University of California
Berkeley

College of Engineering
Department of Electrical Engineering
and Computer Sciences

Professors : N.Morgan / B.Gold
EE225D

Spring,1999

Statistical Pattern Recognition

Lecture 9

Statistical Pattern Recognition

- Many sources of variability in speech signal
- Much left over after deterministic factors
- Powerful statistical math
- More general way of handling discrimination

Statistical Discrimination Functions

- Minimum Error Classifier and Bayes Rule
- Gaussian Classifiers
- Discrete Densities
- Mixture Gaussians
- Neural Networks

Random Variables

Continuous

temperature energy at 500Hz

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

$$\int_{-\infty}^{\infty} p(x, y)dxdy = 1$$

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

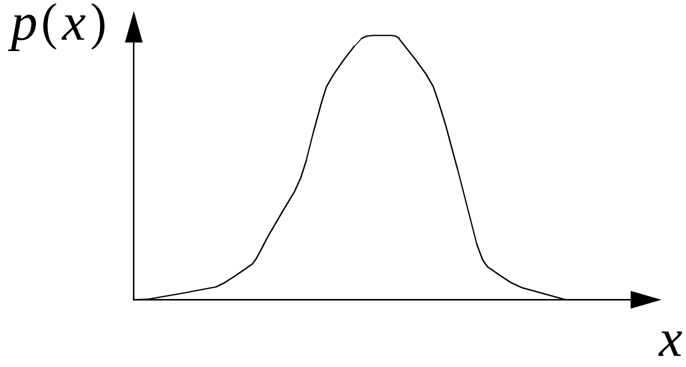
$$p(y|x) = \frac{p(x, y)}{p(x)}$$

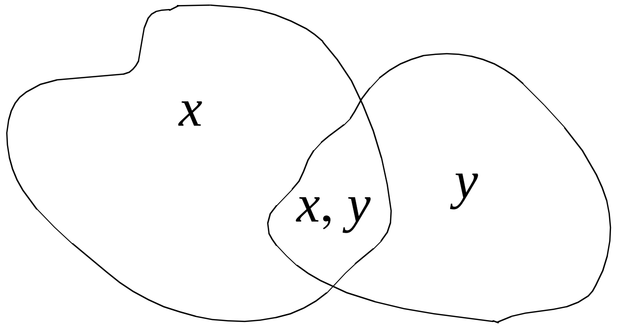
Discrete

(temp < 70°F) or (temp ≥ 70°F)

energy < 60 dBSPL at 500Hz

$$\sum p(x) = 1$$





$$p(x, y) = p(x|y)p(y)$$

$$p(x, y) = p(y|x)p(x)$$

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

likelihood

$$p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{p(x)} \longleftarrow \text{class prior}$$

a posteriori probability

MAP Decision Rule

choose *i* such that

$$p(\omega_i|x) > p(\omega_j|x) \quad \forall \quad j \neq i$$

x observed random variable

ω_i $(\quad \leq i \leq K)$ 1

$$p(\omega_i|x) > p(\omega_j|x) \quad \forall\ (j \neq i)$$

---

2 - class case :

$$p(\text{error}|x) = \begin{cases} p(\omega_1|x) \text{ if } (x \in \omega_2) \\ p(\omega_2|x) \text{ if } (x \in \omega_1) \end{cases}$$

$$p(\text{error}) = \int p(\text{error}, x)dx = \int p(\text{error}|x)p(x)d(x)$$

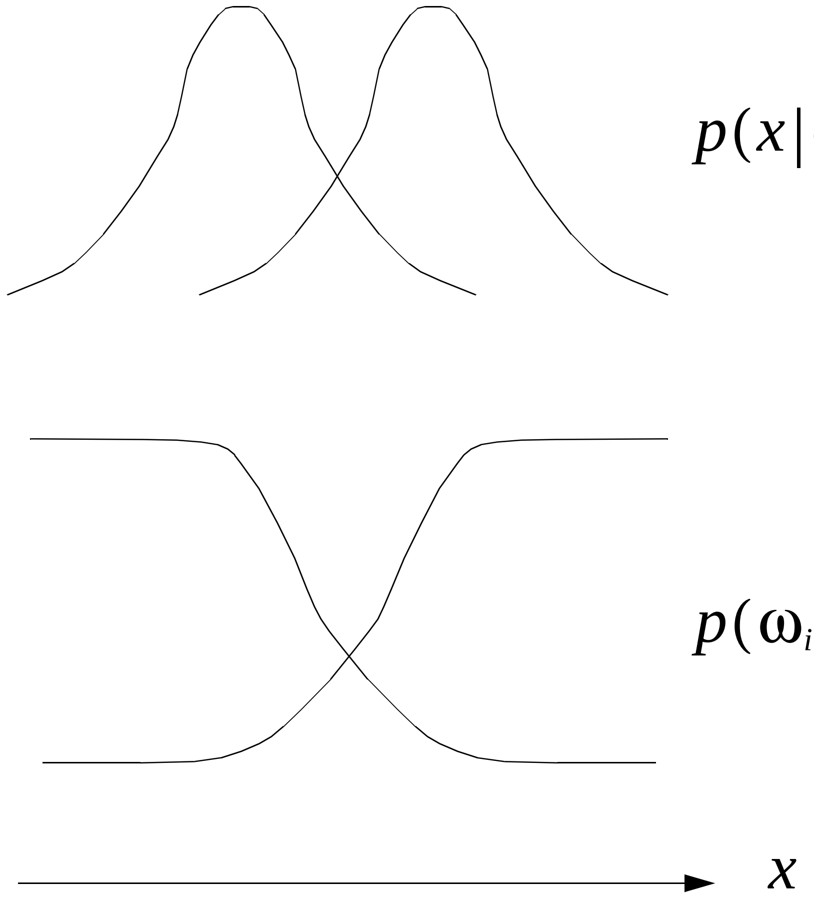
$$p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{p(x)}$$

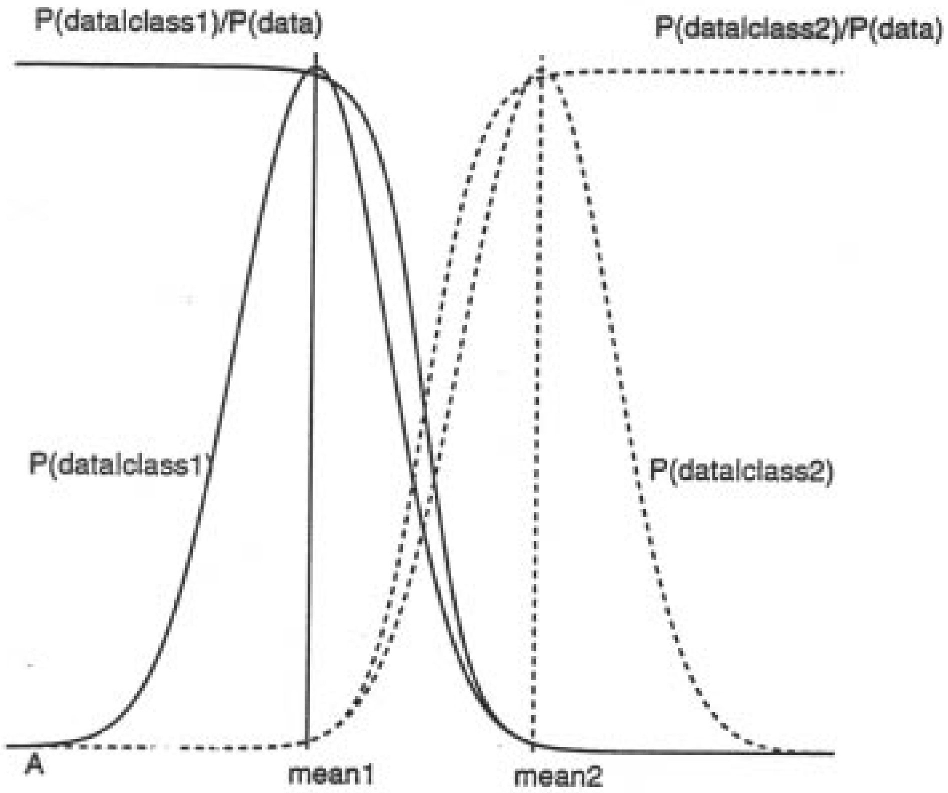
$$\frac{p(\omega_i|x)}{p(\omega_j|x)} = \frac{p(x|\omega_i)p(\omega_i)}{p(x|\omega_j)p(\omega_j)} > 1$$

$$\frac{p(x|\omega_i)}{p(x|\omega_j)} > \frac{p(\omega_j)}{p(\omega_i)}$$

$$\log p(x|\omega_i) + \log p(\omega_i)$$

$$p(x) = \sum_i p(x, \omega_i) = \sum_i p(x|\omega_i)p(\omega_i)$$




P(data|class1)/P(data)
P(data|class2)/P(data)
P(data|class1)
P(data|class2)
A
mean1
mean2

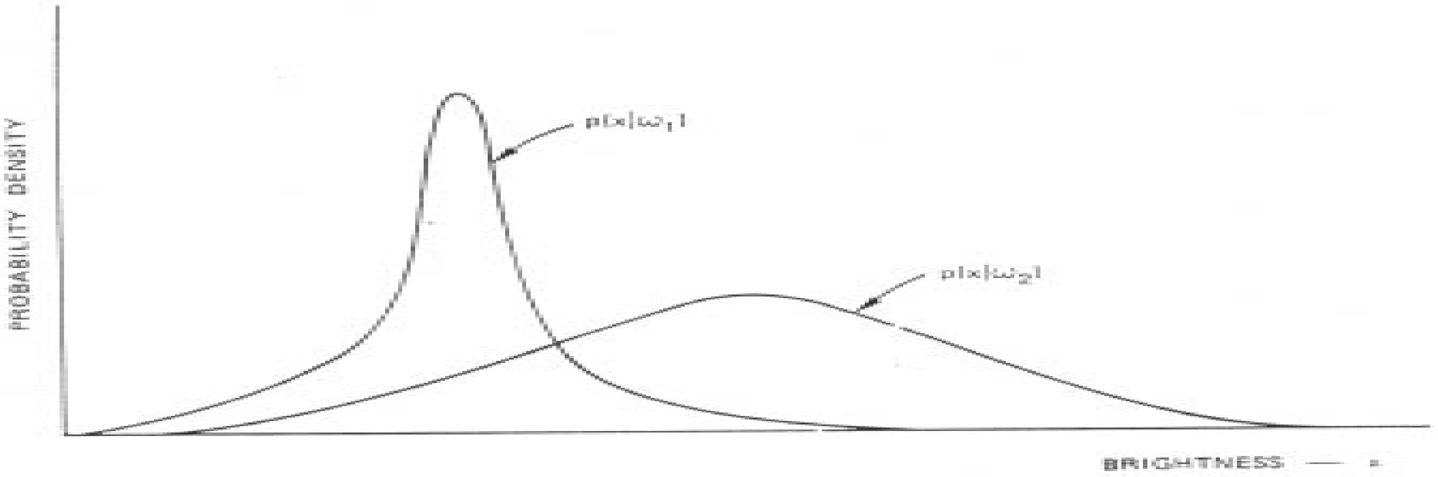


FIGURE 2.1. Hypothetical class-conditional probability density functions.

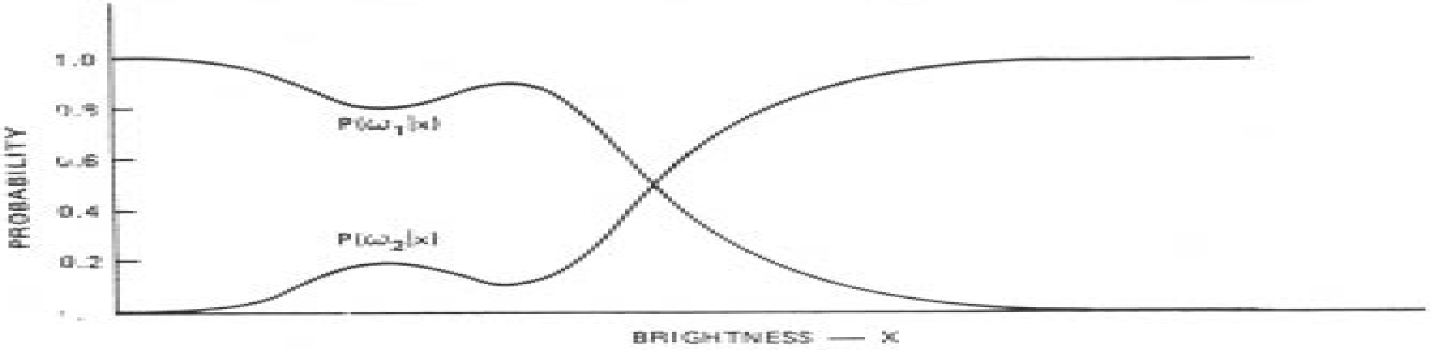


FIGURE 2.2. A posteriori probabilities for $P(\omega_1) = \frac{2}{3}$, $P(\omega_2) = \frac{1}{3}$.

How to Approximate Bayes Classifier

- Parametric form with single pass estimation
- Discretize, count co-occurences
- Iterative training (gaussian mixtures, neural nets)
- Kernel estimation

$$\text{uni} \qquad p(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{x+\mu_i}{\sigma_i}\right)^2\right]$$

$$\text{multi} \quad p(x|\omega_i) = \frac{1}{(2\pi)^{d/2}\left|\sum_i\right|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^T\sum_i^{-1}(x-\mu)\right]$$

$$\sigma^2 = \text{var} = \mathrm{E}[(x-\mu)^2]$$

$$\text{cov}(x,y) = \mathrm{E}[(x-\mu_x)(y-\mu_y)]$$

$$\begin{bmatrix} \ddots & & \text{cov} \quad (i,j) \\ & \sigma^2_{i,j} & \\ & & \end{bmatrix} \qquad \begin{bmatrix} \ddots & & \text{cov} \quad (i,j) \\ & \sigma^2_{i,j} & \\ & & \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix}$$

$$(x - \mu_i)^T \sum_i^{-1} (x - \mu_i)$$

Mahalanobis distance

$$\log p(x|\omega_i) \;=\; K_o + K_i + MD$$

Special Cases

If :

$$\sum_i = \sigma^2 I$$

(all features have the same variance, no correlation)

then a function to minimize is :

$$D_i(x) = \frac{\|x - \mu_i\|^2}{2\sigma^2} - \log p(\omega_i)$$

or

$$D'_i(x) = \|x - \mu_i\|^2 \quad \text{if all } p(\omega_i) \text{ are equal}$$

Minimum distance classifier !

Minimum distance classifiers, using Euclidean distance, are only optimum if (assuming Gaussian):

- Equal priors
- Uncorrelated Features

If priors are unknown, a reasonable guess is minimum distance with Mahalanobis distance.

Special Case 2

$\sum_i = \sum$ all covariance matrices are equal

setting

$$D_i(x) = \frac{1}{2}M^2(x, \mu_i) - \log p(\omega_i)$$

or

$$M^2(x, \mu_i) = (x - \mu_i)^T \sum^{-1} (x - \mu_i)$$

$$= x^T \underset{\underset{\text{indep of } i}{\uparrow}}{\sum^{-1}} x - 2x \sum^{-1} \mu_i + \mu_i^T \underset{\underset{\text{indep of } x}{\uparrow}}{\sum^{-1}} \mu_i$$

$$D_i'(x) = -\omega_i^T x + \omega_{i0}$$

where

$$\omega_i = \sum^{-1}\mu_i$$

and

$$\omega_{i0} = \frac{1}{2}\mu_i^T\sum^{-1}\mu_i - \log p(\omega_i)$$

Linear Classifier !

(Hyperplane decision boundry)

General Gaussian Case

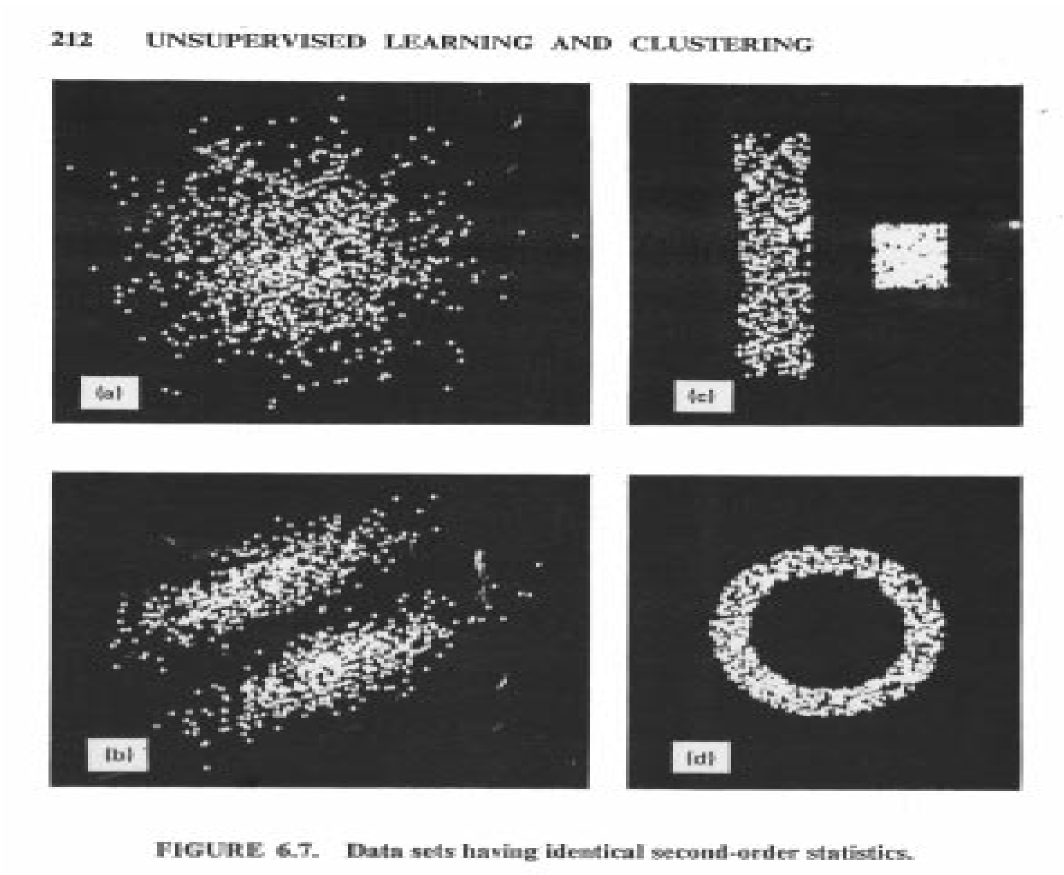
Σ_i arbitrary

$$D_i(x) = x^T W_i x + \omega_i^T x + \omega_{i0}$$

Quadratic Classifier

Gaussians are completely specified by 1st and 2nd order statistics.

Is this enough ?



212 UNSUPERVISED LEARNING AND CLUSTERING

FIGURE 6.7. Data sets having identical second-order statistics.