# Parameter Estimation

# Introduction

– Data availability in a Bayesian framework

- We could design an optimal classifier if we knew:
  – $P(\omega_i)$ (priors)
  – $p(\mathbf{x} \mid \omega_i)$ (class-conditional densities)

  Unfortunately, we rarely have this complete information!

– Design a classifier from a training sample

- No problem with prior estimation
- Samples are often too small for class-conditional estimation (large dimension of feature space!)

A priori information about the problem

Normality of $p(\mathbf{x} \mid \omega_i)$

$$p(\mathbf{x} \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

Characterized by 2 parameters

# Estimation techniques

Maximum-Likelihood (ML) and the Bayesian estimations

Results are nearly identical, but the approaches are different

# Maximum likelihood

- Parameters in ML estimation are fixed but unknown!
- Best parameters are obtained by maximizing the probability of obtaining the samples observed
- Parameters are chosen in a way that they best support/ describe the training data.

# Bayesian learning

- Bayesian methods view the parameters as random variables having some known distribution

- Observation of the samples converts this to a posterior density, thereby revising our opinion about the true values of the parameters.

- In the Bayesian case, we shall see that a typical effect of observing additional samples is to sharpen the a posteriori density function, causing it to peak near the true values of the parameters. This phenomenon is known as *Bayesian learning.*

- In either approach, we use $P(\omega_i \mid \mathbf{x})$ for our classification rule!

- Bayes Theorem is the key…!

# Maximum-Likelihood Estimation

- Maximum-Likelihood Estimation

    - Has good convergence properties as the sample size increases
    - Simpler than any other alternative techniques

  – General principle

    - Assume we have c classes and

      $p(\mathbf{x} \mid \omega_j) \sim N(\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$

      $p(\mathbf{x} \mid \omega_j) \equiv P(\mathbf{x} \mid \omega_j, \theta_j)$ where:

# Maximum likelihood

- Use the information provided by the training samples to estimate $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$,

- Each $\theta_i$ $(i = 1, 2, \ldots, c)$ is associated with each category

- Suppose that *D* contains n samples, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$

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

- ML estimate of $\theta$ is, by definition the value that maximizes $p(D \mid \theta)$
  "It is the value of $\theta$ that best agrees with the actually observed training sample"

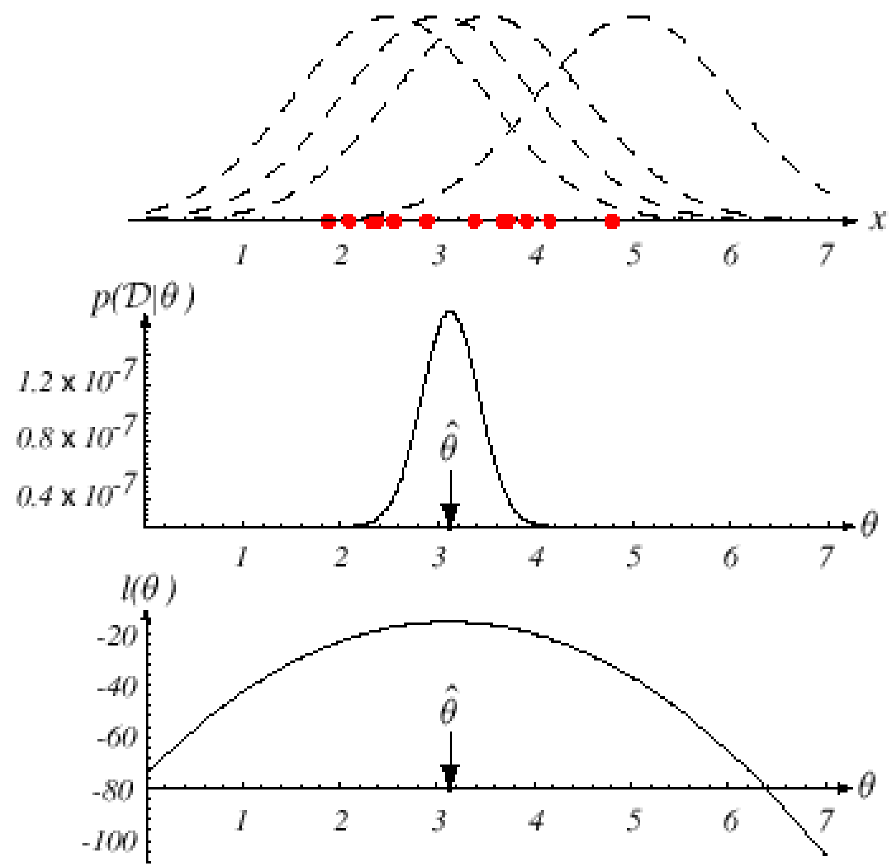



**FIGURE 3.1.** The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of $\theta$ whereas the conditional density $p(x|\theta)$ is shown as a function of $x$. Furthermore, as a function of $\theta$, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. 

# Maximum likelihood estimation

– Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let $\nabla_\theta$ be the gradient operator

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

– We define $l(\theta)$ as the log-likelihood function

$$l(\boldsymbol{\theta}) \equiv \ln p(\mathcal{D}|\boldsymbol{\theta})$$

– New problem statement: determine $\theta$ that maximizes the log-likelihood

# Maximum likelihood estimation

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}),$$

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

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

**Set of necessary conditions :**

$$\boxed{\nabla_{\boldsymbol{\theta}} l = 0.}$$

# Gaussian Case : unknown μ, **known Σ**

p(**x** | μ) ~ N(μ, Σ)

(Samples are drawn from a multivariate normal population)

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

and

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

θ = μ therefore:

- The ML estimate for μ must satisfy:

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

- Multiplying by Σ and rearranging, we obtain:

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

Just the arithmetic average of the samples of the training samples!

If $p(\mathbf{x} \mid \omega_j)$ $(j = 1, 2, \ldots, c)$ is supposed to be Gaussian in a *d*-dimensional feature space; then we can estimate the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_c)^t$ and perform an optimal classification!

# Gaussian Case: *unknown $\mu$ and $\sigma$*

– Gaussian Case: *unknown $\mu$ and $\sigma$*

$$\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$$

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

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

Solving :

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

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

Combining above equations, one obtains:

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

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

# The Gaussian Case: Unknown μ and Σ

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

and

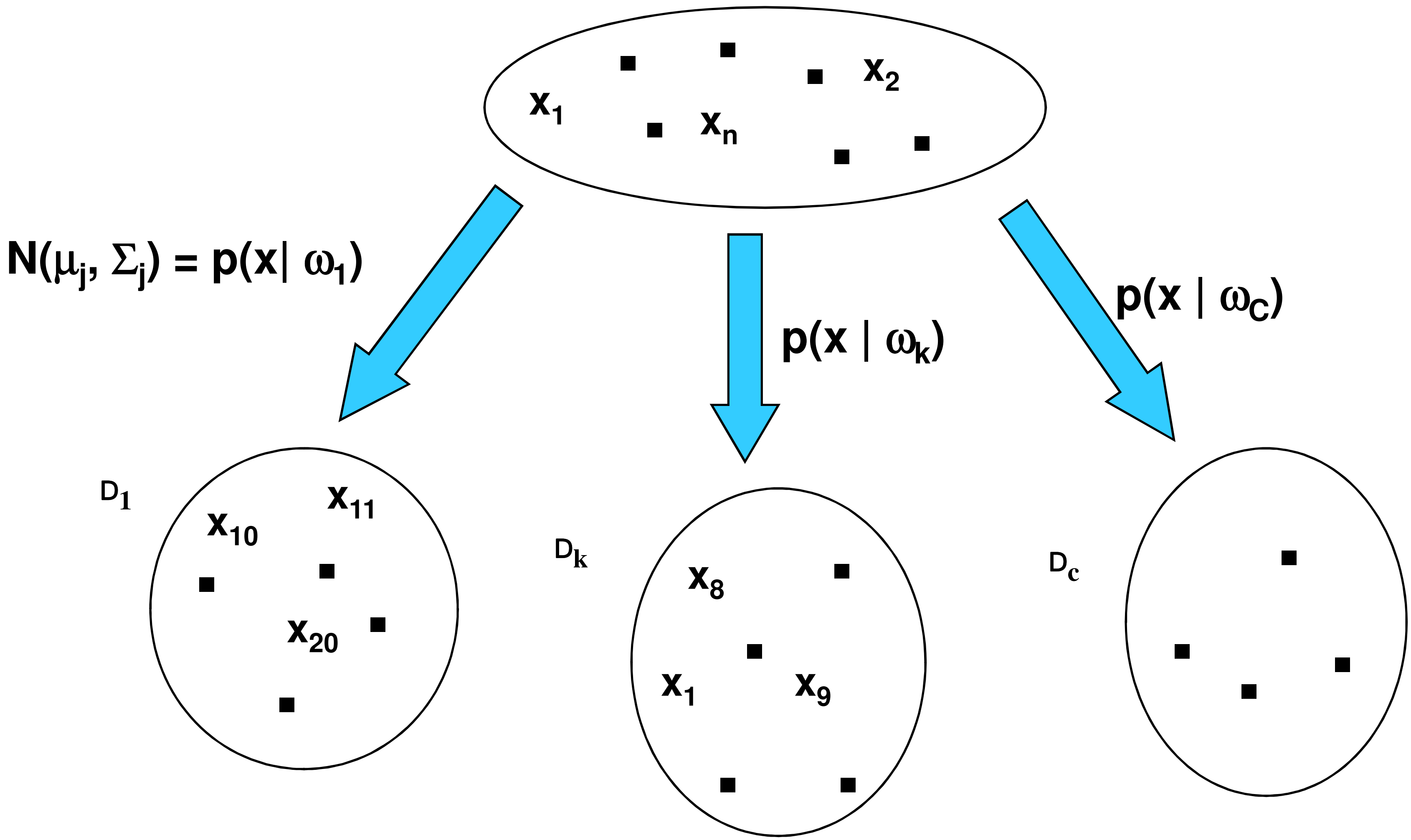
$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t.$$

# ML Problem Statement

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

$p(x_1, \ldots, x_n \mid \theta) = \prod p(\mathbf{x}_k \mid \theta)$; $|D| = n$

Our goal is to determine (value of $\theta$ that makes this sample the most representative!)



$x_1$
$x_n$
$x_2$
$N(\mu_j, \Sigma_j) = p(x| \omega_1)$
$p(x | \omega_k)$
$p(x | \omega_C)$
$D_1$
$x_{10}$
$x_{11}$
$x_{20}$
$D_k$
$x_8$
$x_1$
$x_9$
$D_c$