0

These slides are edited by Roz from originals at

- http://rii.ricoh.com/~stork/DHS.html

Powerpoint lecture slides - DHSch3part2.ppt

1

Chapter 3: Maximum-Likelihood and Bayesian Parameter Estimation (part 2)

- Bayesian Estimation (BE)
- Bayesian Parameter Estimation: Gaussian Case
- Bayesian Parameter Estimation: General Estimation
- Problems of Dimensionality

2

- Bayesian Estimation (Bayesian learning to pattern classification problems)
  - In MLE θ was supposed fix
  - In BE θ is a random variable
  - The computation of posterior probabilities $P(\omega_i \mid x)$ lies at the heart of Bayesian classification
  - Goal: compute $P(\omega_i \mid x, \mathcal{D})$

    Given the sample $\mathcal{D}$, Bayes formula can be written

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

3

- To demonstrate the preceding equation, use:

$$P(x, D \mid \omega_i) = P(x \mid D\omega_i).P(D \mid \omega_i)$$

$$P(x \mid D) = \sum_j P(x, \omega_j \mid D)$$

$P(\omega_i) = P(\omega_i \mid D)$ **(Training sample provides this!)**

Thus :

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

4

- Bayesian Parameter Estimation: Gaussian Case

Goal: Estimate θ using the a-posteriori density $P(\theta \mid D)$

- The univariate case: $P(\mu \mid D)$

  μ is the only unknown parameter

$$P(x \mid \mu) \sim N(\mu, \sigma^2)$$
$$P(\mu) \sim N(\mu_0, \sigma_0^2)$$

(μ_0 and σ_0 are known!)

5

$$P(\mu \mid D) = \frac{P(D \mid \mu).P(\mu)}{\int P(D \mid \mu).P(\mu)d\mu} \qquad (1)$$
$$= \alpha \prod_{k=1}^{k=n} P(x_k \mid \mu).P(\mu)$$

- Reproducing density

$$P(\mu \mid D) \sim N(\mu_n, \sigma_n^2) \qquad (2)$$

Identifying (1) and (2) yields:

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.\mu_0$$
$$and \quad \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

6

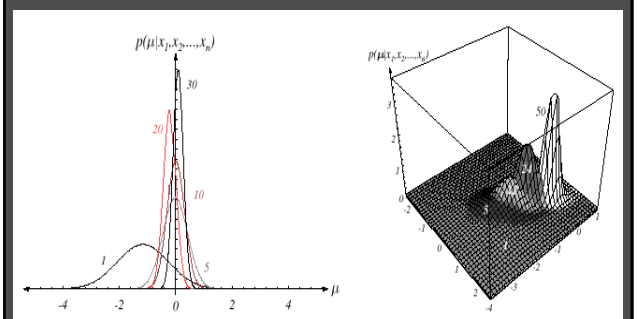

FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001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

7

- The univariate case $P(x \mid D)$
  - $P(\mu \mid D)$ computed
  - $P(x \mid D)$ remains to be computed

$$P(x \mid D) = \int P(x \mid \mu).P(\mu \mid D)d\mu \text{ is Gaussian}$$

It provides:

$$P(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

(Desired class-conditional density $P(x \mid D_j, \omega_j)$)

Therefore: $P(x \mid D_j, \omega_j)$ together with $P(\omega_j)$ and using Bayes formula, we obtain the Bayesian classification rule:

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

8

Bayesian Parameter Estimation: General Theory

- P(x | θ) computation can be applied to any situation in which the unknown density can be parametrized: the basic assumptions are:
  - The form of P(x | θ) is assumed known, but the value of θ is not known exactly
  - Our knowledge about θ is assumed to be 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)

9

The basic problem is:

"Compute the posterior density P(θ | D)"

then "Derive P(x | D)"

Using Bayes formula, we have:

$$P(\theta \mid D) = \frac{P(D \mid \theta).P(\theta)}{\int P(D \mid \theta).P(\theta)d\theta},$$

And by independence assumption:

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

10

- Problems of Dimensionality
  - Problems involving 50 or 100 features (binary valued)
    - Classification accuracy depends upon the dimensionality and the amount of training data
    - Case of two classes multivariate normal with the same covariance

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

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

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