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

- 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:

$$q = (m_j, S_j) = (m_j^1, m_j^2, \ldots, s_j^{11}, s_j^{22}, \mathbf{cov}(x_j^m, x_j^n)\ldots)$$

- 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, $x_1, x_2, \ldots, x_n$

  (samples drawn independently)

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

$P(D \mid \theta)$ 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"

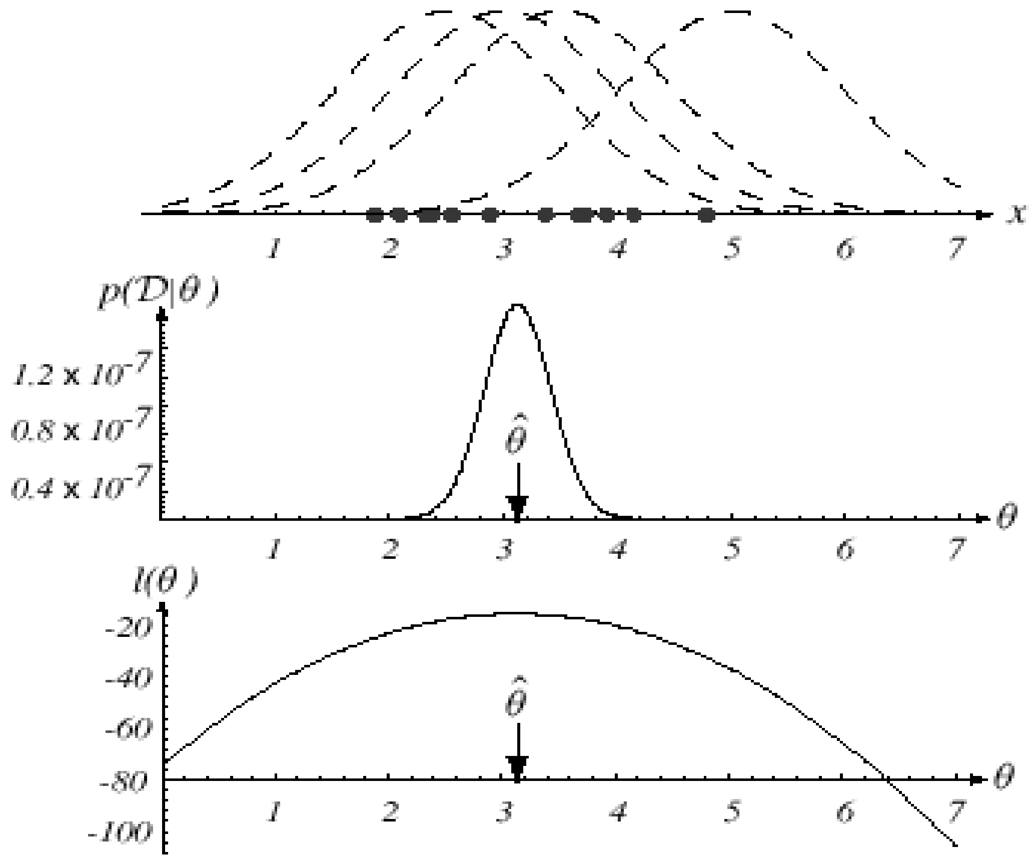


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ 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*. Copyright © 2001 by John Wiley & Sons, Inc.

- 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 $\ell(\theta)$ as the log-likelihood function

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

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

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

Set of necessary conditions for an optimum is:

$$(\nabla_q \mathbf{l} = \sum_{k=1}^{k=n} \nabla_q \ln P(x_k \mid q))$$

$$\nabla_\theta \ell = 0$$

- Example of a specific case: unknown μ

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

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

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

$\theta = \mu$ therefore:

- 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!

- ML Estimation:
  - Gaussian Case: *unknown m and s*
    (one dimension) $q = (q_1, q_2) = (\mu, \sigma^2)$

$$\mathbf{l} = \ln P(x_k \mid q) = -\frac{1}{2}\ln(2\pi q_2) - \frac{1}{2q_2}(x_k - q_1)^2$$

$$\nabla_\theta \mathbf{l} = \begin{pmatrix} \dfrac{\partial}{\partial q_1}(\ln P(x_k \mid q)) \\ \dfrac{\partial}{\partial q_2}(\ln P(x_k \mid q)) \end{pmatrix} = 0$$

$$\begin{cases} \dfrac{1}{q_2}(x_k - q_1) = 0 \\ -\dfrac{1}{2q_2} + \dfrac{(x_k - q_1)^2}{2q_2^2} = 0 \end{cases}$$

Summation:

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

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

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

- Bias

  - ML estimate for σ^2 is biased

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

  - An elementary unbiased estimator for Σ is:

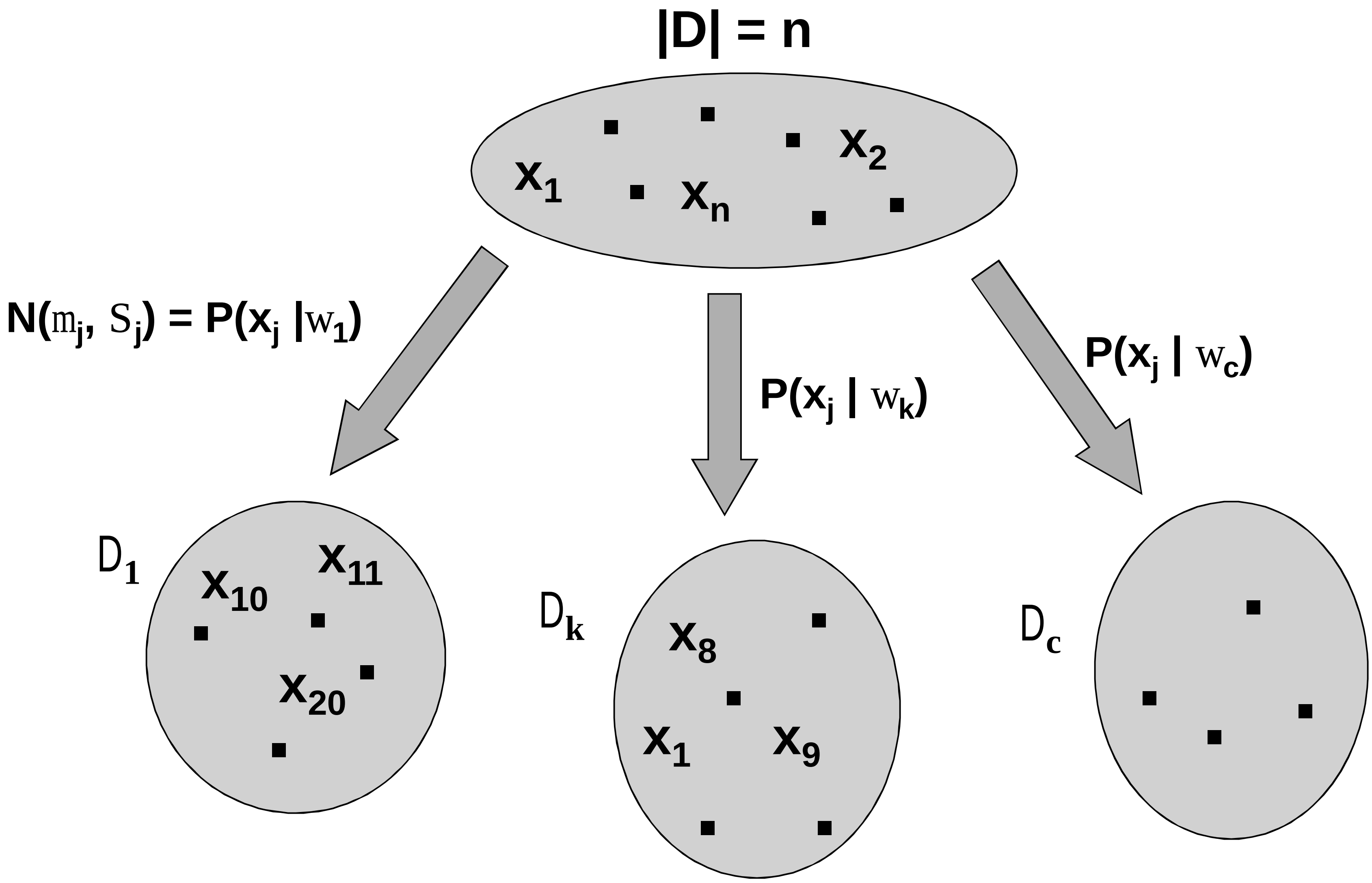
$$C = \underbrace{\frac{1}{n-1}\sum_{k=1}^{k=n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t}_{\text{Sample covariance matrix}}$$

- 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 $\hat{q}$ (value of θ that makes this sample the most representative!)


|D| = n
x1
x2
xn
N(mj, Sj) = P(xj |w1)
P(xj | wk)
P(xj | wc)
D1
x10
x11
x20
Dk
x8
x1
x9
Dc

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

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

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

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

summary

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

$$\theta=\left[\mu \quad \sigma^2\right]^T$$

$$\ln\left(P(X_k|\theta)\right)=-\frac{1}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2}\frac{(X_k-\mu)^2}{\sigma^2}$$

$$\frac{\partial}{\partial\mu}\ln\left(P(X_k|\theta)\right)=\frac{(X_k-\mu)}{\sigma^2}$$

$$\frac{\partial}{\partial(\sigma^2)}\ln\left(P(X_k|\theta)\right)=-\frac{1}{2\sigma^2}+\frac{(X_k-\mu)^2}{2(\sigma^2)^2}$$

$$\nabla_\theta l=\sum_{k=1}^{N}\nabla_\theta\ln\left(P(X_k|\theta)\right)=0$$

$$\sum_{k=1}^{N}\frac{(X_k-\mu)}{\sigma^2}=0 \rightarrow \mu_{ML}=\frac{1}{N}\sum_{k=1}^{N}X_k=\overline{\mu}$$

$$\frac{1}{2\sigma^2}\sum_{k=1}^{N}\left(\frac{(X_k-\mu)^2}{\sigma^2}-1\right)=0 \rightarrow \sigma_{ML}^2=\frac{1}{N}\sum_{k=1}^{N}(X_k-\overline{\mu})^2$$