EE 6885 Statistical Pattern Recognition

Fall 2005
Prof. Shih-Fu Chang
http://www.ee.columbia.edu/~sfchang

Lecture 4 (9/19/05)

- Reading
  - Bayesian Classifiers for Multi-Variate Gaussian
    - Textbook DHS Chapter 2.6
  - Model Parameter Estimation
    - Textbook DHS Chapter 3
- Homework #1 due 2005-09-21, Wed

Discriminant Function for Gaussians

- If consider classification error as “loss” → mini error classification

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

- For Gaussian

$$P(x \mid \omega_i) = N(\mu_i, \Sigma_i)$$

$$P(\mathbf{x} \mid w_i) = \frac{1}{(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left(\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right)$$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Decision Boundaries for case I $\Sigma_i = \sigma^2 I$

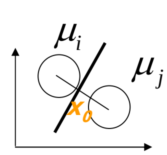
- Decision boundary between class i and j

$$g_i(x) = g_j(x) \qquad g_i(x) = \frac{1}{\sigma^2}\mu_i^t x + w_{i0}$$

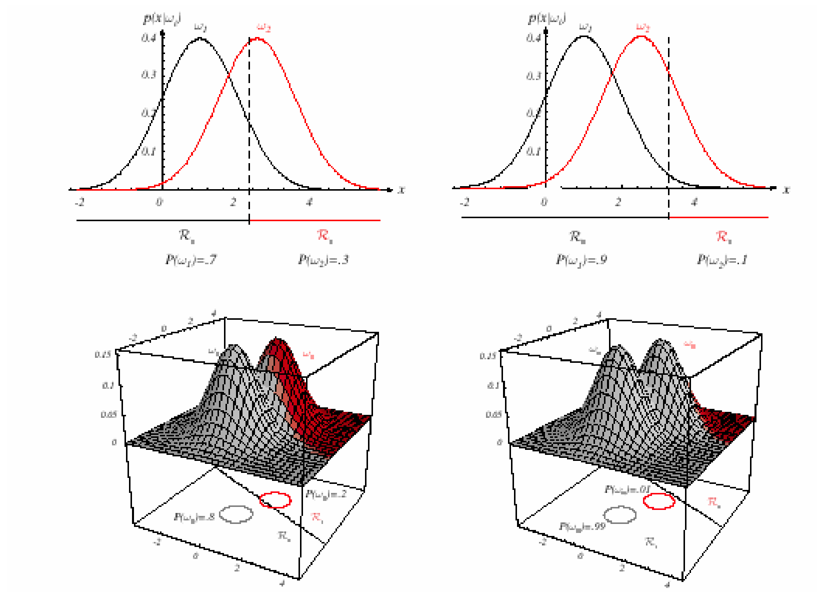
$$w^t(x - x_0) = 0 \qquad w = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{|\mu_i - \mu_j|^2} \ln\frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$$

- A hyperplane perpendicular to $\mu_i - \mu_j$ passing through point x_0



- If $P(\omega_i) = P(\omega_j)$ then $x_0 = \frac{1}{2}(\mu_i + \mu_j)$ i.e., midpoint between μ_i and μ_j



Case 2 $\Sigma_i = \Sigma$

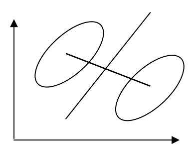
- Hyper-ellipsoidal shapes of equal size and shape

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| + \ln P(\omega_i)$$

$$g_i(x) = w_i^t x + w_{i0} \qquad \text{a hyperplane with } \textit{bias} \;\; w_{i0}$$

$$w_i = \Sigma^{-1}\mu_i^t \qquad w_{i0} = -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i + \ln P(\omega_i)$$

- Decision boundary



$$w^t(x - x_0) = 0 \qquad w = \Sigma^{-1}(\mu_i - \mu_j)$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i-\mu_j)^t \Sigma^{-1}(\mu_i-\mu_j)}(\mu_i - \mu_j)$$

- Not perpendicular
  to the line connecting the means

Case 2 decision boundaries

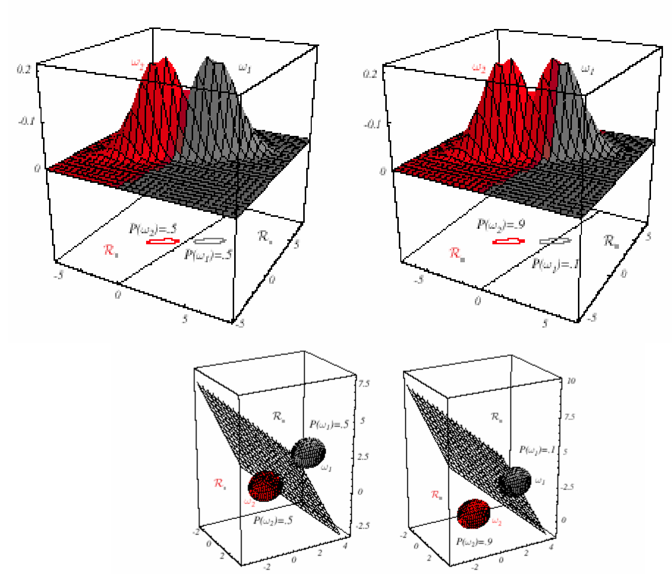
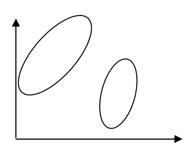


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork

Case 3 $\Sigma_i = \text{arbitrary}$



- Discriminant functions are quadratic

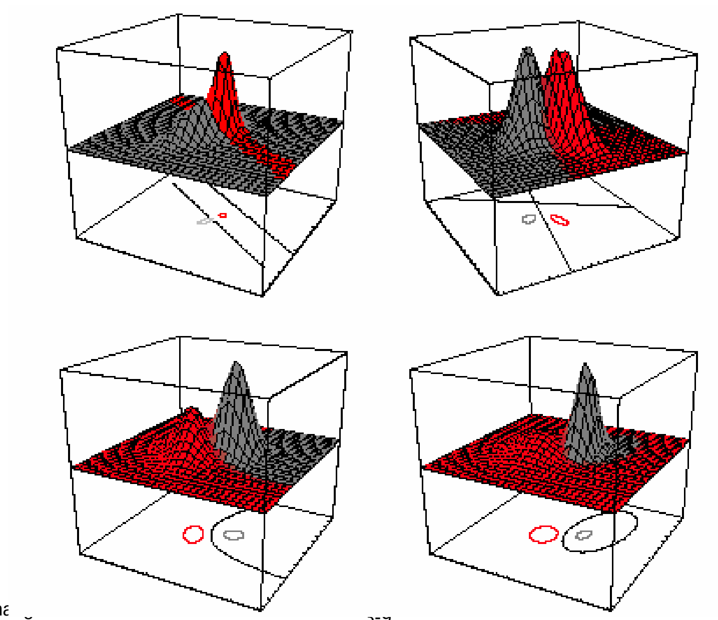
$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$

$$where: \quad \mathrm{W_i} = -\frac{1}{2}\Sigma_i^{-1} \qquad \mathrm{w_i} = \Sigma_i^{-1}\mu_i$$

$$\mathrm{w}_{i0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- The decision surfaces are hyperquadratic.
  The geometry form can be
  - Hyperplane, pairs of hyperplanes, hypersphere, hyperellipsoid, hyperparaboloids, hyperhyperboloids



EE6887-Cha 3-9

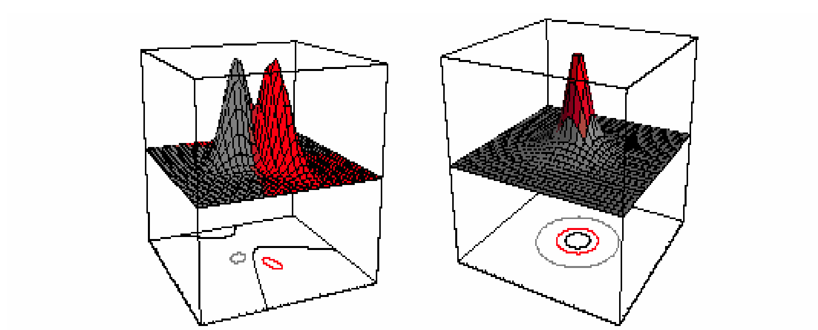
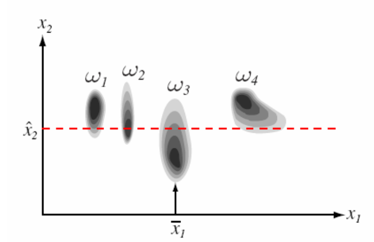


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

EE6887-Chang 3-10

Missing Features

- Bayesian Classifier
  - Decide ω_1 *if* $P(\omega_1 / x) > P(\omega_2 / x)$
  - But what if x is not completely observable?
  - If x_1 is missing, what's the most intuitive classification of ω if equal priors?



mean of x_1

X_1 missing

Missing Features (Cont.)

- Let $x=[x_g, x_b]$, x_g: good features, x_b: bad features

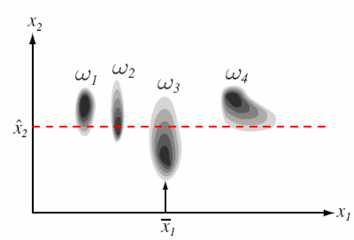
$$\text{compute } P(w_i \mid x_g) = \frac{p(w_i, x_g)}{p(x_g)} = \frac{\int p(w_i, x_g, x_b)dx_b}{p(x_g)}$$

$$= \frac{\int p(x_g, x_b \mid w_i)p(w_i)dx_b}{\int p(x_g, x_b)dx_b}$$

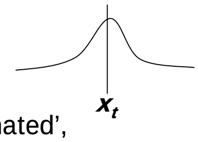
Joint prob. of (ω_i, x_g, x_b) marginalized over x_b

i.e. integrated over the x_b dimension.

Q: which class on the right should we choose?



Noisy Feature



- Let $x=[x_g, x_b]$, x_g: good features, x_b : bad features
- $x_b = x_t$ + noise, x_t is the original feature being 'contaminated', x_t is 'hidden'
- Given good features x_g and noisy feature x_b, find the right class ω_i .

$$P(w_i \mid x_g, x_b) = \frac{p(w_i, x_g, x_b)}{p(x_g, x_b)} = \frac{\int p(w_i, x_g, x_b, x_t)dx_t}{\int p(x_g, x_b, x_t)dx_t}$$

marginalized over x_t

$$= \frac{\int p(x_g, x_b, x_t \mid w_i)p(w_i)dx_t}{\int p(x_g, x_b, x_t)dx_t} = \frac{\int p(x_g, x_t \mid w_i)p(x_b \mid x_g, x_t, w_i)p(w_i)dx_t}{\int p(x_b \mid x_g, x_t)p(x_g, x_t)dx_t}$$

$$= \frac{\int p(x_g, x_t \mid w_i)p(x_b \mid x_t)p(w_i)dx_t}{\int p(x_b \mid x_t)p(x_g, x_t)dx_t}$$

need noise distribution $p(x_b \mid x_t)$, likelihood $p(x_g, x_t \mid w_i)$, prior $p(w_i)$

Joint prob. of (ω_i, x_g, x_b, x_t) marginalized over x_t