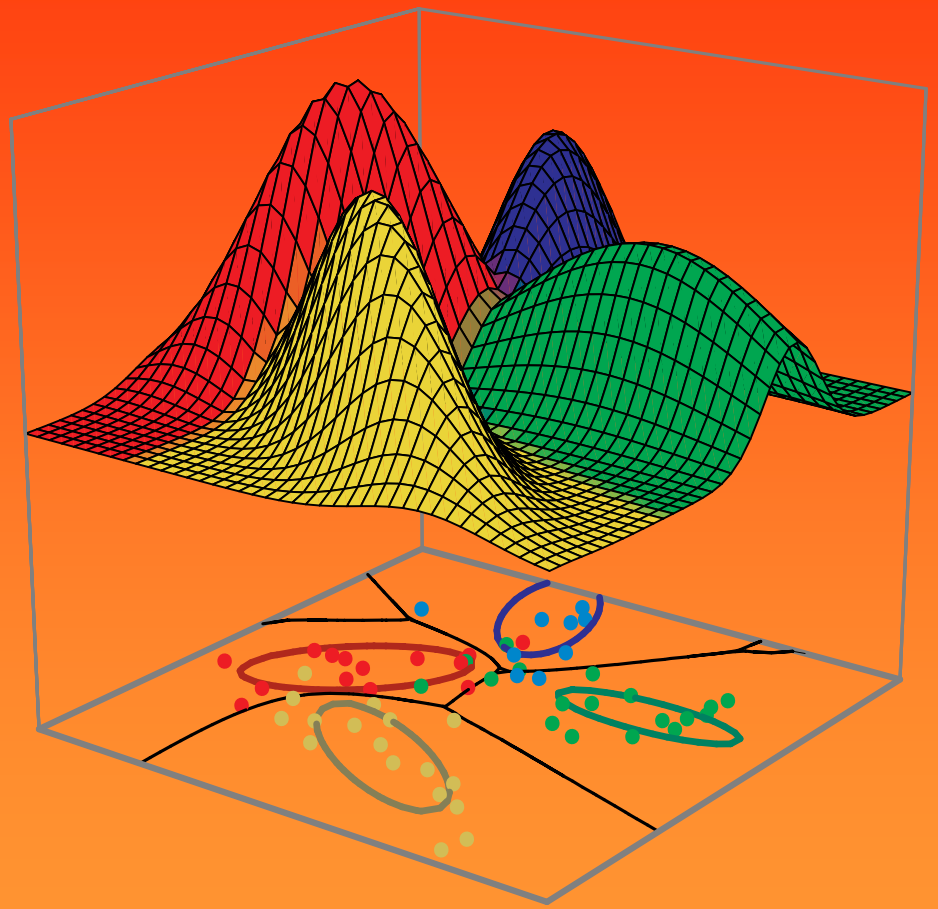

Pattern Classification

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 2 (Part 2): Bayesian Decision Theory (Sections 2.3-2.5)

- Minimum-Error-Rate Classification
- Classifiers, Discriminant Functions and Decision Surfaces
- The Normal Density

Minimum-Error-Rate Classification

- Actions are decisions on classes

  If action α_i is taken and the true state of nature is ω_j then:

  the decision is correct if $i = j$ and in error if $i \neq j$

- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

- Introduction of the zero-one loss function:

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

Therefore, the conditional risk is:

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

$$= \sum_{j \neq 1} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

"The risk corresponding to this loss function is the average probability error"

- The Bayes decision rule depends on minimizing risk
- Minimizing the risk requires selecting the i that maximizes $P(\omega_i \mid x)$
  (since $R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$)
- For Minimum error rate
  - Decide ω_i *if* $P(\omega_i \mid x) > P(\omega_j \mid x) \ \forall j \neq i$

- Regions of decision and zero-one loss function, therefore (using the likelihood ratio formula:

$$\textbf{Let } \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\cdot\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda \textbf{ then decide } \omega_1 \textbf{ if : } \frac{P(x\mid\omega_1)}{P(x\mid\omega_2)}>\theta_\lambda$$

- If λ is the zero-one loss function which means:

$$\lambda=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$$

$$\textbf{then } \theta_\lambda=\frac{P(\omega_2)}{P(\omega_1)}=\theta_a$$

$$\textbf{if } \lambda=\begin{pmatrix}0 & 2\\ 1 & 0\end{pmatrix} \textbf{ then } \theta_\lambda=\frac{2P(\omega_2)}{P(\omega_1)}=\theta_b$$

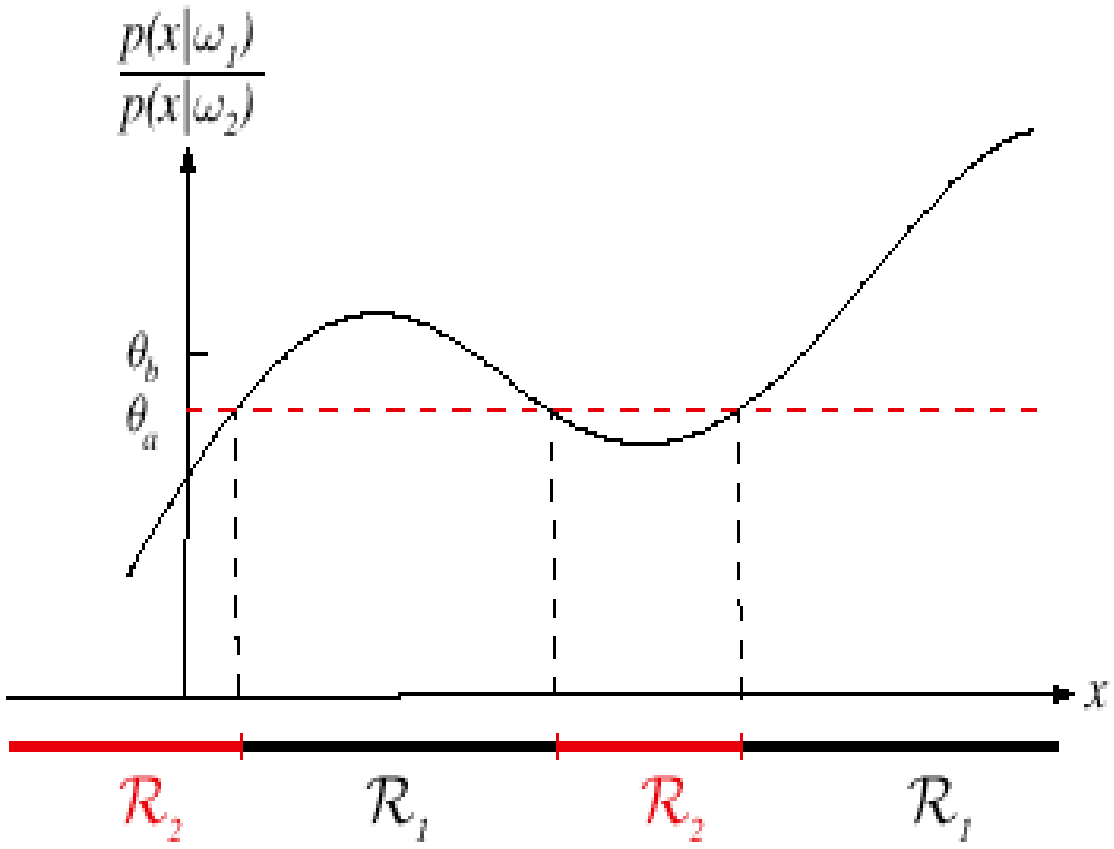

FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classifiers, Discriminant Functions and Decision Surfaces

- The multi-category case
  - Set of discriminant functions $g_i(x)$, $i = 1,\ldots, c$
  - The classifier assigns a feature vector x to class ω_i if:

$$g_i(x) > g_j(x) \ \forall j \neq i$$

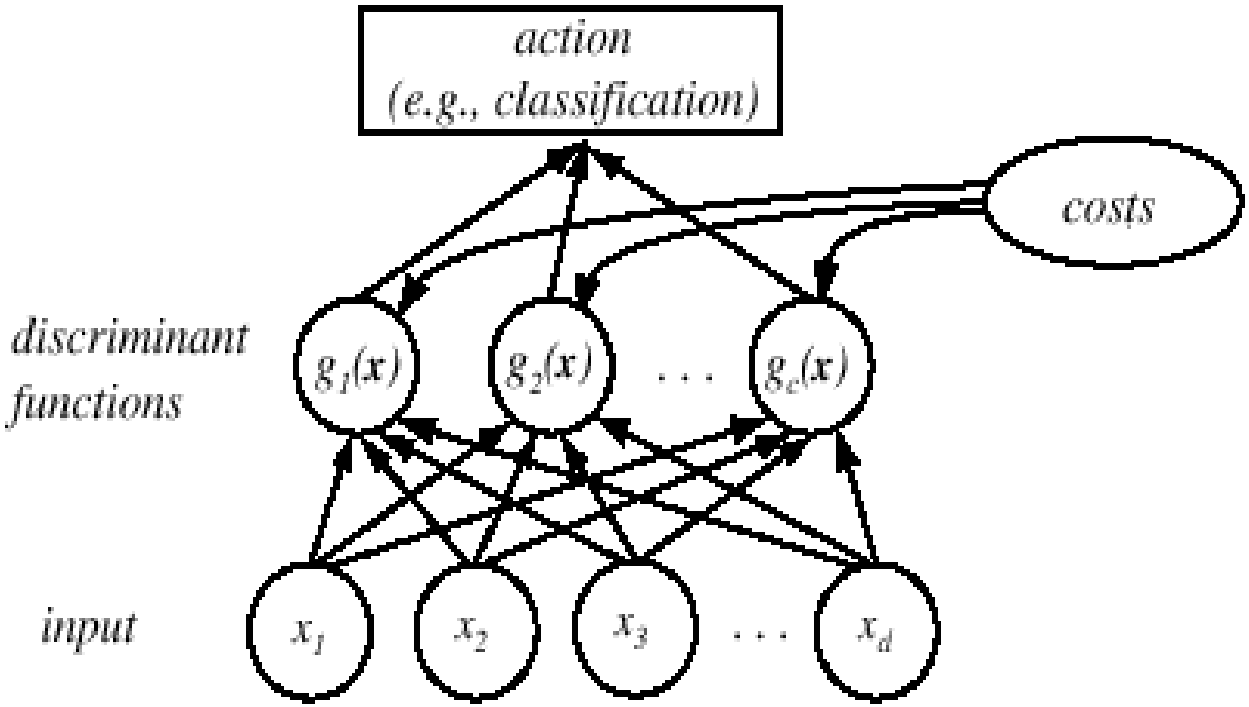

FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- *Let* $g_i(x) = - R(\alpha_i \mid x)$
  (max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take

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

  (max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv P(x \mid \omega_i)\, P(\omega_i)$$

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

(ln: natural logarithm!)

- Feature space divided into c decision regions

  $$\text{if } g_i(x) > g_j(x) \ \forall j \neq i \text{ then } x \text{ is in } \mathcal{R}_i$$

  ($\mathcal{R}_i$ means assign x to ω_i)

- The two-category case
  - A classifier is a "dichotomizer" that has two discriminant functions g_1 and g_2

    Let $g(x) \equiv g_1(x) - g_2(x)$

    Decide ω_1 if $g(x) > 0$; Otherwise decide ω_2

- The computation of g(x)

$$g(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$$

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

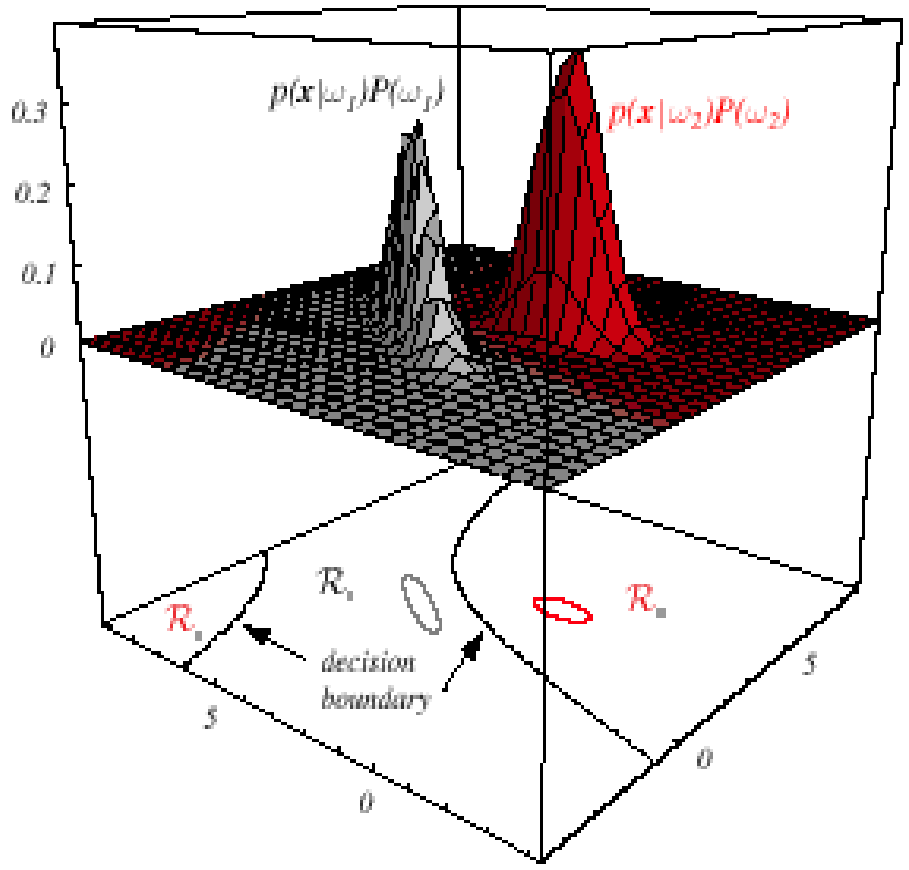

FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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

The Normal Density

- Univariate density
  - Density which is analytically tractable
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

Where:

μ = mean (or expected value) of x

σ^2 = expected squared deviation or variance

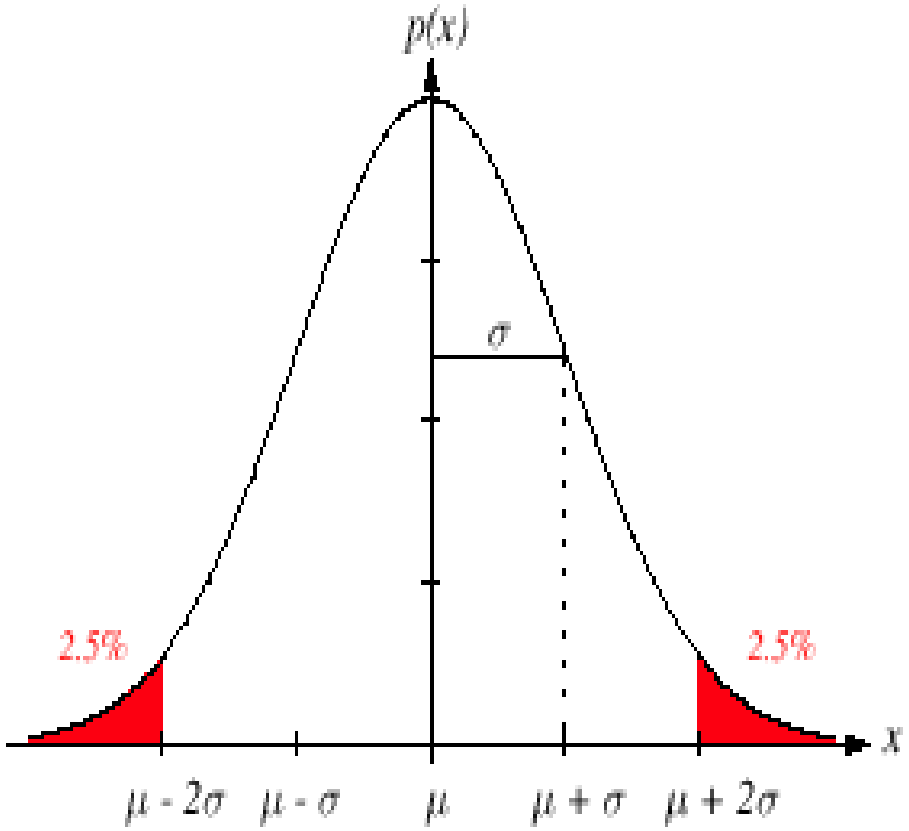

FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Multivariate density
  - Multivariate normal density in d dimensions is:

$$P(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right]$$

where:

$x = (x_1, x_2, ..., x_d)^t$ (t stands for the transpose vector form)

$\mu = (\mu_1, \mu_2, ..., \mu_d)^t$ mean vector

$\Sigma = d*d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively