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 a_i is taken and the true state of nature is w_j then: the decision is correct if $i = j$ and in error if $i \neq j$

  (i,j=1,...,c)

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

- Introduction of the zero-one loss (symmetrical) function:

$$l(a_i, w_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

  - Note: no loss assigned to a correct classification, while a loss of 1 is associated with misclassification (i.e. all errors are equally costly).
- Therefore, the conditional risk is

$$R(a_i \mid x) = \sum_{j=1}^{j=c} l(a_i \mid w_j) P(w_j \mid x)$$

$$= \sum_{j \neq i} P(w_j \mid x) = 1 - P(w_i \mid x)$$

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

- Minimize the risk requires maximize $P(w_i \mid x)$ (since $R(a_i \mid x) = 1 - P(w_i \mid x)$)

- For Minimum error rate

  - Decide w_i *if* $P(w_i \mid x) > P(w_j \mid x)$ $"j \neq I$

- Conclusion: to minimize the risk, select the class maximizing the posterior probability (as recommended by the Bayes decision rule)!

- Regions of decision and zero-one loss function, therefore:

$$\textbf{Let } \frac{l_{12} - l_{22}}{l_{21} - l_{11}} \cdot \frac{P(w_2)}{P(w_1)} = q_l \textbf{ then decide } w_1 \textbf{ if : } \frac{P(x|w_1)}{P(x|w_2)} > q_l$$

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

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

$$\textbf{then } q_l = \frac{P(w_2)}{P(w_1)} = q_a$$

$$\textbf{if } l = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \textbf{ then } q_l = \frac{2P(w_2)}{P(w_1)} = q_b$$

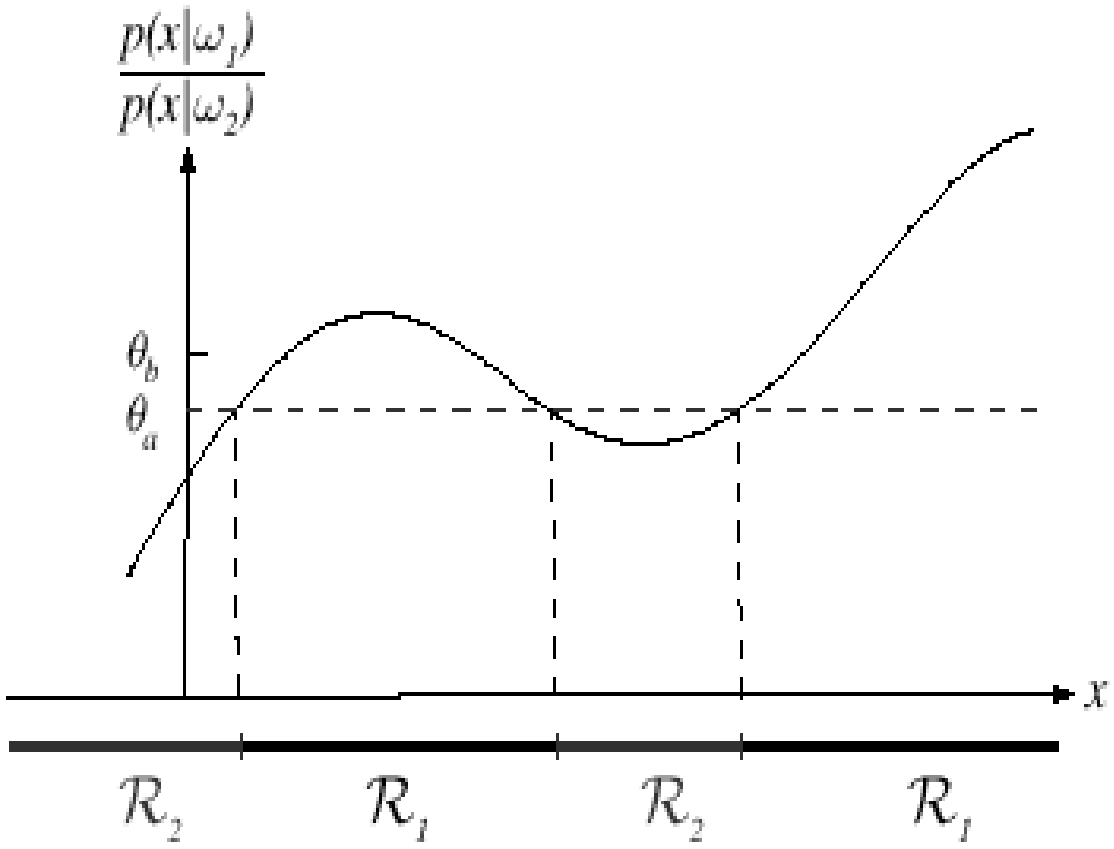

FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, 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

  - Bayesian decision rule recommends evaluating the potential of a class alternative through the value of its posterior probability
  - More generally, a discriminant function is a function employed for differentiating (or discriminating) between classes.
  - Set of discriminant functions $g_i(x), i = 1,..., c$

  - The classifier assigns a feature vector x to class w_i if:

$$g_i(x) > g_j(x) \quad "j \neq i$$

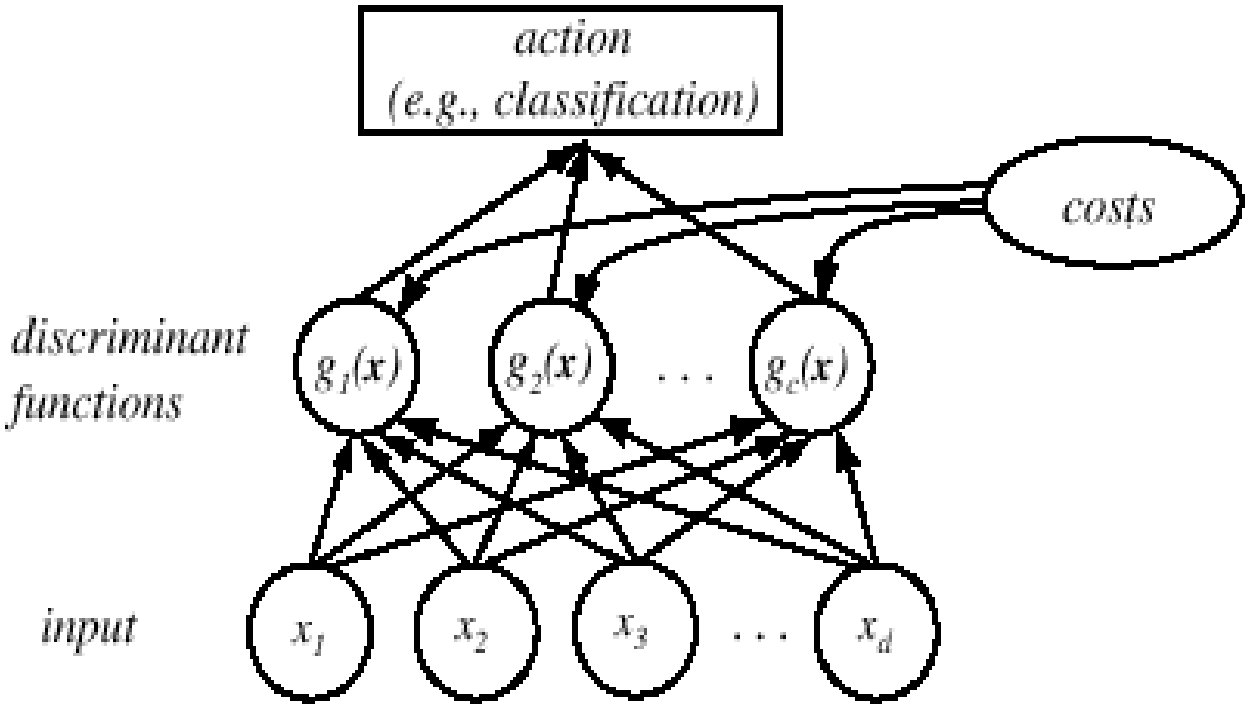

FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction 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.

Discriminant Functions

- *Let* $g_i(x) = - R(a_i \mid x)$
  (max. discriminant corresponds to min. risk!)
- For the minimum error rate, we take

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

(max. discrimination corresponds to max. posterior!)

Or $g_i(x) \propto P(x \mid w_i)\, P(w_i)$

Note: the probability of evidence, $P(x)$, is a normalization factor that can be ignored (since it is the same for all class alternatives).

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

(ln: natural logarithm!)

- Feature space divided into c decision regions

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

(R_i means assign x to w_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 w_1 if $g(x) > 0$; Otherwise decide w_2

- The computation of g(x) using posterior probability

$$g(x) = P(w_1 \mid x) - P(w_2 \mid x)$$

$$\equiv \ln \frac{P(x \mid w_1)}{P(x \mid w_2)} + \ln \frac{P(w_1)}{P(w_2)}$$

- Each discriminant function generates a set of decision regions, $R_1,R_2,\ldots,R_n$ (not needing to be contiguous). Decision regions are separated by decision boundaries. The condition satisfied at the decision boundary between R_k and R_j is: $g_k(X)=g_j(X)$

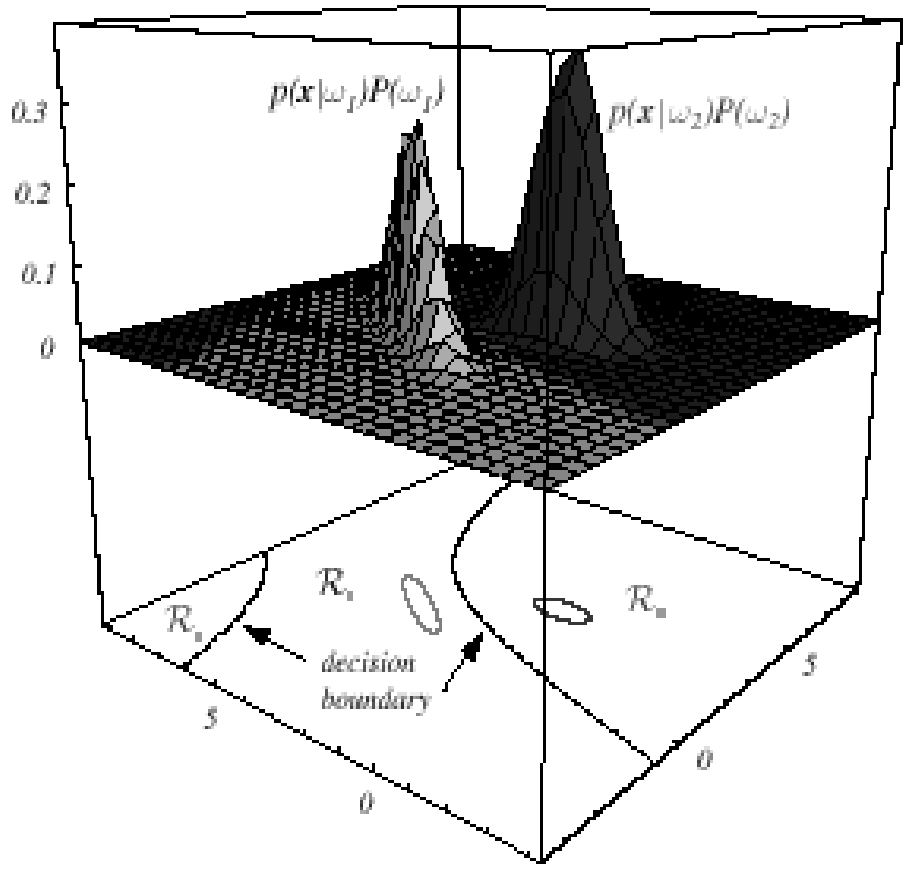

FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. 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

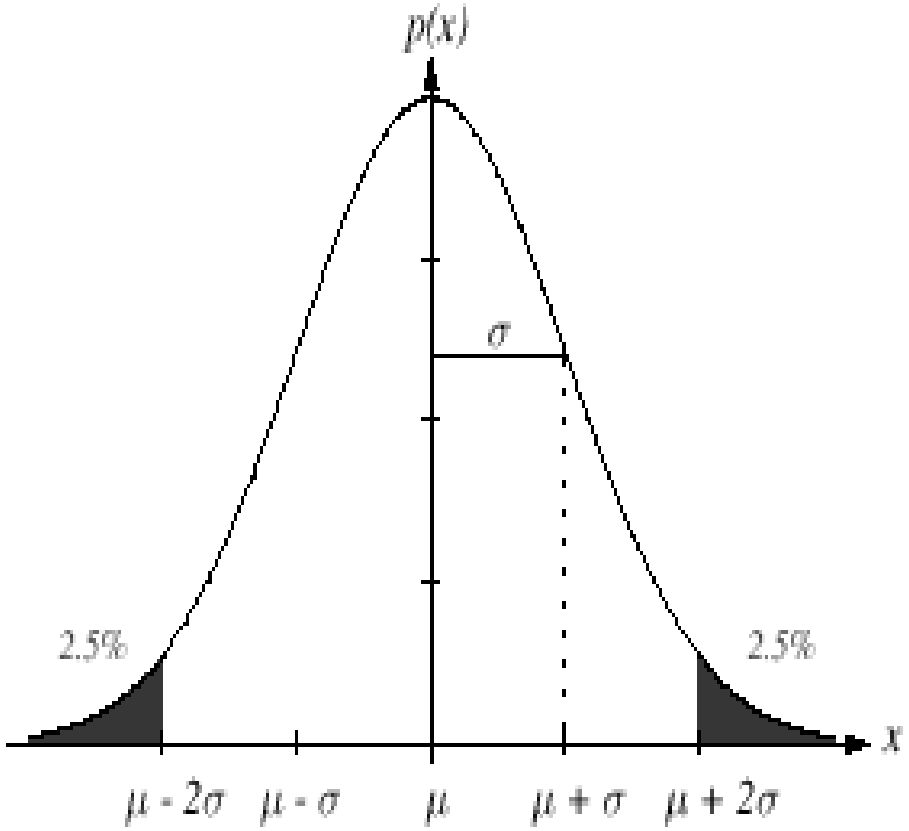

FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

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

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

where:

$x = (x_1, x_2, \ldots, x_d)^T$ (T stands for the transpose vector form)

$m = (m_1, m_2, \ldots, m_d)^T$ mean vector

$S = d*d$ covariance matrix

$|S|$ and S^{-1} are determinant and inverse respectively

Why the normal distribution is important

- If the likelihood probabilities are normally distributed, then a number of simplification can be made:
  - Most important: discriminant functions can be simplified.
  - The decision boundaries will have shapes and positions depending upon the prior probabilities, the means and the covariances of the distributions in questions.
- Three distinct cases will be studied:
  - Features are statistically independent and each feature has the same variance
  - Covariance matrices are arbitrary, but equal to each other for all classes
  - Covariance matrices are arbitrary, and different for each class