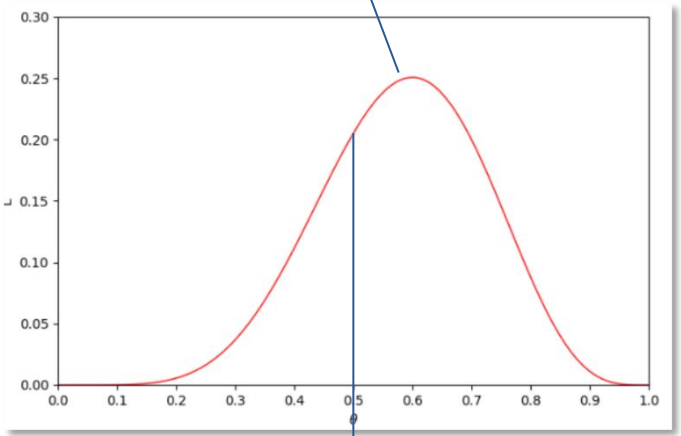
$$L(\theta) = \binom{10}{6} \theta^6 (1-\theta)^4$$



θ=0.5也是有可能的，
虽然可能性小一点

极大似然估计

参数θ为0.6时，概率较大

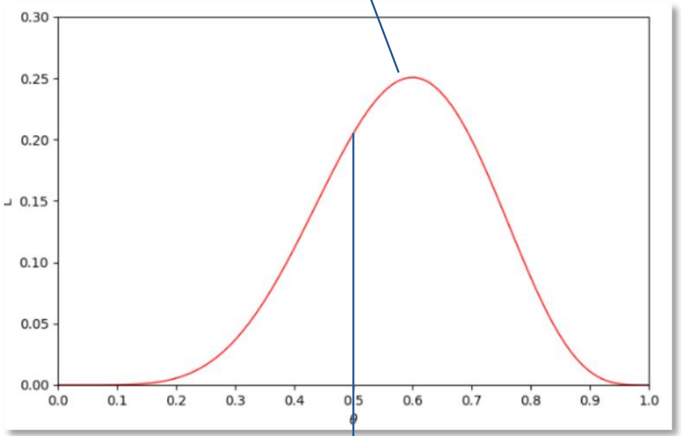


$$L(\theta) = \binom{10}{6}\theta^6(1-\theta)^4$$

θ=0.5也是有可能的，
虽然可能性小一点

极大似然估计

参数θ为0.6时，概率较大



θ=0.5也是有可能的，
虽然可能性小一点

$$L(\theta) = \binom{10}{6}\theta^6(1-\theta)^4$$

似然函数是推测参数的分布。

而求最大似然估计的问题，就变成了求似然函数的极值。

求似然函数的极值 → 对似然函数求导

极大似然估计

- 从特殊到一般

- 最大似然估计针对多次实验。用$x_1, x_2, \dots, x_N$表示每次实验结果，因为每次实验都是独立的，所以似然函数可以写作：

$$L(\theta) = p(\boldsymbol{x}_1|\theta)p(\boldsymbol{x}_2|\theta)\dots p(\boldsymbol{x}_N|\theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$$

极大似然估计

MIMA

- 从特殊到一般

- 最大似然估计针对多次实验。用$x_1, x_2, \ldots, x_N$表示每次实验结果，因为每次实验都是独立的，所以似然函数可以写作：

$$L(\theta) = p(\boldsymbol{x}_1|\theta)p(\boldsymbol{x}_2|\theta)\ldots p(\boldsymbol{x}_N|\theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$$

似然函数

- 则此时可写为：

$$\hat{\theta} = arg \max_{\theta} L(\theta)$$

$$\hat{\theta} = arg \max_{\theta} \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$$

极大似然估计

$$\hat{\theta} = arg \max_{\theta} \; L(\theta)$$

$$\hat{\theta} = arg \max_{\theta} \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$$

通常，使用对数似然

$$\hat{\theta} = arg \max_{\theta} \; LL(\theta)$$

$$LL(\theta) = log \, L(\theta)$$

$$\hat{\theta} = arg \max_{\theta} \; \sum_{n=1}^{N} log \, p(\boldsymbol{x}_n|\theta)$$

Maximum-Likelihood Estimation

MIMA

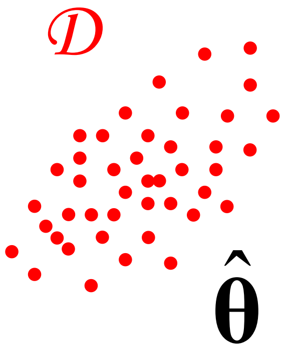
- Because each class is considered individually, the subscript used before will be dropped.

Maximum-Likelihood Estimation

- Because each class is considered individually, the subscript used before will be dropped.

- Now the problem becomes:

Given a sample set $\mathcal{D}$, whose elements are drawn independently from a population possessing a known parameter form, say $p(x|\theta)$, we want to choose a $\hat{\theta}$ that will make $\mathcal{D}$ to occur most likely.



Maximum-Likelihood Estimation (Cont.)

MIMA

- Criterion of ML

$$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$$

- By the independence assumption, we have:

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = p(\mathbf{x}_1 \mid \boldsymbol{\theta})\, p(\mathbf{x}_2 \mid \boldsymbol{\theta}) \cdots p(\mathbf{x}_n \mid \boldsymbol{\theta})$$

- The Likelihood function:

$$L(\boldsymbol{\theta} \mid \mathcal{D}) = p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

- The maximum-likelihood estimation:

$$\boxed{\hat{\boldsymbol{\theta}} = \arg\max_{\theta} L(\theta \mid D)}$$

Maximum-Likelihood Estimation (Cont.)

MIMA

- Often, we resort to maximize the log-likelihood function

$$l(\boldsymbol{\theta} \mid \mathcal{D}) = \ln L(\boldsymbol{\theta} \mid \mathcal{D}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta} \mid \mathcal{D})$$

$$\Updownarrow$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \mathcal{D})$$

why?

Maximum-Likelihood Estimation (Cont.)

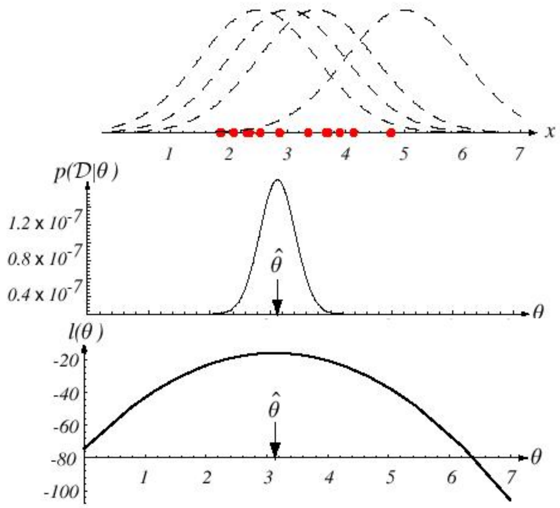
- Often, we resort to maximize the log-likelihood function

$$l(\boldsymbol{\theta} \mid \mathcal{D}) = \ln L(\boldsymbol{\theta} \mid \mathcal{D}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta} \mid \mathcal{D})$$

$$\Updownarrow$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \mathcal{D})$$



Maximum-Likelihood Estimation (Cont.)

MIMA

- Find the extreme values using the method in differential calculus.

- Gradient Operator
  - Let $f(\theta)$ be a continuous function, where $\theta=(\theta_1, \theta_2,\ldots, \theta_n)^T$.

Gradient Operator

$$\nabla_{\boldsymbol{\theta}} = \left(\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \cdots, \frac{\partial}{\partial \theta_n} \right)^T$$

- Find the extreme values by solving

$$\nabla_{\boldsymbol{\theta}} f = 0$$

高斯分布的极大似然估计

- 情况一：

均值未知　方差已知

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

获得均值

- 情况二：

均值未知　方差未知

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

获得均值、方差

高斯分布的极大似然估计

MIMA

均值未知
方差已知

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$L(\boldsymbol{\mu}|\mathcal{D})=p(\mathcal{D}|\boldsymbol{\mu})=\prod_{k=1}^{n}p(\mathbf{x}_k|\boldsymbol{\theta})$$ (似然函数)

$$=\frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}}\prod_{k=1}^{n}\exp\left[-\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})\right]$$

高斯分布的极大似然估计

MIMA

均值未知
方差已知

$$L(\boldsymbol{\mu} \mid \mathcal{D}) = \frac{1}{(2\pi)^{nd/2} |\boldsymbol{\Sigma}|^{n/2}} \prod_{k=1}^{n} \exp\left[-\frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \right]$$

$$l(\boldsymbol{\mu} \mid \mathcal{D}) = \ln L(\boldsymbol{\mu} \mid \mathcal{D})$$

(对数似然函数)

$$= -\ln(2\pi)^{nd/2} |\boldsymbol{\Sigma}|^{n/2} - \frac{1}{2} \sum_{k=1}^{n} (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu})$$

$$\nabla_{\boldsymbol{\mu}} l(\boldsymbol{\mu} \mid \mathcal{D}) = \sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) = 0$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$

高斯分布均值的最大似然估计等于样本的均值

高斯分布的极大似然估计

MIMA

均值未知
方差未知

$$\boldsymbol{\theta} = (\theta_1, \theta_2)^T = (\mu, \sigma^2)^T$$

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

$$L(\boldsymbol{\theta} \mid \mathcal{D}) = p(\mathcal{D} \mid \boldsymbol{\theta})$$ (似然函数)

$$= \prod_{k=1}^{n} p(x_k \mid \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \prod_{k=1}^{n} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]$$

高斯分布的极大似然估计

均值未知
方差未知

$$L(\boldsymbol{\theta}\,|\,\mathcal{D}) = \frac{1}{(2\pi)^{n/2}\sigma^n}\prod_{k=1}^{n}\exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]$$

$$\boldsymbol{\theta} = (\theta_1, \theta_2)^T = (\mu, \sigma^2)^T$$

$$l(\boldsymbol{\theta}\,|\,\mathcal{D}) = \ln L(\boldsymbol{\theta}\,|\,\mathcal{D})$$

(对数似然函数)

$$= -\ln(2\pi)^{n/2}\theta_2^{\ n/2} - \frac{1}{2\theta_2}\sum_{k=1}^{n}(x_k-\theta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}\,|\,\mathcal{D}) = \begin{bmatrix} \dfrac{1}{\theta_2}\displaystyle\sum_{k=1}^{n}(x_k-\theta_1) \\ -\dfrac{n}{2\theta_2} + \displaystyle\sum_{k=1}^{n}\dfrac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix} = \mathbf{0}$$

$$\hat{\mu} = \hat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n}x_k$$

$$\hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})^2$$

The Gaussian Case I

MIMA

- Case I: unknown μ, and Σ is known

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \quad 1$$

$$L(\boldsymbol{\mu} \mid \mathcal{D}) = p(\mathcal{D} \mid \boldsymbol{\mu}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta}) \quad 2$$

(Likelihood function)

$$= \frac{1}{(2\pi)^{nd/2} |\boldsymbol{\Sigma}|^{n/2}} \prod_{k=1}^{n} \exp\left[-\frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \right] \quad 3$$

$$l(\boldsymbol{\mu} \mid \mathcal{D}) = \ln L(\boldsymbol{\mu} \mid \mathcal{D}) \quad 4$$

$$= -\ln(2\pi)^{nd/2} |\boldsymbol{\Sigma}|^{n/2} - \frac{1}{2} \sum_{k=1}^{n} (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \quad 5$$

The Gaussian Case I

MIMA

$$l(\boldsymbol{\mu} \mid \mathcal{D}) = \ln L(\boldsymbol{\mu} \mid \mathcal{D}) \quad (1)$$

$$= -\ln(2\pi)^{nd/2} \mid \boldsymbol{\Sigma} \mid^{n/2} - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \quad (2)$$

$$\nabla_{\boldsymbol{\mu}} l(\boldsymbol{\mu} \mid \mathcal{D}) = \sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) = 0 \quad (3)$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \quad (4)$$

Sample Mean!

Intuitive Result: Maximum estimate for the unknown μ is just the arithmetic average of training samples---sample mean.

The Gaussian Case II

MIMA

- Case II: both μ and ∑ are unknown
- Consider univariate case

$$\boldsymbol{\theta} = (\theta_1, \theta_2)^T = (\mu, \sigma^2)^T$$ 0

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

Likelihood function

$$L(\boldsymbol{\theta} \mid \mathcal{D}) = p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(x_k \mid \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2} \sigma^n} \prod_{k=1}^{n} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2} \right]$$ 2

$$l(\boldsymbol{\theta} \mid \mathcal{D}) = \ln L(\boldsymbol{\theta} \mid \mathcal{D}) = -\ln(2\pi)^{n/2} \sigma^n - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2$$ 3

$$= -\ln(2\pi)^{n/2} \theta_2^{\ n/2} - \frac{1}{2\theta_2} \sum_{k=1}^{n} (x_k - \theta_1)^2$$ 4

The Gaussian Case II

MIMA

$$l(\boldsymbol{\theta} \mid \mathcal{D}) = -\ln(2\pi)^{n/2}\theta_2^{\,n/2} - \frac{1}{2\theta_2}\sum_{k=1}^{n}(x_k - \theta_1)^2$$ 1

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta} \mid \mathcal{D}) = \begin{bmatrix} \dfrac{1}{\theta_2}\displaystyle\sum_{k=1}^{n}(x_k - \theta_1) \\ -\dfrac{n}{2\theta_2} + \displaystyle\sum_{k=1}^{n}\dfrac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix} = \mathbf{0}$$ 2

Unbiased Estimator:

$$E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

Consistent Estimator:

$$\lim_{n\to\infty} E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

$$\left\{\begin{aligned} \hat{\mu} &= \hat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n}x_k \\ \hat{\sigma}^2 &= \hat{\theta}_2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2 \end{aligned}\right.$$

3 Arithmetic average of n vectors

4 Arithmetic average of n matrices

$$(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

MLE for Normal Population

MIMA

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

Sample Mean

$$E[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

$$E[\hat{\boldsymbol{\Sigma}}] = \frac{n-1}{n}\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$$

$$\mathbf{C} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

Sample Covariance Matrix

$$E[\mathbf{C}] = \boldsymbol{\Sigma}$$

Contents

MIMA

Bayesian Estimation

- Settings
  - The parametric form of the likelihood function for each category is known.
  - However, θ_j is considered to be random variables instead of being fixed (but unknown) values.

In this case, we can no longer make a single ML estimate $\hat{\boldsymbol{\theta}}$ *and then infer* $P(\omega_i \mid \mathbf{x})$ *based on* $P(\omega_i)$ *and* $p(\mathbf{x} \mid \omega_i)$

How can we proceed? → Fully exploit training examples!

Posterior Probabilities from sample

MIMA

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_c\}$$ 0

$$P(\omega_i \mid \mathbf{x}) = P(\omega_i \mid \mathbf{x}, \mathcal{D})$$ 1

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \omega_i, \mathcal{D}) P(\omega_i \mid \mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, \mathcal{D}) P(\omega_j \mid \mathcal{D})} = P(\omega_i \mid \mathbf{x}, \mathcal{D})$$ 2

Assumptions: $P(\omega_i \mid \mathcal{D}) = P(\omega_i)$ $P(\mathbf{x} \mid \omega_i, \mathcal{D}) = P(\mathbf{x} \mid \omega_i, \mathcal{D}_i)$ 3

Each class can be considered independently

$$P(\omega_i \mid \mathbf{x}, \mathcal{D}) = \frac{P(\mathbf{x} \mid \omega_i, \mathcal{D}_i) P(\omega_i)}{\sum_{j=1}^{c} P(\mathbf{x} \mid \omega_j, \mathcal{D}_j) P(\omega_j)}$$ 4

Problem Formulation

MIMA

$$P(\omega_i \mid \mathbf{x}, \mathcal{D}) = \frac{P(\mathbf{x} \mid \omega_i, \mathcal{D}_i) P(\omega_i)}{\sum_{j=1}^{c} P(\mathbf{x} \mid \omega_j, \mathcal{D}_j) P(\omega_j)}$$

1

The key problem is to determine, $P(\mathbf{x} \mid \omega_i, \mathcal{D}_i)$ *,treat each class independently, the problem becomes* $P(\mathbf{x} \mid \mathcal{D})$

This is always the central problem of Bayesian Learning.

Class-Conditional Density Estimation

MIMA

Assume $p(\mathbf{x})$ is unknown but knowing it has a fixed form with parameter vector $\boldsymbol{\theta}$.

θ :Random variable w.r.t. parametric form

x is independent of D given θ

$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid D) d\boldsymbol{\theta} \quad 1$$

$$= \int p(\mathbf{x} \mid \boldsymbol{\theta}, D) p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta} \quad 2$$

$$= \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta} \quad 3$$

Class-Conditional Density Estimation

MIMA

Assume $p(\mathrm{x})$ is unknown but knowing it has a fixed form with parameter vector θ.

θ :Random variable w.r.t. parametric form

x is independent of D given θ

$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid D) d\boldsymbol{\theta} \quad 1$$

$$= \int p(\mathbf{x} \mid \boldsymbol{\theta}, D) p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta} \quad 2$$

$$= \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta} \quad 3$$

$$= p(\mathbf{x} \mid \omega_i, \mathcal{D}) \quad 4$$

$$p(\mathbf{x} \mid \mathcal{D}) \approx p(\mathbf{x} \mid \hat{\boldsymbol{\theta}}) \quad 5$$

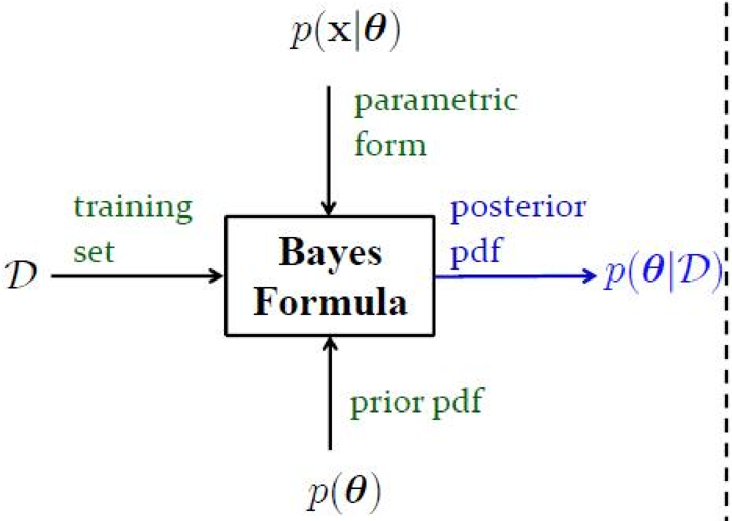
The form of distribution is assumed known

The posterior density we want to estimate

Bayesian Estimation: General Procedure

MIMA

Phase I: $p(\boldsymbol{\theta} \mid \mathcal{D}) = ?$



$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta}, \mathcal{D})}{p(\mathcal{D})}$$

$$= \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}, \mathcal{D})d\boldsymbol{\theta}}$$

$$= \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})d\boldsymbol{\theta}}$$

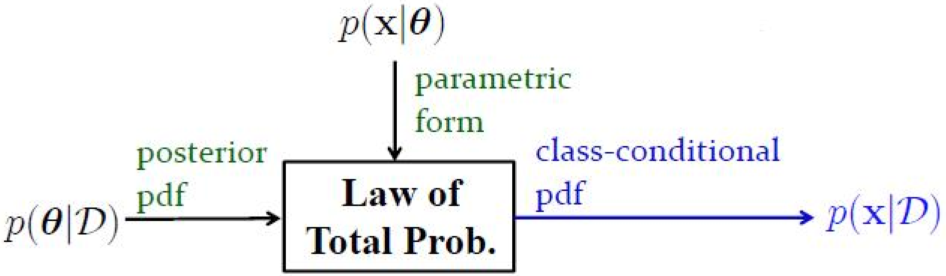
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$

Bayesian Estimation: General Procedure

MIMA

Phase II:

$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta} \qquad \boxed{1}$$



Phase III:

$$P(\omega_i \mid \mathbf{x}, \mathcal{D}) = \frac{P(\mathbf{x} \mid \omega_i, \mathcal{D}_i) P(\omega_i)}{\sum_{j=1}^{c} P(\mathbf{x} \mid \omega_j, \mathcal{D}_j) P(\omega_j)} \qquad \boxed{2}$$

The Gaussian Case

MIMA

- The univariate Gaussian: unknown μ

Phase I:

$$p(\mu),\quad p(x\,|\,\mu),\quad D \quad \Rightarrow \quad p(\mu\,|\,D)$$

1

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

2

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]$$

3

Other form of prior pdf could be assumed as well.

The Gaussian Case

MIMA

Phase I:

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right] \quad p(x \mid \mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \boxed{1}$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \alpha \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \qquad \boxed{2}$$

$$p(\mu \mid \mathcal{D}) = \alpha \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right] \cdot \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right] \qquad \boxed{3}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{x_k-\mu}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right] \qquad \boxed{4}$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right] \qquad \boxed{5}$$

The Gaussian Case

MIMA

Phase I:

$p(\mu \mid \mathcal{D})$ *is an exponential function of a quadratic function of* μ; *thus* $p(\mu \mid \mathcal{D})$ *is also a normal.*

$$p(\mu \mid \mathcal{D}) \sim N(\mu_n, \sigma_n^2)$$

$$p(\mu \mid \mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2} \left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right]$$

$$= \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2\sigma_n^2} \left(\mu^2 - 2\mu_n \mu + \mu_n^2 \right) \right]$$

$$p(\mu \mid \mathcal{D}) = \alpha'' \exp\left[-\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left(\frac{1}{\sigma^2} \sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2} \right) \mu \right] \right]$$

Comparison

The Gaussian Case

MIMA

- Equating the coefficients in both form, then, we have:

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \qquad \hat{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

The Gaussian Case

Phase II:
$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta}$$

$$\underline{\underline{p(\mu \mid D)}}, \quad \underline{\underline{p(x \mid \mu)}} \Rightarrow p(x \mid D)$$

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

$$p(\mu \mid \mathcal{D}) \sim N(\mu_n, \sigma_n^2)$$

How would $p(x \mid \mathcal{D})$ look like in this case?

The Gaussian Case

MIMA

$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x} \mid u) p(u \mid \mathcal{D}) d\boldsymbol{\theta}$$

$$\begin{cases} p(x \mid \mu) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right] \\ p(\mu \mid \mathcal{D}) \sim N(\mu_n, \sigma_n^2) \end{cases}$$

$$p(x \mid \mathcal{D}) = \frac{1}{2\pi\sigma\sigma_n} \int \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] \underbrace{\int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2 \mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] d\mu}_{=?}$$

$p(x \mid \mathcal{D})$ is an exponential function of a quadratic function of x; thus, it is also a normal pdf.

The Gaussian Case

MIMA

$$p(\mathbf{x}\mid\mathcal{D}) = \int p(\mathbf{x}\mid u)\,p(u\mid\mathcal{D})\,d\boldsymbol{\theta}$$

$$\begin{cases} p(x\mid\mu) = \dfrac{1}{\sqrt{2\pi}\sigma}\exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right] \\ p(\mu\mid\mathcal{D}) \sim N(\mu_n, \sigma_n^2) \end{cases}$$

$$\boxed{p(x\mid\mathcal{D}) \sim N(\mu_n, \sigma^2+\sigma_n^2)}$$

$$p(x\mid\mathcal{D}) = \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]\underbrace{\int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu-\frac{\sigma_n^2 x+\sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right]d\mu}_{=?}$$

$p(x\mid\mathcal{D})$ is an exponential function of a quadratic function of x; thus, it is also a normal pdf.

The Gaussian Case

MIMA

Phase III:

$$P(\omega_i \mid \mathbf{x}, \mathcal{D}) = \frac{P(\mathbf{x} \mid \omega_i, \mathcal{D}_i) P(\omega_i)}{\sum_{j=1}^{c} P(\mathbf{x} \mid \omega_j, \mathcal{D}_j) P(\omega_j)}$$

Summary

- Key issue
  - Estimate prior and class-conditional pdf from training set
  - Basic assumption on training examples: i.i.d.

- Two strategies to key issue
  - Parametric form for class-conditional pdf
    - Maximum likelihood estimation
    - Bayesian estimation
  - No parametric form for class-conditional pdf

Summary

MIMA

- Maximum likelihood estimation
  - Settings: parameters as fixed but unknown values
  - The objective function: log-likelihood function
  - The gradient for the objective function should be zero
  - Gaussian

- Bayesian estimation
  - Settings: parameters as random variables
  - General procedure: I, II, III
  - Gaussian case