EE 6885 Statistical Pattern Recognition

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

Lecture 3 (9/14/05)

- Reading
  - Bayesian Decision Theory
    - Textbook DHS Chapter 2.5 – 2.6
- Homework #1 due 2005-09-21, Wed
- TA office hours
  - Tuesdays 9:20-10:50am, CEPSR Rm 712
- Basic Matlab sample code (course web site)

Minimal Risk Decision Rule

- Decide ω_1, if $R(\alpha_1 / x) < R(\alpha_2 / x)$
- *i.e.* Decide ω_1, if $\dfrac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \dfrac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \dfrac{P(\omega_2)}{P(\omega_1)}$
- If $\lambda_{12} = \lambda_{21} = 1$, $\lambda_{11} = \lambda_{22} = 0$, *i.e., loss = 'classification error'*

$$\text{Decide } \omega_1, \text{ if } \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}, \text{ i.e. } P(\omega_1 \mid x) > P(\omega_2 \mid x)$$

Max A Posterioi (MAP) classifier

- If $P(\omega_1) = P(\omega_2)$,

ML classifier

$$\text{Decide } \omega_1, \text{ if } \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > 1, \text{ i.e. } P(x \mid \omega_1) > P(x \mid \omega_2)$$

(Likelihood Ratio)

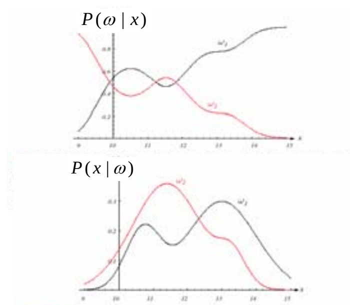


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

EE6887-Chang

3-3

Discriminant Function and Decision Regions

- Decide ω_1, if $g_i(x) > g_j(x) \quad \forall j \neq i$
  - E.g., minimal risk classifier $g_i(x) = -R(\alpha_i \mid x)$

    e.g., $g_i(x) = P(\omega_i \mid x) \qquad g_i(x) = P(x \mid \omega_i) P(\omega_i)$

    $g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$
- Decision Region, R_i

  x is in R_i, if $g_i(x) > g_j(x) \ \forall j \neq i$
- Two-Category case

  Decide ω_1, if $g(x) = g_1(x) - g_2(x) > 0$

$$g(x) = \ln \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

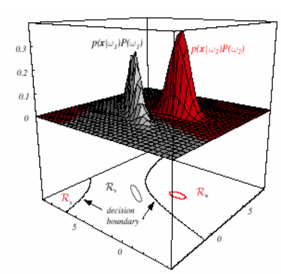


FIGURE 2.6. In this two-dimensional two-category classifier, the probability are Gaussian, the decision boundary consists of two hyperbolas, and thus the region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is 1 that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

EE6887-Chang

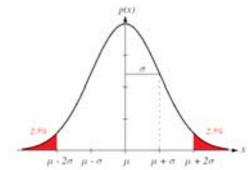
3-4

Gaussian Distribution

- Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)}$$

- Given the same mean and variance, Gaussian has the max entropy
- Sum of a large number of small, independent random variables approaches Gaussian



$\Pr[|x-\mu| \le \sigma] \cong 0.68$

$\Pr[|x-\mu| \le 2\sigma] \cong 0.95$

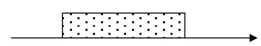
$\Pr[|x-\mu| \le 3\sigma] \cong 0.997$

Mahalanobis distance from x to μ

$$r = |x-\mu| / \sigma$$

$$\boxed{\begin{array}{l} \textit{Entropy of Gaussian}: \\ H_{gau} = 0.5 + \log_2(\sqrt{2\pi}\sigma) \end{array}}$$

Comparison w. uniform



Multi-variate Gaussian

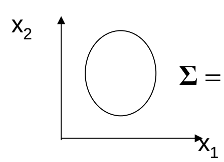
- Multivariate Gaussian, $N(\mu, \Sigma)$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma}|}} e^{\left(\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$$

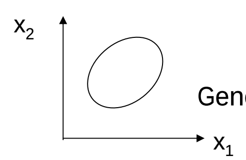
where $\mathbf{x}, \boldsymbol{\mu}$ are D-dimensional vectors

$\mathbf{\Sigma} : D \times D$ matrix

$|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$



$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$



General $\mathbf{\Sigma}$

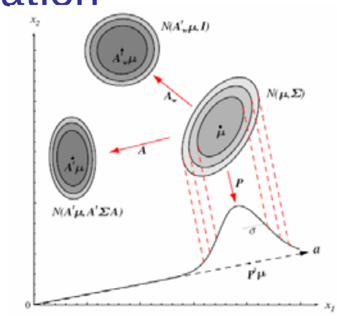
$$(\sigma_{ij})^2 = \Sigma(i,j) = \mathrm{cov}(x(i), x(j))$$
$$= E[(x(i) - E(x(i))(x(j) - E(x(j))]$$

Effect of Linear Transformation

- Linear transformation of Gaussian

$$y = A^t x \qquad y : k \times 1,\ A : d \times k,\ x : d \times 1$$

$$y \sim N(A^t \mu, A^t \Sigma A)$$



- Whitening transform

$$\Sigma = \Phi \Lambda \Phi^t \quad \text{(SVD, Eigenvectors)}$$

$\Phi : [\phi_1 \mid \phi_2 \mid \ldots \mid \phi_d]$ columns are orthogonal ev $\quad$ Λ: diag. matrix of eigenvalues

$$A^t \Sigma A = A^t \Phi \Lambda \Phi^t A = I$$

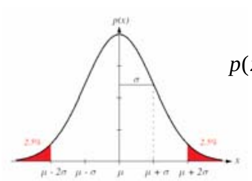
Whitening Trans. $\quad A_w = \Phi \Lambda^{-1/2}$

also PCA Transform

$$y = A_w{}^t x \quad \sim \quad N(A_w{}^t \mu,\ I)$$

Mahalanobis Distance

- Mahalanobis dist in 1-D



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)}$$

Mahalanobis distance from x to μ

$$r = |x-\mu| / \sigma$$

$$\Pr[|x-\mu| \le \sigma] = \Pr[r \le 1] \cong 0.68$$

$$\Pr[|x-\mu| \le 2\sigma] = \Pr[r \le 2] \cong 0.95$$

$$\Pr[|x-\mu| \le 3\sigma] = \Pr[r \le 3] \cong 0.997$$

- Multi-Dimensional case

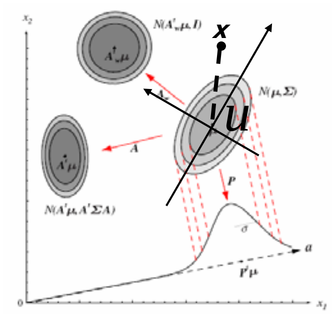
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma}|}} \exp\left(\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\, \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)$$

$$= \frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma}|}} \exp\left(\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\, \Phi\Lambda^{-1}\Phi^t (\mathbf{x}-\boldsymbol{\mu})\right)$$

$$= \frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma}|}} \exp\left(\frac{-1}{2}\underbrace{(A_w{}^t(\mathbf{x}-\boldsymbol{\mu}))^t (A_w{}^t(\mathbf{x}-\boldsymbol{\mu}))}_{r^2}\right)$$

r is the Mahalanobis distance

- Malalanobis distance from point x to the mean of a Gaussian



EE6887-Chang 3-9

Discriminant Function for Gaussians

- If consider classification error as "loss"

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

- For Gaussian

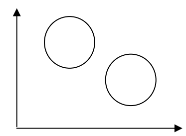
$$P(x \mid \omega_i) = N(\mu_i, \Sigma_i)$$

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

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

EE6887-Chang 3-10

Case I $\Sigma_i = \sigma^2 I$



- Statistical indep. features
  circles of same size

$$|\Sigma| = \sigma^{2d} \qquad \Sigma^{-1} = \sigma^{-2} I$$

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

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

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

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

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

Decision Boundaries

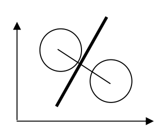
- Decision boundary between class i and j

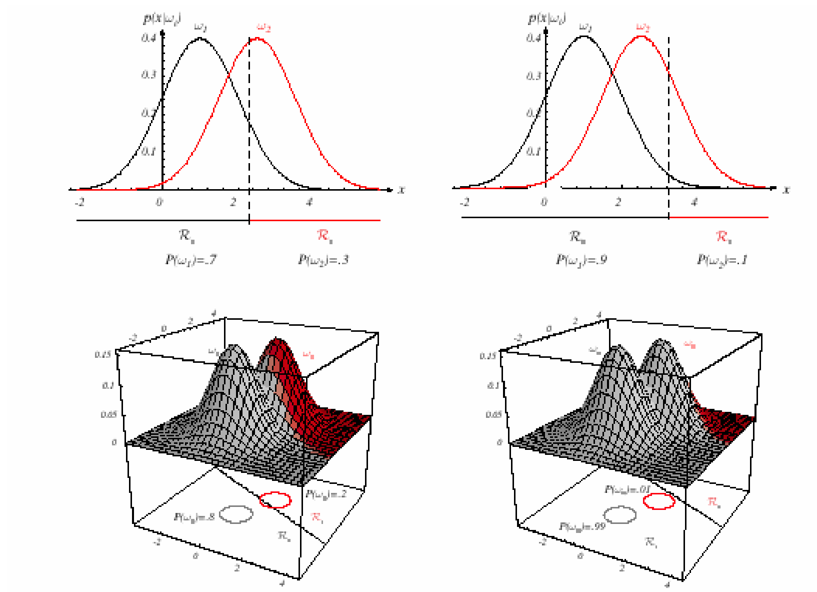
$$\boxed{g_i(x) = \frac{1}{\sigma^2}\mu_i^t x + w_{i0}} \qquad g_i(x) = g_j(x)$$

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

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

- A hyperplane perpendicular to $\mu_i - \mu_j$
  passing through point x_0
- If $\mathrm{P}(\omega_i) = \mathrm{P}(\omega_j)$ then $x_0 = \frac{1}{2}(\mu_i + \mu_j)$
  i.e., midpoint between μ_i and μ_j





EE6887-Chang 3-13

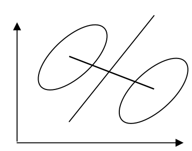
Case 2 $\Sigma_i = \Sigma$

- Hyperellipsoidal shapes of equal size and shape

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

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

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



- 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

EE6887-Chang 3-14

Case 2 decision boundaries

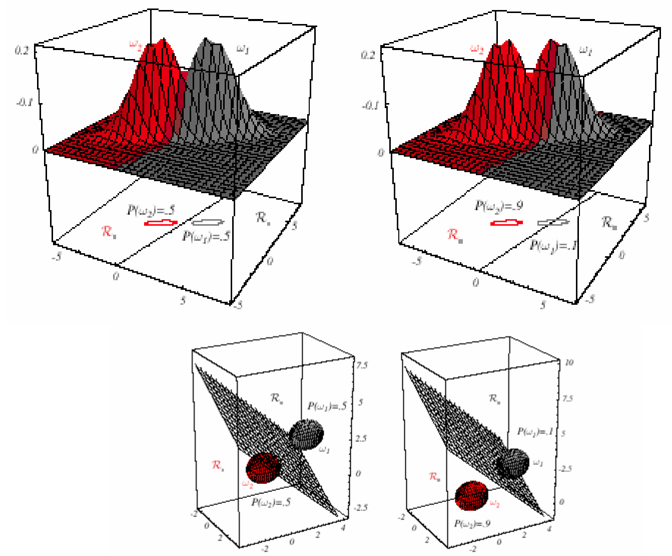


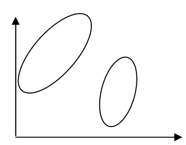
FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

Case 3 $\Sigma_i =$ arbitrary

- Discriminant functions are quadratic

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

where:

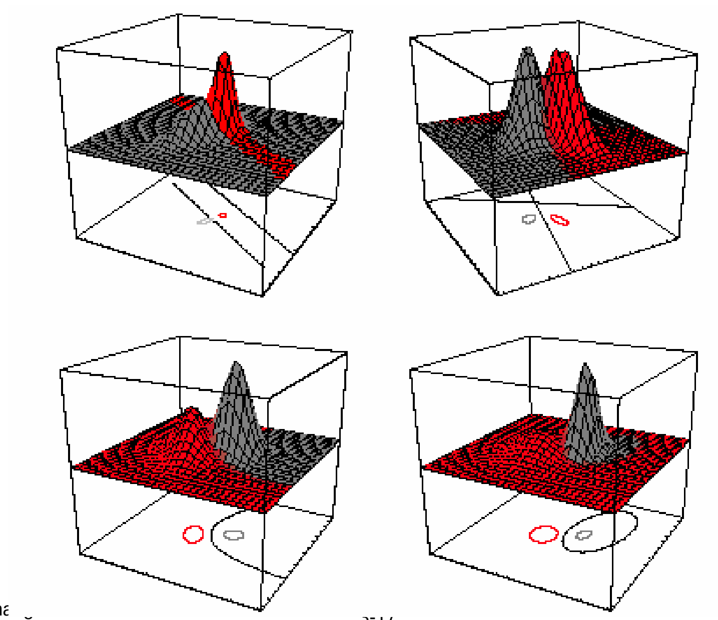


$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

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

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



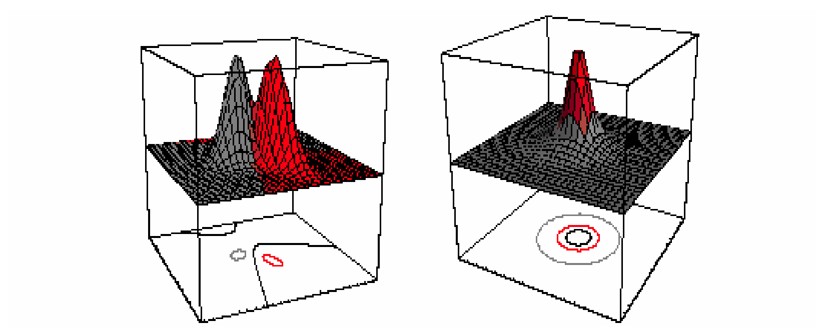


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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