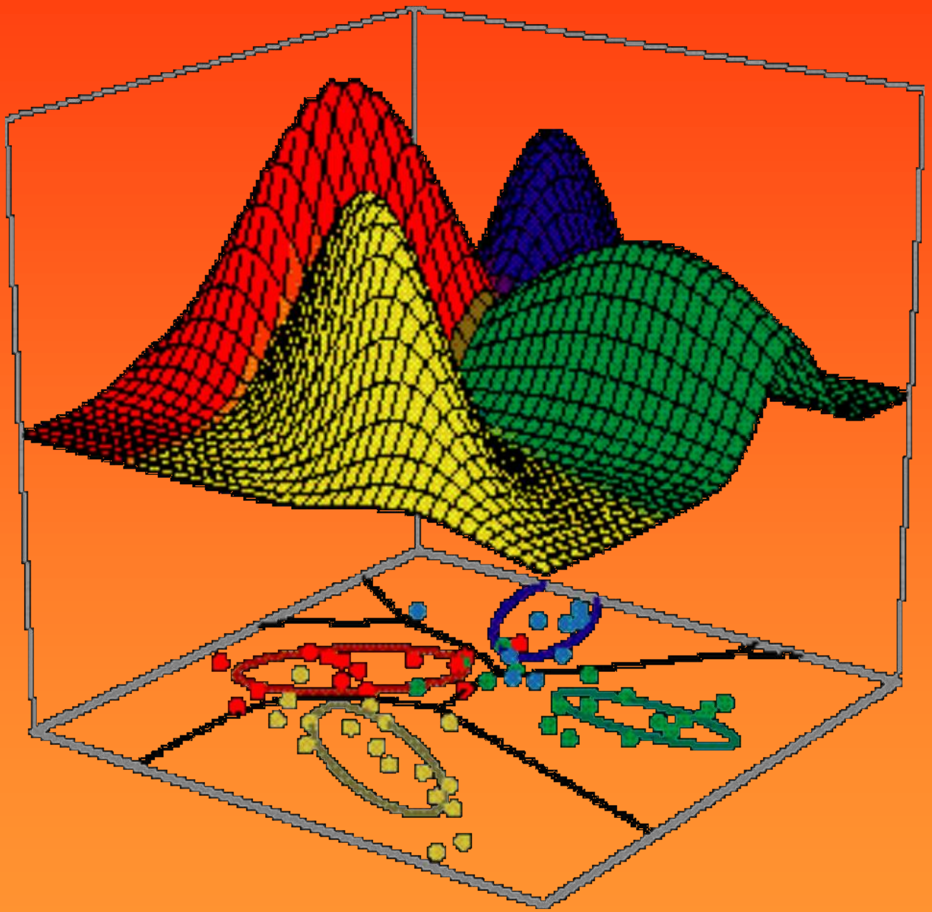


Pattern Classification

All materials in these slides were taken from Pattern Classification (2nd ed) **by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000**
with the permission of the authors and the publisher

Chapter 3: Maximum-Likelihood & Bayesian Parameter Estimation (3.1,3.2)

- Introduction
- Maximum-Likelihood Estimation
  - The Gaussian Case 1:unknown μ
  - The Gaussian Case 2: unknown μ and σ
  - Bias
-

3.1 Introduction

- Data availability in a Bayesian framework
  - We could design an optimal classifier if we knew:
    - $P(\omega_i)$ (priors)
    - $P(x \mid \omega_i)$ (class-conditional densities)
    - Unfortunately, we rarely have both complete information!

- Design a classifier from a training sample
  - No problem with the estimation of prior probabilities
  - Samples are often too few for the estimation of class-conditional densities
  - Complexity for large dimension of feature space

- To simplify above problem
  - Normality of $P(x \mid \omega_i)$
  - $P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$: Characterized by 2 parameters
  - The problem is changed from estimating $P(x \mid \omega_i)$ to estimating μ_i, Σ_i

- Estimation techniques
  - Maximum-Likelihood (ML) and the Bayesian estimations
  - Results are nearly identical, but the approaches are conceptually different

- Parameters in ML estimation are fixed but unknown!
- Best parameters are obtained by maximizing the probability of obtaining the samples observed
- Bayesian methods view the parameters as random variables having some known prior distribution. Training data allow us to convert a distribution on this variable into a posterior probability density

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

3.2 Maximum-Likelihood Estimation

- M-L Estimation
  - Has good convergence properties as the sample size increase
  - Simpler than any other alternative techniques
- General principle
  - Assume we have c classes and

$p(x \mid \omega_j) \sim N(\mu_j, \Sigma_j)$

$p(x \mid \omega_j) \equiv p(x \mid \omega_j, \theta_j)$ where:

$$\theta_j = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, \ldots, \sigma_j^{11}, \sigma_j^{22} \ldots)$$

- Use the information provided by the training samples $D = (D_1, D_2, \ldots, D_c)$ to estimate

  $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, each θ_i ($i = 1, 2, \ldots, c$) is associated with each category

  assume D_i give no information about θ_j if i<>j

  So Handle each class separately to simplify our notation

  Suppose that D contains n samples, $x_1, x_2, \ldots, x_n$

$$p(D \mid \theta) = \prod_{k=1}^{k=n} p(x_k \mid w, \theta) = \prod_{k=1}^{k=n} p(x_k \mid \theta) = F(\theta)$$

$p(D \mid \theta)$ is called the likelihood of θ

- ML estimation of θ is, by definition, the value $\hat{\theta}$ that maximizes $p(D \mid \theta)$

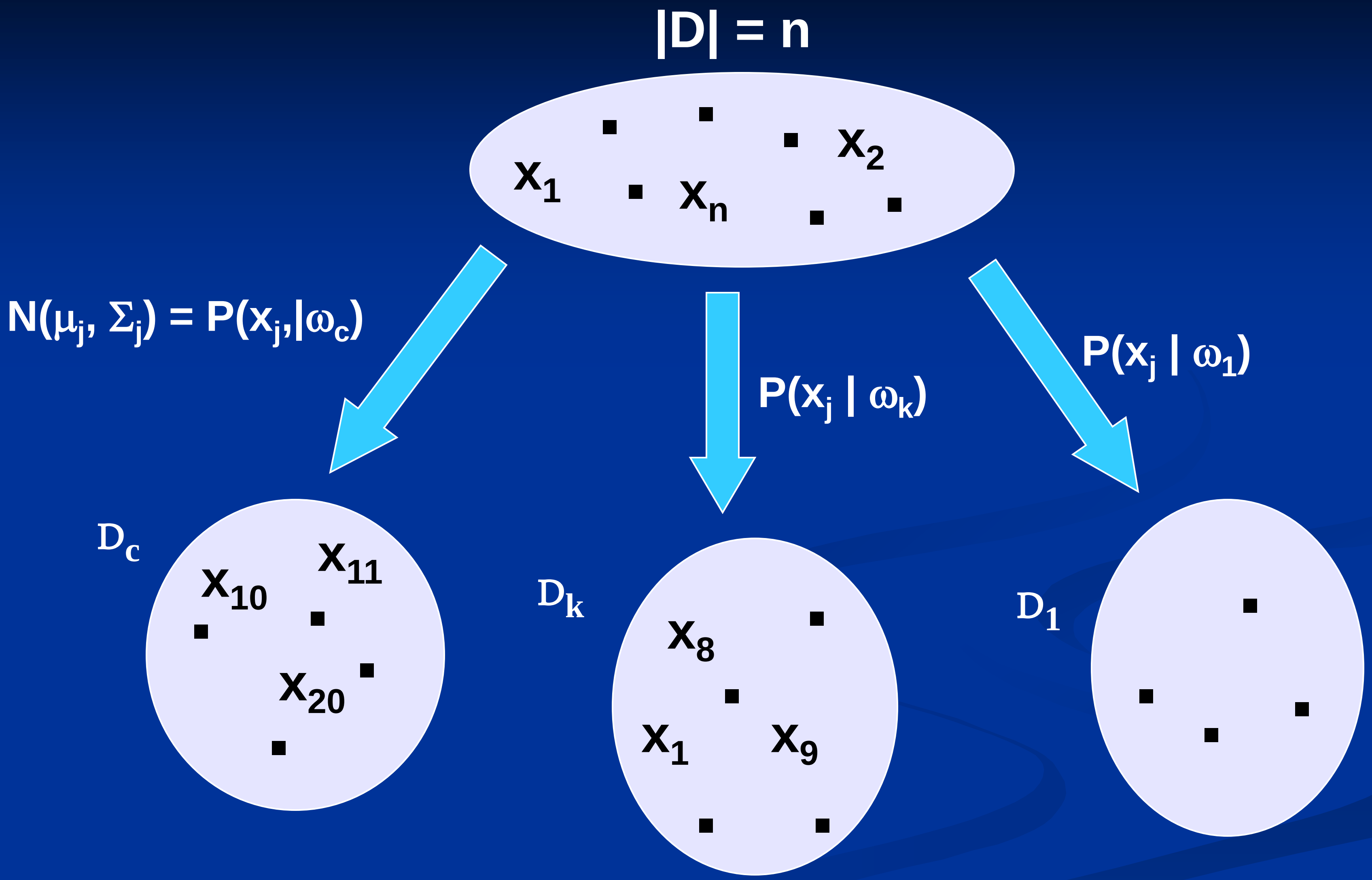
  "It is the value of θ that best agrees with the actually observed training sample"

ML Problem Statement

- Let $D = \{x_1, x_2, \ldots, x_n\}$

  $p(x_1,\ldots, x_n \mid \theta) = \Pi^{1,\ldots,n} P(x_k \mid \theta)$; $|D| = n$

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



|D| = n
x1
x2
xn
N(μj, Σj) = P(xj,|ωc)
P(xj | ωk)
P(xj | ω1)
Dc
x10
x11
x20
Dk
x8
x1
x9
D1

$\theta = (\theta_1, \theta_2, \ldots, \theta_c)$

Problem: find $\hat{\theta}$ such that:

$$\max_{\theta} P(D \mid \theta) = \max P(x_1, \ldots, x_n \mid \theta)$$

$$= \max \prod_{k=1}^{n} P(x_k \mid \theta)$$

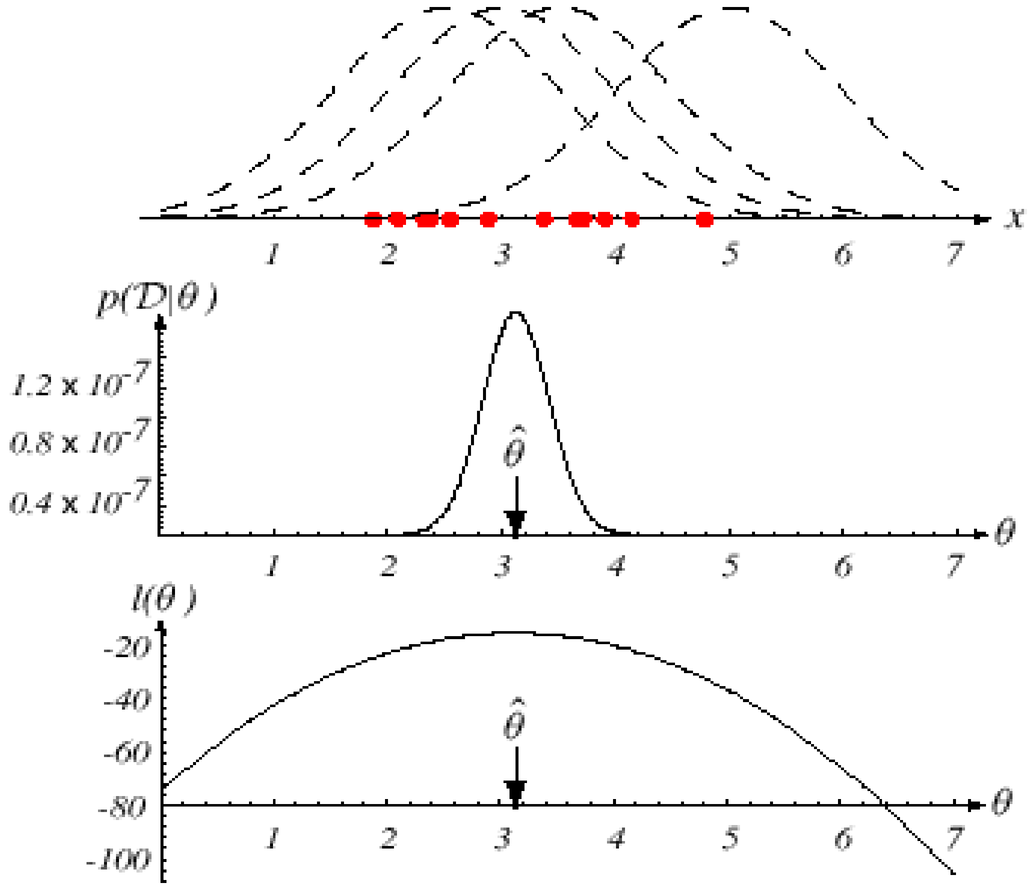


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ 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 θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, 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*.

- Optimal estimation
  - Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let ∇_θ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p} \right]^t$$

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

$$l(\theta) = \ln p(D \mid \theta)$$

  - New problem statement:

determine θ that maximizes the log-likelihood

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

- Set of necessary conditions for an optimum is:

$$(\nabla_\theta l = \sum_{k=1}^{k=n} \nabla_\theta \ln P(x_k \mid \theta))$$

$$\nabla_\theta l = 0$$

- Global maximum, local maximum or minimum, inflection point
- MAP estimators (Max a posteriori)

$$l(\theta)p(\theta)$$

- Example of a specific case 1: unknown μ

  - $p(x_i \mid \mu) \sim N(\mu, \Sigma)$
    (Samples are drawn from a multivariate normal population)

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

$$and \quad \nabla_{\mu} \ln p(x_k \mid \mu) = \Sigma^{-1}(x_k - \mu)$$

  - The ML estimate for μ must satisfy:

$$\sum_{k=1}^{k=n} \Sigma^{-1}(x_k - \hat{\mu}) = 0$$

- Multiplying by Σ and rearranging, we obtain:

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

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

Conclusion:

If $P(x_k \mid \omega_j)$ ($j = 1, 2, \ldots, c$) 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!

- Example of a specific case 2
  - Gaussian Case: *unknown* μ *and* σ

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

$$l = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \dfrac{\partial}{\partial\theta_1}(\ln P(x_k \mid \theta)) \\ \dfrac{\partial}{\partial\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = 0$$

$$\begin{cases} \dfrac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$

Summation:

$$\begin{cases} \sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & (1) \\ -\sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{k=n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

Combining (1) and (2), one obtains:

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

- Bias
  - ML estimate for σ^2 is biased

$$E[\overset{\wedge 2}{\sigma}] = E\left[\frac{1}{n}\Sigma(x_i - \bar{x})^2\right] = \frac{n-1}{n}.\sigma^2 \neq \sigma^2$$

  - An elementary unbiased estimator for Σ is:

$$\underbrace{\mathrm{C} = \frac{1}{\mathrm{n}-1}\sum_{k=1}^{k=n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t}_{Sample\,\mathrm{covariance\,matrix}}$$

  - ML estimate for Σ is biased

$$\overset{\wedge}{\Sigma} = \frac{n-1}{n}C$$

- Absolutely unbiased, asymptotically unbiased
- Prove ML estimate for σ^2 is biased

$$E[x^2] = D[x] + E[x]^2$$

$$E[\sum_{i=1}^{n} x_i^2] = n(\sigma^2 + \mu^2)$$

$$E[\bar{x}^2] = D[\bar{x}] + E(\bar{x})^2 = \frac{1}{n}\sigma^2 + \mu^2$$

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

$$= \frac{1}{n}[n(\sigma^2 + \mu^2) - n(\frac{1}{n}\sigma^2 + \mu^2)] = \frac{n-1}{n}\sigma^2$$