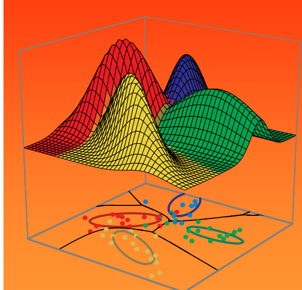
Parameter Estimation

Biometrics
CSE 190-a
Lecture 6

Announcements

- Readings on E-reserves
- Project proposal due today



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
-

- 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

6

- 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!

7

- **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:

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

8

- 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"

9

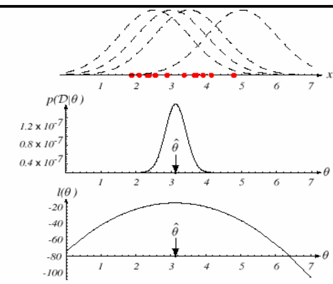


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$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

10

- 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

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

11

Set of necessary conditions for an optimum is:

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

$$\nabla_\theta l = 0$$

12

- Example of a specific case: unknown μ, Σ known
  - $P(x_i \mid \mu) \sim N(\mu, \Sigma)$
    (Samples are drawn from a multivariate normal population)

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

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

$\theta = \mu$ therefore:

- The ML estimate for μ must satisfy:

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

13

- 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!

14

- ML Estimation:
  - Gaussian Case: *unknown μ and σ*
    $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

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

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

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

15

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:

$$\hat{\theta}_1 = \hat{\mu} = \sum_{k=1}^{k=n} \frac{x_k}{n} \quad ; \quad \hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum_{k=1}^{k=n}(x_k - \mu)^2}{n}$$

16

- 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:

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

17

- 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!)

18

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

19

Bayesian Estimation

- In MLE θ was supposed fix
- In BE θ 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

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

20

- We assume that
  - Samples D_i provide info about class i only, where $D=\{D_{1, \dots,} D_c\}$
  - $P(\omega_i) = P(\omega_i \mid D_i)$ (i.e., samples D_i determine the prior on ω_i)

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

- Goal: compute $p(\omega_i \mid x, \mathrm{D}_i)$

21

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

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

is $p(x \mid \omega_i, \mathrm{D}_i)$ the class conditional density, but this involves a parameter θ that is a random variable.

22

- If we knew θ we would be done! But we don't know it.
- We do know that
  - θ has a known prior p(θ)
  - and we have observed samples D_i.
- So we can re-write the ccd as

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

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

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

3

23

Bayesian Parameter Estimation: Gaussian Case

Step I: Estimate θ using the a-posteriori density P(θ | D)

The univariate case: P(μ | D)

μ is the only unknown parameter

$$p(\mathbf{x} \mid \mu) \sim \mathbf{N}(\mu, \sigma^2)$$
$$p(\mu) \sim \mathbf{N}(\mu_0, \sigma_0^2)$$

(μ_0 and σ_0 are known!)

24

- So now we must calculate

$$p(\mu \mid \mathrm{D}) = \frac{p(D \mid \mu)\, p(\mu)}{\int p(D \mid \mu)\, p(\mu)\, d\mu}$$

$$= \alpha \prod_{k=1}^{n} p(x_k \mid \mu)\, p(\mu)$$

- Reproducing density is found as

$$p(\mu \mid \mathrm{D}) \sim \mathbf{N}(\mu_n, \sigma_n^2)$$

where

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\, \mu_0$$

$$\text{and} \quad \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

25

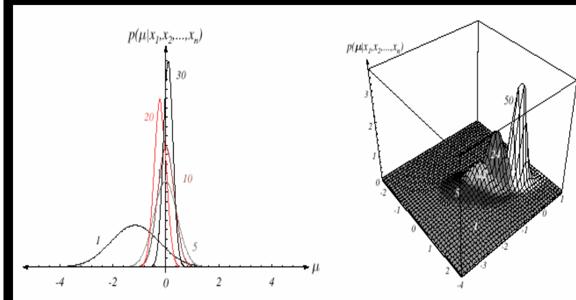


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

26

Bayesian Parameter Estimation: Gaussian Case

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

$$p(x \mid \mathrm{D}) = \int p(x \mid \mu)\, p(\mu \mid \mathrm{D})\, d\mu \ \ \textbf{is Gaussian}$$

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

$$p(x \mid \mathrm{D}) \sim \mathbf{N}(\mu_n, \sigma^2 + \sigma_n^2)$$

27

Bayesian Parameter Estimation: Gaussian Case

Step III: We do this for each class and combine P(x | D_j, ω_j) with P(ω_j) along with Bayes rule to get

$$\underset{\omega_j}{\text{Max}} \left[P(\omega_j \mid x, D) \right] \equiv \underset{\omega_j}{\text{Max}} \left[P(x \mid \omega_j, D_j) . P(\omega_j) \right]$$

28

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

29

The basic problem is:

Step I: Compute the posterior density P(θ | D)

Step II: Derive P(x | D)

Step III: Compute p(ω | x, D)

Using Bayes formula, we have:

$$p(\theta \mid \mathrm{D}) = \frac{p(\mathrm{D} \mid \theta)\, p(\theta)}{\int p(\mathrm{D} \mid \theta)\, p(\theta)\, d\theta}$$

And by an independence assumption:

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

30

Why Don't We Always Acquire More Features?

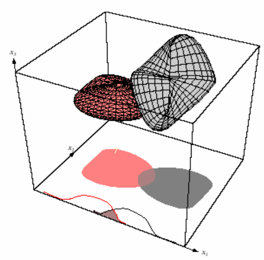


FIGURE 3.3. Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace—here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace—there can be greater overlap of the projected distributions, and hence greater Bayes error. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

31

Problems of Dimensionality

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

$$P(\text{error}) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{\frac{-u^2}{2}} du$$

$$\text{where}: \; r^2 = (\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)$$

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

32

- If features are independent then:

$$\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$$

$$r^2 = \sum_{i=1}^{i=d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$

- 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 !

7