**Announcements**

- Readings on E-reserves
- Project proposal due today

---

**Parameter Estimation**

Biometrics  
CSE 190-a  
Lecture 6

---

**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 $\mu$ and $\sigma$
- Bias

---

**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 $P(o_i \mid x)$ 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

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

$$P(x \mid o_j) = P(x \mid o_j, \theta)$$ where:

$$\theta = (\mu_1, \Sigma_1, \mu_2, \Sigma_2, \ldots, \mu_c, \Sigma_c, \text{cov}(x_1, x_2), \text{cov}(x_1, x_3), \ldots)$$

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, $x_1, x_2, \ldots, x_n$

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

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

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"

Optimal estimation

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ and let $\nabla_\theta$ 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 $\hat{\theta}$ that maximizes the log-likelihood

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

Set of necessary conditions for an optimum is:

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

$$\nabla_\theta l = 0$$
Example of a specific case: unknown $\mu$, $\Sigma$ known

$P(x | \mu) = N(\mu, \Sigma)$ (Samples are drawn from a multivariate normal population)

$$z = \mu$$

Therefore:

- The ML estimate for $\mu$ must satisfy:

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

Multiplying by $\Sigma$ and rearranging, we obtain:

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

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

Conclusion:

If $P(x_k | \omega_j)$ ($j = 1, 2, \ldots, c$) 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 $\mu$ and $\sigma$

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

$$\begin{align*}
\theta_1 &= \frac{1}{n} \sum_{k=1}^{n} x_k \\
\theta_2 &= \frac{1}{n} \sum_{k=1}^{n} (x_k - \theta_1)^2
\end{align*}$$

Bias

- ML estimate for $\sigma^2$ is biased

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

An elementary unbiased estimator for $\Sigma$ is:

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

Sample covariance matrix

Appendix: ML Problem Statement

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

$$P(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} P(x_k | \theta)$$

$|D| = n$

Our goal is to determine $\hat{\theta}$ (value of $\theta$ that makes this sample the most representative!)
Bayesian Parameter Estimation (part 2)

- Bayesian Estimation (BE)
- Bayesian Parameter Estimation: Gaussian Case
- Bayesian Parameter Estimation: General Estimation
- Problems of Dimensionality
- Computational Complexity
- Component Analysis and Discriminants
- Hidden Markov Models

Bayesian Estimation

- In MLE $\theta$ was supposed fix
- In BE $\theta$ is a random variable
- The computation of posterior probabilities $P(\omega_i \mid x)$ that is used for classification lies at the heart of Bayesian classification
- Given the sample $D$, Bayes formula can be written

$$P(\omega_i \mid x, D) = \frac{p(x \mid \omega_i, D)P(\omega_i \mid D)}{\sum_{j=1}^{c} p(x \mid \omega_j, D)P(\omega_j \mid D)}$$

Bayesian Parameter Estimation: Gaussian Case

**Step 1**: Estimate $\theta$ using the a-posteriori density $P(\theta \mid D)$

- $\mu$ is the only unknown parameter
- $p(x \mid \mu) \sim N(\mu, \sigma^2)$
- $p(\mu) \sim N(\mu_0, \sigma_0^2)$

($\mu_0$ and $\sigma_0$ are known!)

We assume that
- Samples $D_i$ provide info about class $i$ only, where $D=(D_1, \ldots, D_c)$
- $P(\omega_i \mid D_i)$ (i.e., samples $D_i$ determine the prior on $\omega_i$)

Goal: compute $p(\omega_i \mid x, D_i)$

So now what do we do??? Well, the only term we don’t know on the right-side of

$$P(\omega_i \mid x, D_i) = \frac{p(x \mid \omega_i, D_i)P(\omega_i)}{\sum_{j=1}^{c} p(x \mid \omega_j, D_i)P(\omega_j)}$$

is $p(x \mid \omega_i, D_i)$ the class conditional density, but this involves a parameter $\theta$ that is a random variable.

If we knew $\theta$ we would be done! But we don’t know it.

We do know that
- $\theta$ has a known prior $p(\theta)$
- and we have observed samples $D_i$

So we can re-write the ccd as

$$p(x \mid D) = \int p(x, \theta \mid D)d\theta$$

$$= \int p(x \mid \theta, D)p(\theta \mid D)d\theta$$

$$= \int p(x \mid \theta)p(\theta \mid D)d\theta$$
So now we must calculate

\[ p(\mu | D) = \frac{p(D | \mu) p(\mu)}{\int p(D | \mu) p(\mu) d\mu} \]

where

\[ N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(\mu - \mu_0)^2}{2\sigma^2}} \]

and\[ \sigma^2 = \frac{\sigma_0^2 + n \sigma^2}{n + 1} \]

Reproducing density is found as

\[ p(\mu | D) = N(\mu_n, \sigma_n^2) \]

Bayesian Parameter Estimation: Gaussian Case

Step II: \( p(x | D) \) remains to be computed!

\[ p(x | D) = \int p(x | \mu) p(\mu | D) d\mu \]

is Gaussian

So the desired cdf \( p(x | D) \) can be written as

\[ p(x | D) = N(\mu_n, \sigma_n^2) \]

Bayesian Parameter Estimation: General Theory

\[ p(x | D) \] computation can be applied to any situation in which the unknown density can be parameterized.

The basic assumptions are:

- The form of \( p(x | \theta) \) is assumed known, but the value of \( \theta \) is not
- Our knowledge about \( \theta \) is contained in a known prior density \( p(\theta) \)
- The rest of our knowledge \( \theta \) is contained in a set \( D \) of \( n \) random variables \( x_1, x_2, \ldots, x_n \) that follows \( P(x) \)

The basic problem is:

\[ P(x | \theta) \]

\[ P(\theta | D) \]

\[ P(\theta | \theta) \]

Using Bayes formula, we have:

\[ p(\theta | D) = \frac{p(D | \theta) p(\theta)}{\int p(D | \theta) p(\theta) d\theta} \]

And by an independence assumption:

\[ p(D | \theta) = \prod_{x \in D} p(x | \theta) \]
Why Don’t We Always Acquire More Features?

Consider case of two classes multivariate normal with the same covariance:

\[ P(\text{error}) = \frac{1}{\sqrt{2\pi r}} \int e^{-t^2/2} \, dt \]

where:

\[ r^2 = (\mu_1 - \mu_2) \Sigma^{-1} (\mu_1 - \mu_2) \]

\[ \lim_{r \to \infty} P(\text{error}) = 0 \]

Problems of Dimensionality

If features are independent then:

\[ \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2) \]

\[ r^2 = \sum_{i=1}^{d} \left( \frac{\mu_{1i} - \mu_{2i}}{\sigma_i} \right)^2 \]

Most useful features are the ones for which the difference between the means is large relative to the standard deviation.

It has frequently been observed in practice that, beyond a certain point, the inclusion of additional features leads to worse rather than better performance: we have the wrong model!