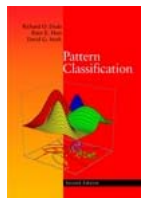
Chapter 3 (part 1): Maximum-Likelihood & Bayesian Parameter Estimation

- 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:

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

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

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

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

$P(D \mid \theta)$ 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"

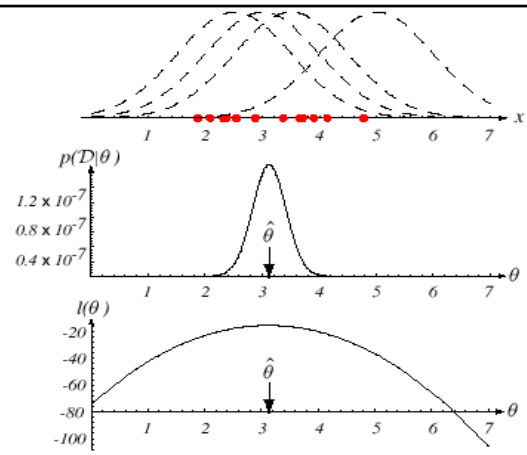


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$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

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

  - 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:

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

$$\nabla_\theta l = 0$$

- 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, …, 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* μ *and* σ

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

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

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

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

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

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

  - An elementary unbiased estimator for Σ is:

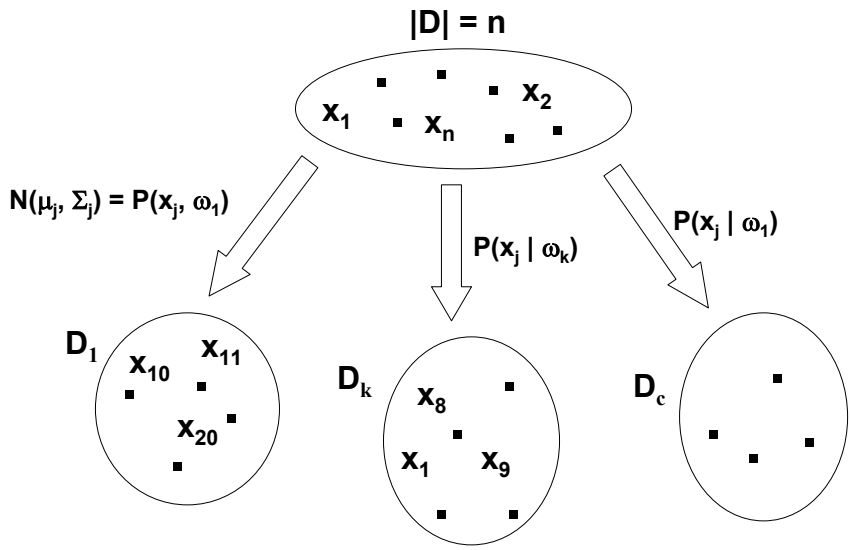
$$C = \frac{1}{n} \sum^{k=n} (x_k - \mu)(x_k - \mu)^t$$

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



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

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

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

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