

# Chapter 3

# Maximum-Likelihood and Bayesian Parameter Estimation



# Exercise

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

$$\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1} \quad \mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i \quad w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

Assumes equal prior probabilities,  
What is the decision boundary?



# Bayes Theorem for Classification

$$P(\omega_j | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_j) \cdot P(\omega_j)}{p(\mathbf{x})} \quad (1 \leq j \leq c) \quad (\text{Bayes Formula})$$

To compute posterior probability  $P(\omega_j | \mathbf{x})$ , we need to know:

Prior probability:  $P(\omega_j)$

Likelihood:  $p(\mathbf{x} | \omega_j)$

The collection of training examples is composed of  $c$  data sets

$\mathcal{D}_j \quad (1 \leq j \leq c)$

- Each example in  $\mathcal{D}_j$  is drawn according to the class-conditional pdf, i.e.  $p(\mathbf{x} | \omega_j)$
- Examples in  $\mathcal{D}_j$  are *i.i.d.* random variables, i.e. **independent and identically distributed** (独立同分布)



# Bayes Theorem for Classification (Cont.)

For prior probability:  no difficulty

$$P(\omega_j) = \frac{|\mathcal{D}_j|}{\sum_{i=1}^c |\mathcal{D}_i|}$$

(Here,  $|\cdot|$  returns the **cardinality**,  
i.e. number of elements, of a set)

For class-conditional pdf:

Ch. 3  **Case I:**  $p(\mathbf{x}|\omega_j)$  has certain **parametric form**

e.g.:  $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$  (**parameters:**  $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ )

$p(\mathbf{x}|\omega_j)$    $\boldsymbol{\theta}_j$  contains “ $d + d(d + 1)/2$ ” free parameters

To show the dependence of  
 $p(\mathbf{x}|\omega_j)$  on  $\boldsymbol{\theta}_j$  **explicitly**:

$p(\mathbf{x}|\omega_j)$    $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$

Ch. 4  **Case II:**  $p(\mathbf{x}|\omega_j)$  doesn't have **parametric form**



# Estimation Under Parametric Form

Parametric class-conditional pdf:  $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$  ( $1 \leq j \leq c$ )

## □ Assumption I: Maximum-Likelihood (ML) estimation (极大似然估计)

View parameters as quantities whose values are **fixed but unknown**



Estimate parameter values by **maximizing the likelihood** (probability) of observing the actual training examples

## □ Assumption II: Bayesian estimation (贝叶斯估计)

View parameters as **random variables** having some known prior distribution



Observation of the actual training examples transforms parameters' **prior distribution into posterior distribution** (via Bayes theorem)

# Maximum-Likelihood Estimation

## Settings

Likelihood function for each category is governed by some **fixed but unknown** parameters, i.e.  $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$  ( $1 \leq j \leq c$ )

**Task:** Estimate  $\{\boldsymbol{\theta}_j\}_{j=1}^c$  from  $\{\mathcal{D}_j\}_{j=1}^c$

### A simplified treatment

Examples in  $\mathcal{D}_j$  gives no information about  $\boldsymbol{\theta}_i$  if  $i \neq j$



Work with each category **separately** and therefore simplify the notations by dropping subscripts w.r.t. categories

**without loss of generality:**  $\mathcal{D}_j \rightarrow \mathcal{D}$  ;  $\boldsymbol{\theta}_j \rightarrow \boldsymbol{\theta}$

# Maximum-Likelihood Estimation (Cont.)

$$\mathbf{x}_k \sim p(\mathbf{x}|\theta)$$

$$(k = 1, \dots, n)$$

$\theta$  : Parameters to be estimated

$\mathcal{D}$  : A set of i.i.d. examples  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$

The objective function

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

The likelihood of  $\theta$  w.r.t. the set of observed examples

The maximum-likelihood estimation

$$\hat{\theta} = \arg \max_{\theta} p(\mathcal{D}|\theta)$$

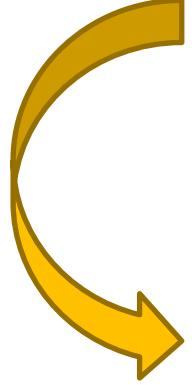
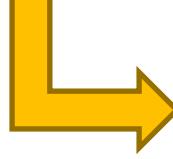
Intuitively,  $\hat{\theta}$  best agrees with the actually observed examples



# Maximum-Likelihood Estimation (Cont.)

## Gradient Operator (梯度算子)

- ✓ Let  $\theta = (\theta_1, \dots, \theta_p)^t \in \mathbf{R}^p$  be a  $p$ -dimensional vector
- ✓ Let  $f : \mathbf{R}^p \rightarrow \mathbf{R}$  be  $p$ -variate real-valued function over


$$\nabla_{\theta} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix} \quad f(\theta) = \theta_1^2 + 3\theta_1\theta_2$$

$$\nabla_{\theta} f = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} 2\theta_1 + 3\theta_2 \\ 3\theta_1 \end{bmatrix}$$

$l(\theta) = \ln p(\mathcal{D}|\theta)$  is named as the **log-likelihood function**

$$\hat{\theta} = \arg \max_{\theta} p(\mathcal{D}|\theta) \quad \longleftrightarrow \quad \hat{\theta} = \arg \max_{\theta} l(\theta)$$

# Maximum-Likelihood Estimation (Cont.)

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

$$\nabla_{\theta} l = \nabla_{\theta} \left( \sum_{k=1}^n \ln p(\mathbf{x}_k | \boldsymbol{\theta}) \right) = \sum_{k=1}^n \nabla_{\theta} \ln p(\mathbf{x}_k | \boldsymbol{\theta})$$

$p$ -dimensional vector with each component being a function over  $\theta$

$p$ -variate real-valued function over  $\theta$  (not over  $x_k$ )

# Necessary conditions for ML estimate $\hat{\theta}$

$$\nabla_{\theta} l \Big|_{\theta=\hat{\theta}} = 0 \text{ (a set of } p \text{ equations)}$$

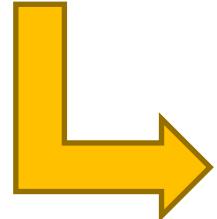


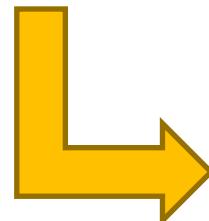
# The Gaussian Case: Unknown $\mu$

$\mathbf{x}_k \sim N(\boldsymbol{\mu}, \Sigma)$   
 $(k = 1, \dots, n)$

suppose  $\Sigma$  is known   $\theta = \{\boldsymbol{\mu}\}$

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


$$\begin{aligned} \ln p(\mathbf{x}_k | \boldsymbol{\mu}) &= -\frac{1}{2} \ln [(2\pi)^d |\Sigma|] - \frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \\ &= -\frac{1}{2} \ln [(2\pi)^d |\Sigma|] - \frac{1}{2} \mathbf{x}_k^t \Sigma^{-1} \mathbf{x}_k + \boldsymbol{\mu}^t \Sigma^{-1} \mathbf{x}_k - \frac{1}{2} \boldsymbol{\mu}^t \Sigma^{-1} \boldsymbol{\mu} \end{aligned}$$


$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k | \boldsymbol{\mu}) = \Sigma^{-1} (\mathbf{x}_k - \boldsymbol{\mu})$$

# The Gaussian Case: Unknown $\Sigma$ (Cont.)

$$l(\boldsymbol{\mu}) = \sum_{k=1}^n \ln p(\mathbf{x}_k | \boldsymbol{\mu})$$

Intuitive result

ML estimate for the unknown  
is just the arithmetic average of  
training samples – **sample mean**

↓  $\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k | \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$

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

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

↓  $\nabla_{\boldsymbol{\mu}} l = \mathbf{0}$  (necessary condition  
for ML estimate  $\hat{\boldsymbol{\mu}}$ )

Multiply  $\boldsymbol{\Sigma}$  on  
both sides

$$\sum_{k=1}^n \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0} \quad \longleftrightarrow \quad \sum_{k=1}^n (\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0}$$



# The Gaussian Case: Unknown $\mu$ and $\Sigma$

$$\mathbf{x}_k \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ (k = 1, \dots, n)$$

$\mu$  and  $\Sigma$  unknown  $\Rightarrow \boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$

Consider *univariate case*

$$p(x_k|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right] \quad \left(\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}\right)$$

→  $\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$

→  $\nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$



# The Gaussian Case: Unknown $\boldsymbol{\theta}$ and $\Sigma$ (Cont.)

$$l(\boldsymbol{\theta}) = \sum_{k=1}^n \ln p(x_k | \boldsymbol{\theta})$$

$$\sum_{k=1}^n \frac{1}{\hat{\theta}_2} (x_k - \hat{\theta}_1) = 0$$

$$\nabla_{\boldsymbol{\theta}} \ln p(x_k | \boldsymbol{\theta}) =$$

$$\begin{bmatrix} \frac{1}{\theta_2} (x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$-\sum_{k=1}^n \frac{1}{\hat{\theta}_2} + \sum_{k=1}^n \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0$$

$$\nabla_{\boldsymbol{\theta}} l = \left[ \begin{array}{c} \sum_{k=1}^n \frac{1}{\theta_2} (x_k - \theta_1) \\ \sum_{k=1}^n \left( -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \right) \end{array} \right]$$

$\nabla_{\boldsymbol{\theta}} l = 0$  (necessary condition  
for ML estimate  $\hat{\theta}_1$  and  $\hat{\theta}_2$ )



# The Gaussian Case: Unknown $\mu$ and $\Sigma$ (Cont.)

$$\sum_{k=1}^n \frac{1}{\hat{\theta}_2} (x_k - \hat{\theta}_1) = 0 \rightarrow \sum_{k=1}^n (x_k - \hat{\theta}_1) = 0 \rightarrow \hat{\theta}_1 = \frac{1}{n} \sum_{k=1}^n x_k$$

$$-\sum_{k=1}^n \frac{1}{\hat{\theta}_2} + \sum_{k=1}^n \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \rightarrow \hat{\theta}_2 = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{\theta}_1)^2$$

## ML estimate in *univariate* case

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

# The Gaussian Case: Unknown $\Sigma$ and (Cont.)

ML estimate in *multivariate case*

Intuitive  
result as well!

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



*Arithmetic average of  
n vectors  $\mathbf{x}_k$*

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



*Arithmetic average  
of n matrices  
 $(\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t$*

# Biased/Unbiased Estimator

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

↓

$$\mathcal{E}[\hat{\Sigma}] = \mathcal{E} \left[ \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t \right] = \frac{n-1}{n} \Sigma$$

**Biased estimator (有偏估计) of  $\Sigma$**

→ C =  $\frac{n}{n-1} \hat{\Sigma}$     **Unbiased estimator (无偏估计) of  $\Sigma$**

$$\lim_{n \rightarrow \infty} \mathcal{E}[\hat{\Sigma}] = \Sigma$$

**Asymptotically unbiased estimator  
(渐进无偏估计) of  $\Sigma$**



# Bayesian Estimation

## Settings

- The **parametric form** of the likelihood function for each category is known  $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$  ( $1 \leq j \leq c$ )
- However,  $\boldsymbol{\theta}_j$  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  $\hat{\boldsymbol{\theta}}_j$  and then infer  $P(\omega_j|\mathbf{x})$  based on  $P(\omega_j)$  and  $p(\mathbf{x}|\omega_j, \hat{\boldsymbol{\theta}}_j)$



How can we  
proceed under  
this situation

Fully exploit training examples!

$$P(\omega_j|\mathbf{x}) \longrightarrow P(\omega_j|\mathbf{x}, \mathcal{D}^*)$$

$$(\mathcal{D}^* = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_c)$$

# Bayesian Estimation (Cont.)

$$P(\omega_j | \mathbf{x}, \mathcal{D}^*) = \frac{p(\omega_j, \mathbf{x}, \mathcal{D}^*)}{p(\mathbf{x}, \mathcal{D}^*)} = \frac{p(\omega_j, \mathbf{x}, \mathcal{D}^*)}{\sum_{i=1}^c p(\omega_i, \mathbf{x}, \mathcal{D}^*)}$$

  $p(\omega_j, \mathbf{x}, \mathcal{D}^*) = p(\mathcal{D}^*) \cdot p(\omega_j, \mathbf{x} | \mathcal{D}^*) = p(\mathcal{D}^*) \cdot P(\omega_j | \mathcal{D}^*) \cdot p(\mathbf{x} | \omega_j, \mathcal{D}^*)$

$$P(\omega_j | \mathbf{x}, \mathcal{D}^*) = \frac{p(\mathcal{D}^*) \cdot P(\omega_j | \mathcal{D}^*) \cdot p(\mathbf{x} | \omega_j, \mathcal{D}^*)}{p(\mathcal{D}^*) \cdot \sum_{i=1}^c P(\omega_i | \mathcal{D}^*) \cdot p(\mathbf{x} | \omega_i, \mathcal{D}^*)}$$

**Two assumptions**

$$P(\omega_j | \mathcal{D}^*) = P(\omega_j)$$

$$p(\mathbf{x} | \omega_j, \mathcal{D}^*) = p(\mathbf{x} | \omega_j, \mathcal{D}_j)$$

$$= \frac{P(\omega_j | \mathcal{D}^*) \cdot p(\mathbf{x} | \omega_j, \mathcal{D}^*)}{\sum_{i=1}^c P(\omega_i | \mathcal{D}^*) \cdot p(\mathbf{x} | \omega_i, \mathcal{D}^*)} \quad \text{Eq.22 [pp.91]}$$

$$= \frac{P(\omega_j) \cdot p(\mathbf{x} | \omega_j, \mathcal{D}_j)}{\sum_{i=1}^c P(\omega_i) \cdot p(\mathbf{x} | \omega_i, \mathcal{D}_i)} \quad \text{Eq.23 [pp.91]}$$



# Bayesian Estimation (Cont.)

$$P(\omega_j | \mathbf{x}, \mathcal{D}^*) = \frac{P(\omega_j) \cdot p(\mathbf{x} | \omega_j, \mathcal{D}_j)}{\sum_{i=1}^c P(\omega_i) \cdot p(\mathbf{x} | \omega_i, \mathcal{D}_i)}$$

Key problem

Determine  $p(\mathbf{x} | \omega_j, \mathcal{D}_j)$

Treat each class  
independently

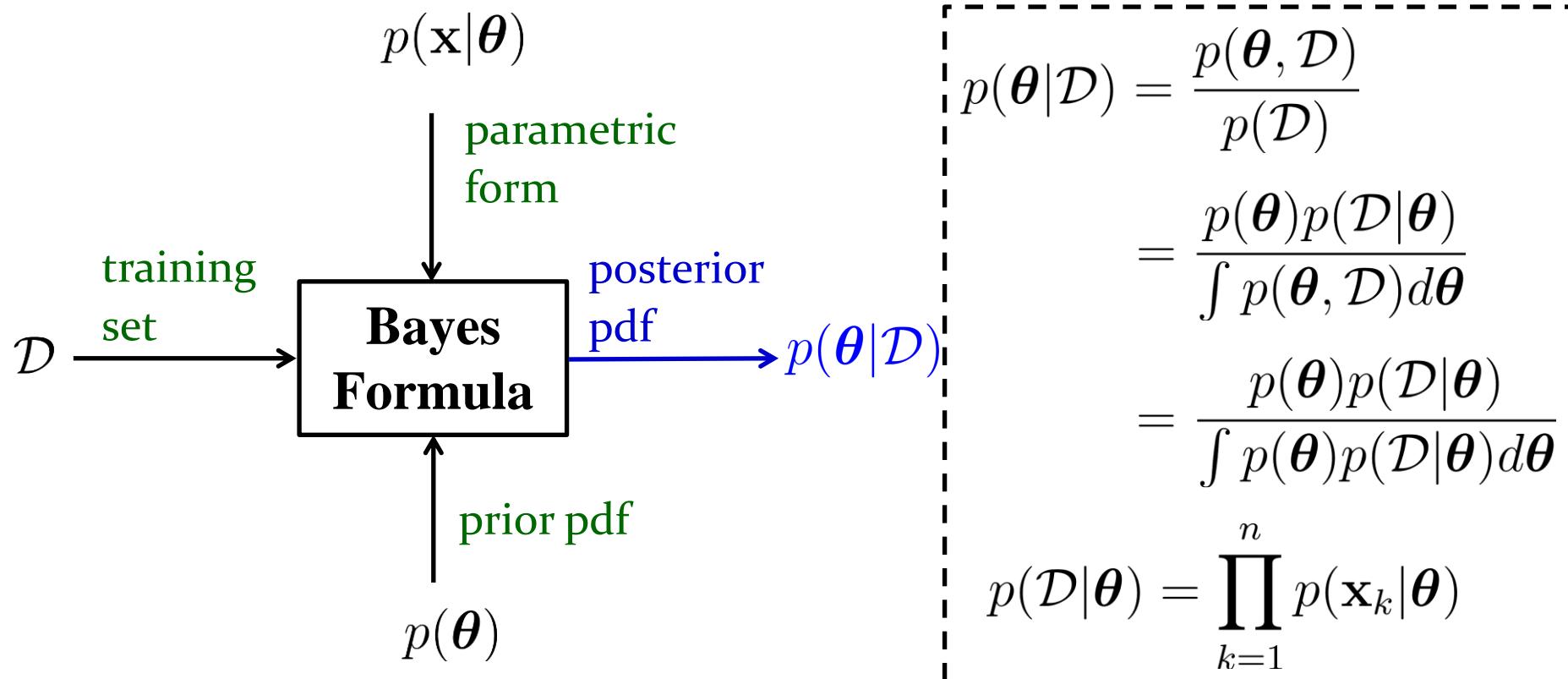


Simplify the *class-conditional pdf*  
notation  $p(\mathbf{x} | \omega_j, \mathcal{D}_j)$  as  $p(\mathbf{x} | \mathcal{D})$

$$\begin{aligned} p(\mathbf{x} | \mathcal{D}) &= \int p(\mathbf{x}, \boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta} \quad (\boldsymbol{\theta} : \text{random variables w.r.t. parametric form}) \\ &= \int p(\mathbf{x} | \boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta} \\ &= \int p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta} \quad (\mathbf{x} \text{ is independent of } \mathcal{D} \text{ given } \boldsymbol{\theta}) \end{aligned}$$

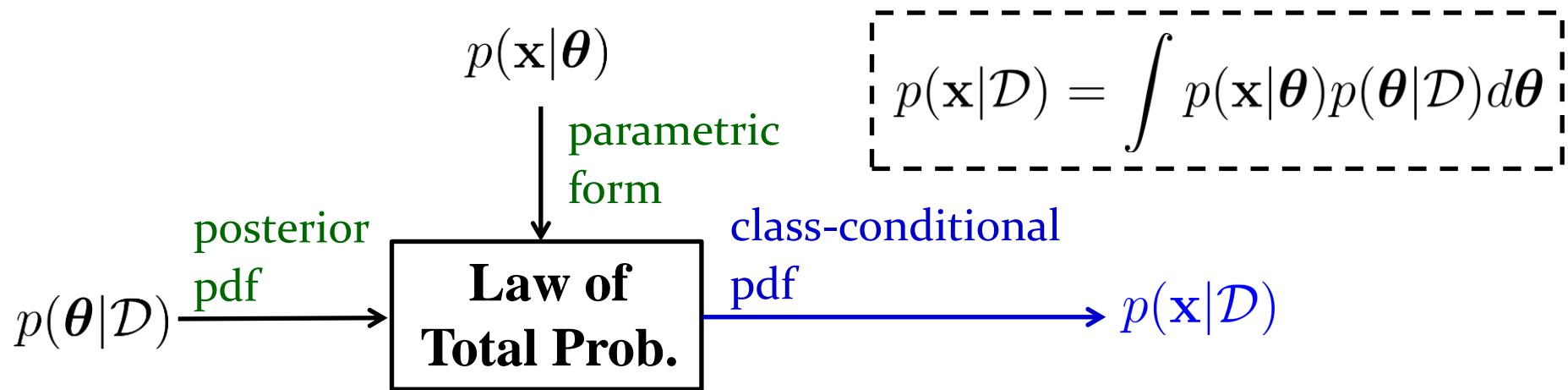
# Bayesian Estimation: The General Procedure

**Phase I:** prior pdf  $\rightarrow$  posterior pdf (for  $\theta$ )



# Bayesian Estimation: The General Procedure

**Phase II:** posterior pdf (for  $\theta$ )  $\rightarrow$  class-conditional pdf (for  $x$ )



**Phase III:**  $P(\omega_j|x, \mathcal{D}^*) = \frac{P(\omega_j) \cdot p(x|\omega_j, \mathcal{D}_j)}{\sum_{i=1}^c P(\omega_i) \cdot p(x|\omega_i, \mathcal{D}_i)}$

# The Gaussian Case: Unknown $\mu$

**Consider univariate case:**  $\theta = \{\mu\}$  ( $\sigma^2$  is known)

**Phase I:** prior pdf  $\rightarrow$  posterior pdf (for  $\theta$ )

$$\frac{p(\mu)}{\text{T}} + \frac{p(x|\mu)}{\text{T}} + \mathcal{D} \xrightarrow{\text{Large yellow arrow}} p(\mu|\mathcal{D})$$

$\xrightarrow{\quad} p(x|\mu) \sim N(\mu, \sigma^2)$  Gaussian parametric form

$\xrightarrow{\quad} p(\mu) \sim N(\mu_0, \sigma_0^2)$

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



# The Gaussian Case: Unknown $\mu$ (Cont.)

$$\begin{aligned} p(\mu|\mathcal{D}) &= \frac{p(\mu, \mathcal{D})}{p(\mathcal{D})} = \frac{p(\mu)p(\mathcal{D}|\mu)}{\int p(\mu)p(\mathcal{D}|\mu) d\mu} \\ &= \alpha p(\mu) p(\mathcal{D}|\mu) \quad \left( \int p(\mu)p(\mathcal{D}|\mu) d\mu \text{ is a constant not related to } \mu \right) \\ &= \alpha p(\mu) \prod_{k=1}^n p(x_k|\mu) \quad (\text{examples in } \mathcal{D} \text{ are } i.i.d.) \end{aligned}$$

$$p(\mu) \sim N(\mu_0, \sigma_0^2)$$

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

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

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



# The Gaussian Case: Unknown $\mu$

(Cont.)

$$p(\mu|\mathcal{D}) = \alpha p(\mu) \prod_{k=1}^n p(x_k|\mu)$$

$p(\mu|\mathcal{D})$  is an exponential function of a quadratic function of  $\mu$

$\Rightarrow$   $p(\mu|\mathcal{D})$  is a normal pdf as well

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

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

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

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



# The Gaussian Case: Unknown $\mu$ (Cont.)

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

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

Equating the  
coefficients in  
both form:

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2} \sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2}$$



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

$$\mu_n = \frac{\sigma_n^2}{\sigma^2} \sum_{k=1}^n x_k + \frac{\sigma_n^2}{\sigma_0^2} \mu_0$$

# The Gaussian Case: Unknown $\mu$ (Cont.)

**Phase II:** posterior pdf (for  $\theta$ )  $\rightarrow$  class-conditional pdf (for  $x$ )

$p(\mu|\mathcal{D}) + p(x|\mu) \xrightarrow{\text{Large Yellow Arrow}} p(x|\mathcal{D})$   
 $\xrightarrow{\text{Left Arrow}} p(x|\mu) \sim N(\mu, \sigma^2)$   
 $\xrightarrow{\text{Right Arrow}} p(\mu|\mathcal{D}) \sim N(\mu_n, \sigma_n^2)$

How would  $p(x|D)$  look like in this case?

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

$$\mu_n = \frac{\sigma_n^2}{\sigma^2} \sum_{k=1}^n x_k + \frac{\sigma_n^2}{\sigma_0^2} \mu_0$$



# The Gaussian Case: Unknown $\mu$ (Cont.)

Then, phase III  
follows naturally  
for prediction

$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})d\mu \quad \text{Eq.25 [pp.92]}$$

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

$$= \beta \cdot \exp \left[ -\frac{1}{2} \frac{(x-\mu_n)^2}{\sigma^2 + \sigma_n^2} \right] \quad \text{Eq.36 [pp.95]}$$

$p(x|\mathcal{D})$  is an exponential  
function of a quadratic  
function of  $x$



$p(x|\mathcal{D})$  is a  
normal pdf  
as well

$$p(x|\mathcal{D}) \sim$$

$$N(\mu_n, \sigma^2 + \sigma_n^2)$$



# The Gaussian Case: Unknown $\mu$ (Multivariate)

$\theta = \{\mu\}$  ( $\Sigma$  is known)



$$p(\mathbf{x}|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

$$p(\boldsymbol{\mu}|\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \quad p(\mathbf{x}|\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0 \left( \boldsymbol{\Sigma}_0 + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k + \frac{1}{n} \boldsymbol{\Sigma} \left( \boldsymbol{\Sigma}_0 + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_0$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0 \left( \boldsymbol{\Sigma}_0 + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \frac{1}{n} \boldsymbol{\Sigma}$$

# A Few Notes on Parametric Techniques

## ML estimation vs. Bayes estimation

- *Infinite examples*

ML estimation

=

Bayes estimation

- *Complexity*

ML estimation

<

Bayes estimation

- *Interpretability*

ML estimation

>

Bayes estimation

- *Prior knowledge*

ML estimation

<

Bayes estimation

## Source of classification error

Bayes error

+

Model error

+

Estimation error



# Summary

- Key issue for PR
  - Estimate prior and class-conditional pdf from training set
  - Basic assumption on training examples: *i.i.d.*
- Two strategies to the key issue
  - Parametric form for class-conditional pdf
    - Maximum likelihood (ML) estimation
    - Bayesian estimation
  - No parametric form for class-conditional pdf



# Summary (Cont.)

- Maximum likelihood estimation
  - Settings: **parameters as fixed but unknown values**
  - The objective function: **Log-likelihood function**
  - Necessary conditions for ML estimation: **gradient for the objective function should be zero vector**
  - The Gaussian case
    - Unknown  $\mu$
    - Unknown  $\mu$  and  $\Sigma$



# Summary (Cont.)

- Bayesian estimation
  - Settings: **parameters as random variables**
  - The general procedure
    - Phase I: *prior pdf*  $\rightarrow$  *posterior pdf* (for  $\theta$ )
    - Phase II: *posterior pdf* (for  $\mu$ )  $\rightarrow$  *class-conditional pdf* (for  $\mathbf{x}$ )
    - Phase III: *prediction* (Eq.22 [pp.91])
  - The Gaussian case
    - Unknown  $\mu$ : univariate and multivariate
  - A recursive view of Bayesian estimation
    - $p(\theta | \mathcal{D}^n) \propto p(\mathbf{x}_n | \theta)p(\theta | \mathcal{D}^{n-1})$

