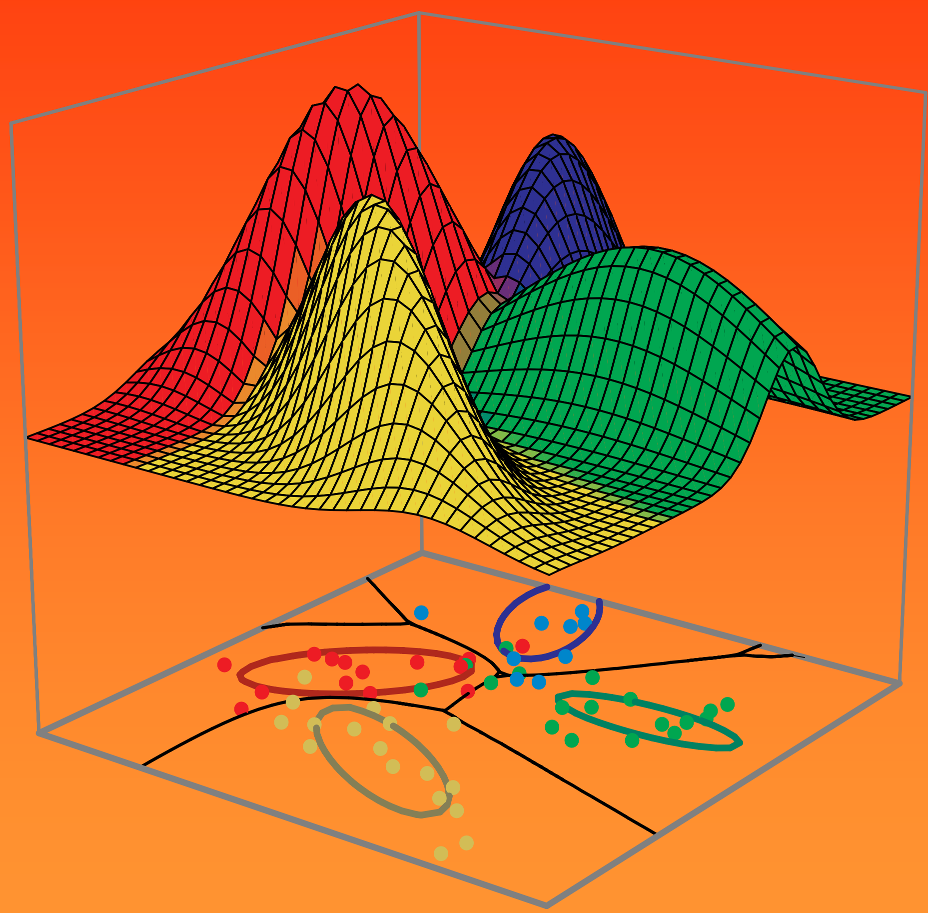


Pattern Classification

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 (part 1)

- Introduction
- Maximum-Likelihood Estimation
  - Example of a Specific Case
  - The Gaussian Case: unknown μ and σ
  - Bias
- Appendix: ML Problem Statement

- 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 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
  - Do we know something about the distribution?
  - → find parameters to characterize the distribution

- Example: Normality of $P(x \mid \omega_i)$

$$P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

  - Characterized by 2 parameters

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

- 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 distribution

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

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

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

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

$$\theta = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, \ldots, \sigma_j^{11}, \sigma_j^{22}, \mathrm{cov}(x_j^m, x_j^n) \ldots)$$

- Use the information provided by the training samples to estimate $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, each θ_i $(i = 1, 2, \ldots, c)$ is associated with each category

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

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

$P(D \mid \theta)$ is called the likelihood of θ w.r.t. the set of samples)

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

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

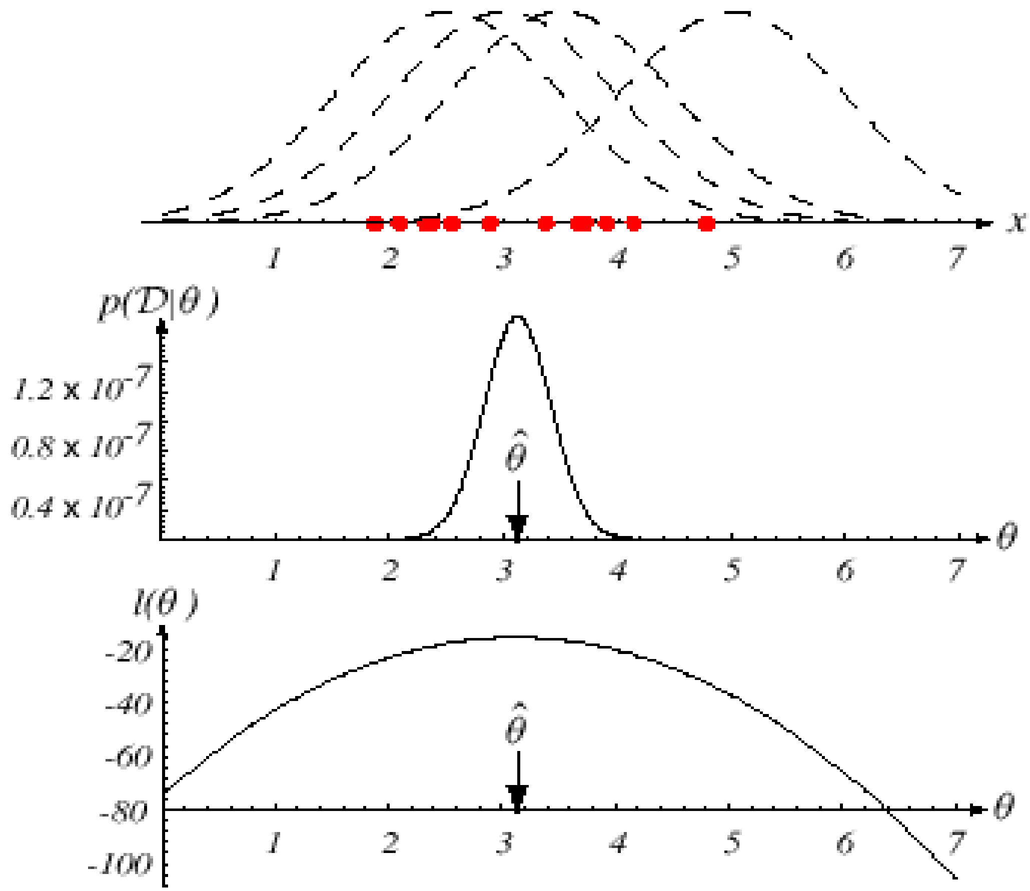


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$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

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

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

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

(recall *D* is the training data)

  - New problem statement:
    determine θ that maximizes the log-likelihood

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

The definition of *l*() is:

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

and

$$(\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{k=n} \nabla_{\boldsymbol{\theta}} \ln P(\mathbf{x}_k \mid \boldsymbol{\theta})) \qquad (\text{eq } 6)$$

Set of necessary conditions for an optimum is:

$$\nabla_{\theta} l = 0 \qquad (\text{eq. } 7)$$

- Example, the Gaussian case: unknown μ
  - We assume we know the covariance
  - $p(x_i \mid \mu) \sim N(\mu, \Sigma)$
    (Samples are drawn from a multivariate normal population)

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

$$\textit{and} \quad \nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \qquad (\text{eq. } 9)$$

$\theta = \mu$ therefore:
The ML estimate for μ must satisfy:

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0} \quad \text{from eqs } 6,7 \,\&\, 9$$

- Multiplying by Σ and rearranging, we obtain:

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

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, Gaussian Case: *unknown μ and Σ*
  - *First consider univariate case: unknown μ and σ*

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

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

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

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

Summation (over the training set):

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

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

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

- The ML estimates for the multivariate case is similar
  - The scalars χ and μ are replaced with vectors
  - The variance σ^2 is replaced by the covariance matrix

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

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

Bias

- ML estimate for σ^2 is biased

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

- Extreme case: n=1, E[ ] = 0 ≠ σ^2

- As *n* increases the bias is reduced

→ this type of estimator is called *asymptotically unbiased*

- An elementary unbiased estimator for Σ is:

$$\underbrace{\mathbf{C} = \frac{1}{n\text{-}1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t}_{\text{Sample covariance matrix}}$$

This estimator is unbiased for all distributions

→ Such estimators are called *absolutely unbiased*

- Our earlier estimator for Σ is biased:

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

In fact it is asymptotically unbiased:

Observe that

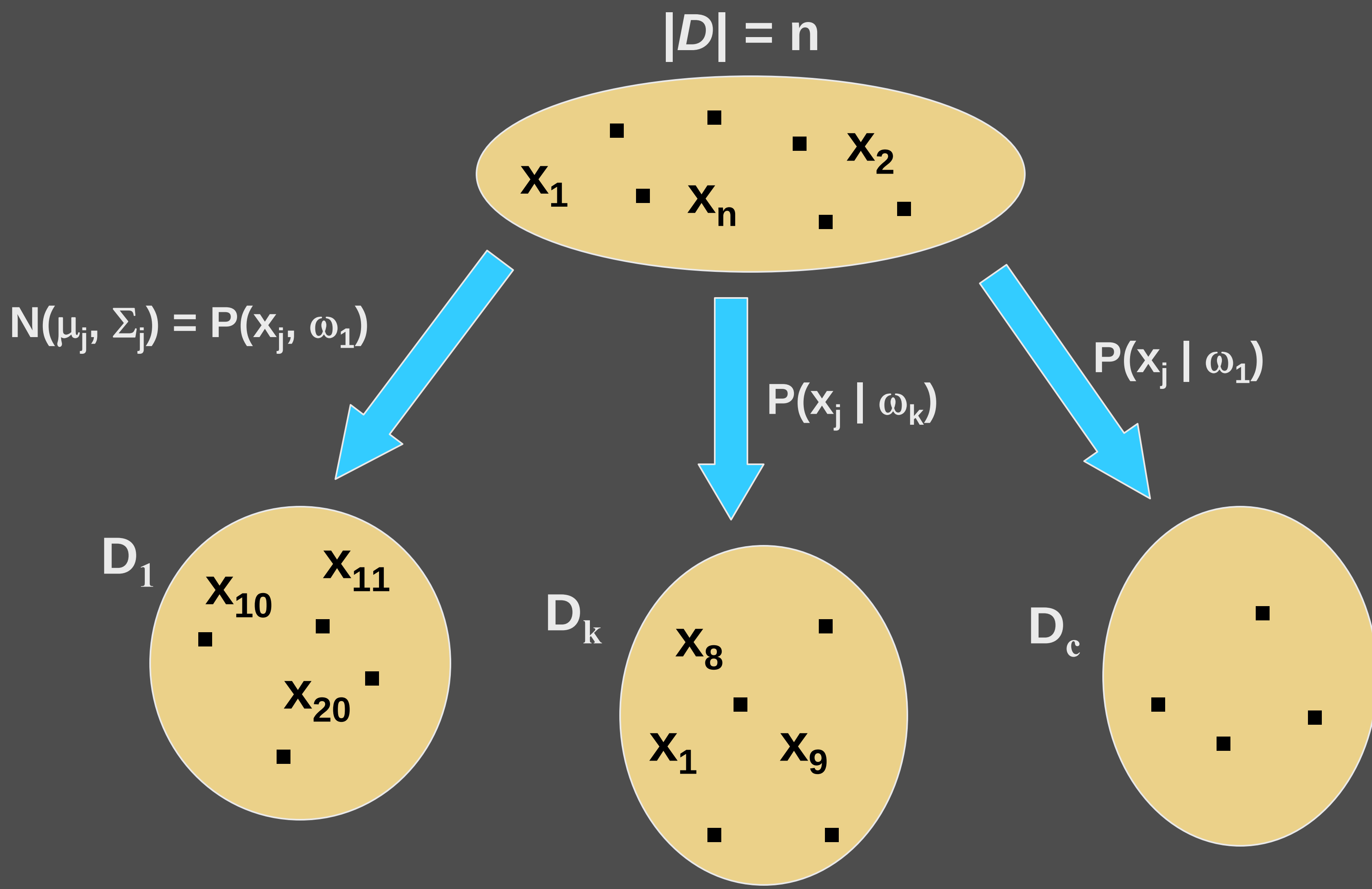
$$\hat{\Sigma} = \frac{n-1}{n} C$$

- Appendix: ML Problem Statement

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

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

    Our goal is to determine $\hat{\theta}$ (value of θ that maximizes the likelihood of this sample set!)


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

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

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

$$\underset{\theta}{\mathrm{Max}}\, P(D \mid \theta) = \mathrm{Max}\, P(x_1, \ldots, x_n \mid \theta)$$

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