4. Bayesian Decision Theory and Bayesian Classifier

Dr R. K. Chaurasiya

Dept. Electronics and Communication Engineering

MANIT Bhopal

Bayesian Decision Theory

- Fundamental statistical approach

- It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known.

Bayesian Decision Theory

- Fundamental statistical approach

- It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known.

- The Basic Idea
  - To minimize errors, choose the least risky class, i.e. the class for which the *expected loss* is smallest

Example: Salmon or Sea bass?

Prior

- State of nature: Class of the fish (ω_1 *OR* ω_2)
  - Because the state of nature is so unpredictable, we consider w to be a variable that must be described probabilistically.
  - State of nature is a random variable

Prior

- State of nature: Class of the fish (ω_1 *OR* ω_2)
  - Because the state of nature is so unpredictable, we consider w to be a variable that must be described probabilistically.
  - State of nature is a random variable
- $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustively)
- priori probability (or simply prior) $P(\omega_1)$ that the next fish is sea bass, and some prior probability $P(\omega_2)$ that it is salmon.

Prior Probability

- The *state of nature (true class for the current input), in the absence of other information*
- Informally, "what percentage of the time state X occurs"
- Example
  - The prior probability that an instance taken from two classes is provided as input, in the absence of any features (e.g. P(cat) = 0.3, P(dog) = 0.7)

- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2
  - *Is it a good classifier? Accuracy?*
- Use of the class –conditional information
- $P(x \mid \omega_1)$ and $P(x \mid \omega_2)$ describe the difference in lightness between populations of sea and salmon

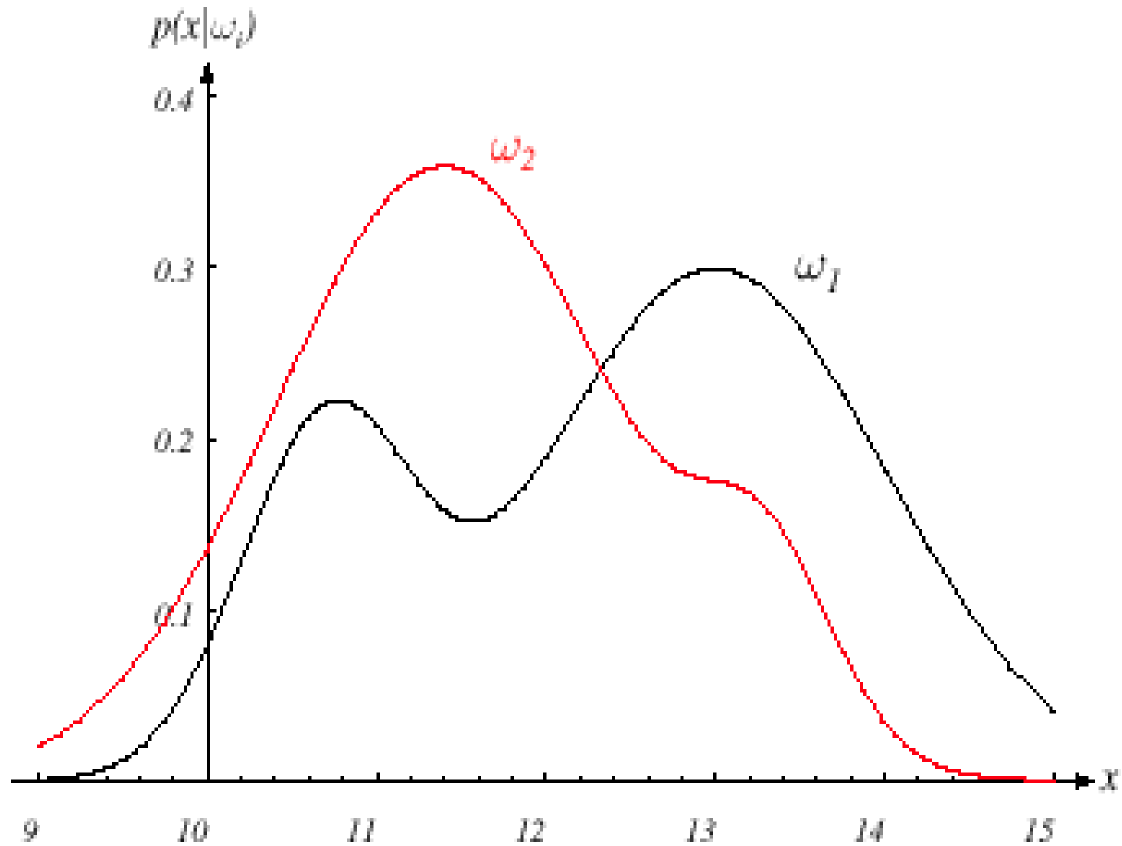


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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.

Bayes Theorem

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(w_j)}{p(x)}$$

$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{evidence}}$$

$$\text{where } \; p(x) = \sum_{j=1}^{c} p(x|\omega_j)P(\omega_j)$$

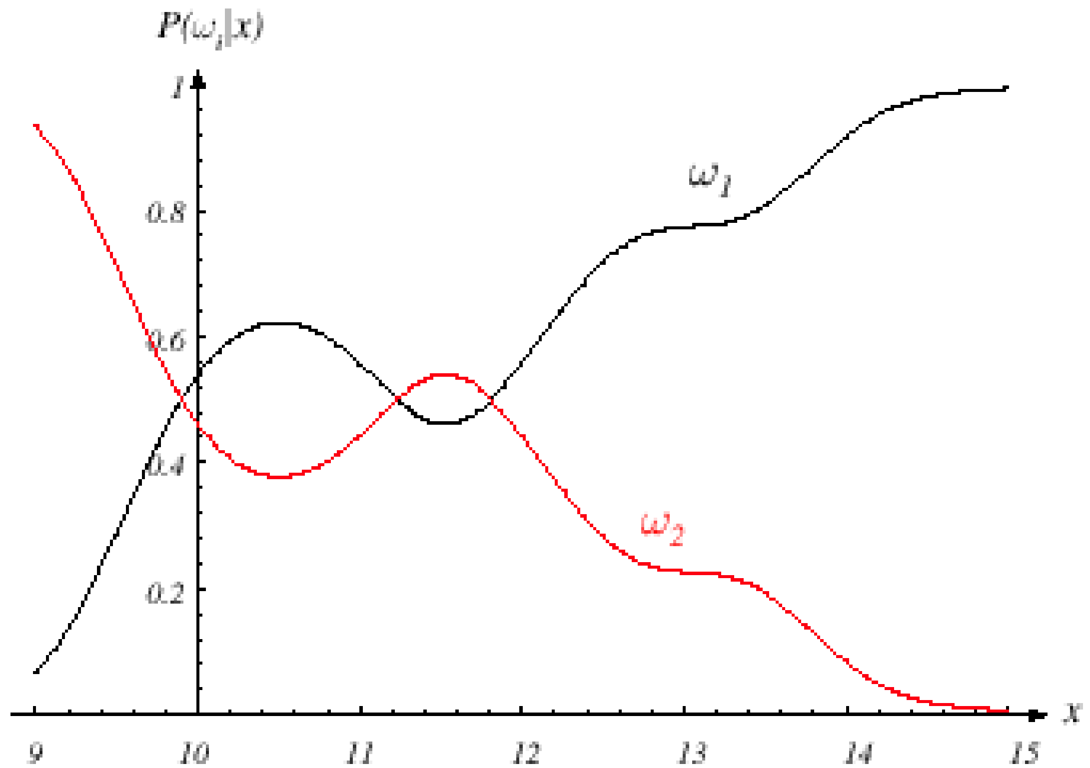


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Decision given the posterior probabilities

X is an observation for which:

if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ $\Rightarrow$ True state of nature $= \omega_1$

if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ $\Rightarrow$ True state of nature $= \omega_2$

- Decision given the posterior probabilities

X is an observation for which:

$$\text{if } P(\omega_1 \mid x) > P(\omega_2 \mid x) \Rightarrow \text{True state of nature} = \omega_1$$
$$\text{if } P(\omega_1 \mid x) < P(\omega_2 \mid x) \Rightarrow \text{True state of nature} = \omega_2$$

Therefore:

whenever we observe a particular x, the probability of error is :

$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$
$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{if we choose } \omega_2 \\ P(\omega_2|x) & \text{if we choose } \omega_1 \end{cases}$$

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)\,dx \quad \text{(average error)}$$

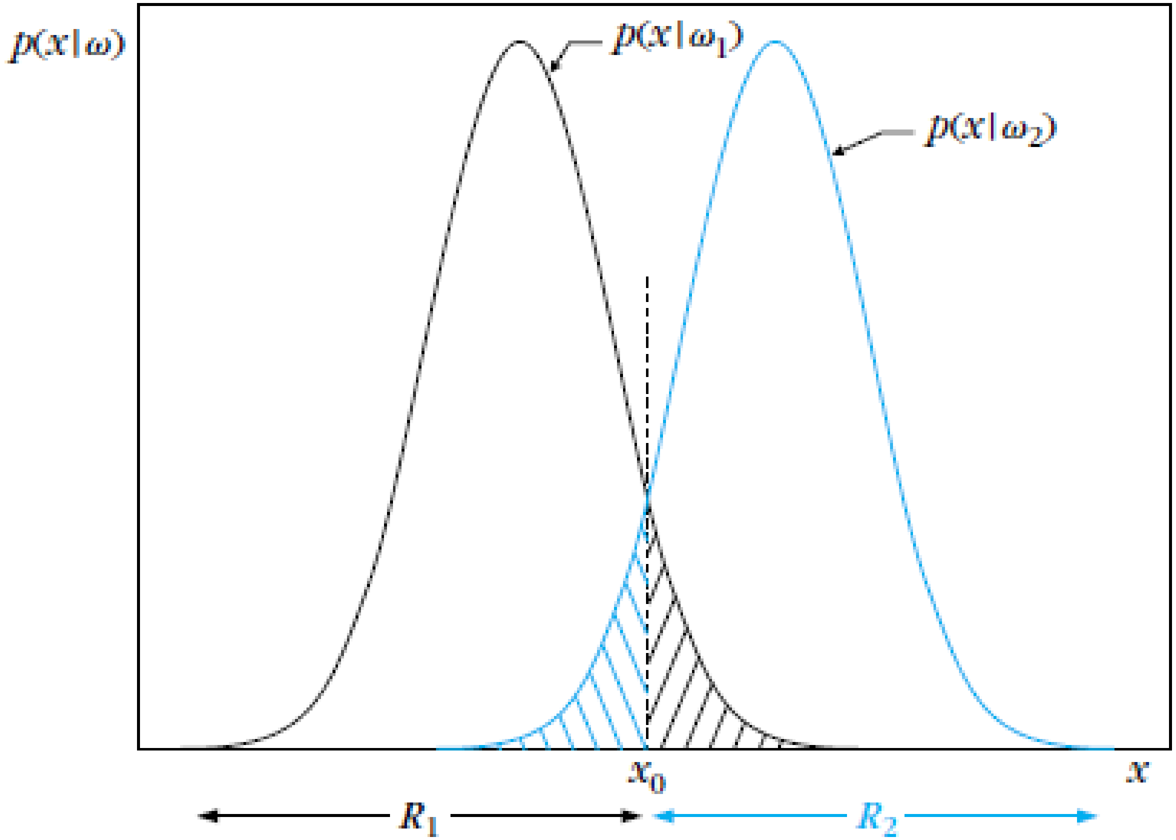
- Minimizing the probability of error
- Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2

  Therefore:

$$P(error \mid x) = \min\,[P(\omega_1 \mid x), P(\omega_2 \mid x)]$$

(Bayes decision)

Minimizing the Probability of Error: Example



$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)\,dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega_1)\,dx$$

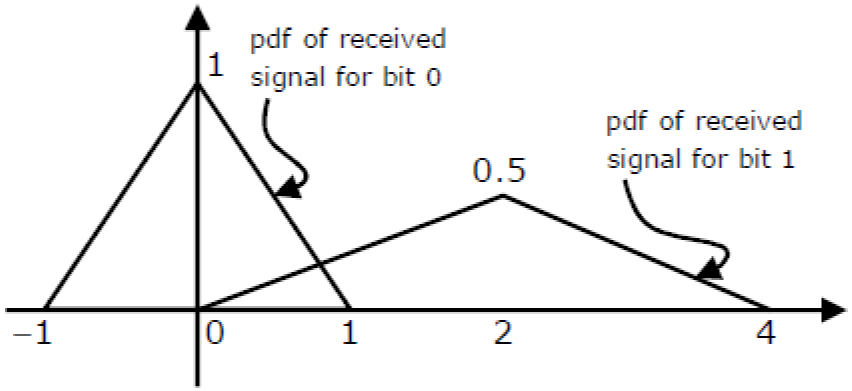
Minimizing the Probability of Error

$$P_e = P(x \in R_2 | \omega_1) P(\omega_1) + P(x \in R_1 | \omega_2) P(\omega_2)$$

$$= P(\omega_1) \int_{R_2} p(x|\omega_1)\, dx + P(\omega_2) \int_{R_1} p(x|\omega_2)\, dx$$

|EC-GATE-2013 PAPER| **Common Data Questions: 48 & 49**

Bits 1 and 0 are transmitted with equal probability. At the receiver, the pdf of the respective received signals for both bits are as shown below.



48. If the detection threshold is 1, the BER will be

(A) $\frac{1}{2}$ (B) $\frac{1}{4}$ (C) $\frac{1}{8}$ (D) $\frac{1}{16}$

49. The optimum threshold to achieve minimum bit error rate (BER) is

(A) $\frac{1}{2}$ (B) $\frac{4}{5}$ (C) 1 (D) $\frac{3}{2}$

Bayesian Decision Theory – Generalization of the preceding ideas

- Use of more than one feature ($\boldsymbol{x}$ instead of x)..!
- Use more than two states of nature..!
- Introduce a loss of function which is more general than the probability of error.
- The loss function states how costly each classification action taken is.

Concept of Risk…

- The classification error probability is not always the best criterion to be adopted for minimization.
- This is because it assigns the same importance to all errors.
- However, there are cases in which some wrong decisions may have more serious implications than others.
- In such cases, it is more appropriate to assign a penalty term to weigh each error.

Minimizing the Average Risk

- Let, R_1, R_2 be the regions in the feature space where we decide in favor of ω_1 and ω_2, respectively.
- The error probability P_e is given by

$$P_e = P(\omega_1) \int_{R_2} p(x|\omega_1)\, dx + P(\omega_2) \int_{R_1} p(x|\omega_2)\, dx$$

- Instead of selecting R_1 and R_2 so that P_e is minimized, we will now try to minimize a modified version of it, that is,

$$r = \lambda_{12} P(\omega_1) \int_{R_2} p(\boldsymbol{x}|\omega_1) d\boldsymbol{x} + \lambda_{21} P(\omega_2) \int_{R_1} p(\boldsymbol{x}|\omega_2) d\boldsymbol{x}$$

Minimizing the Average Risk

- $$r = \lambda_{12} P(\omega_1) \int_{R_2} p(\boldsymbol{x}|\omega_1) d\boldsymbol{x} + \lambda_{21} P(\omega_2) \int_{R_1} p(\boldsymbol{x}|\omega_2) d\boldsymbol{x}$$
- Here each of the two terms that contributes to the overall error probability is weighted according to its significance.
- A penalty term λ_{ij}, known as *loss*, is associated with the decision of misclassifying $\boldsymbol{x}$ from ω_i to class ω_j.
- In general, for a multi-class classification problem, we defibe a **Loss-matrix** L, which has at its (k, i) location the corresponding penalty term.

Minimizing the Average Risk in a 2-class classification problem.

- **General Loss-matrix for 2-class classification problem**

$$L = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$$

- The risk of misclassification of feature vector $\boldsymbol{x}$ to two different classes can be computed as

$$l_1 = \lambda_{11}\, p(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{21}\, p(\boldsymbol{x}|\omega_2)P(\omega_2)$$

$$l_2 = \lambda_{12}\, p(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{22}\, p(\boldsymbol{x}|\omega_2)P(\omega_2)$$

We assign $\boldsymbol{x}$ to ω_1 if $l_1 < l_2$, that is,

$$(\lambda_{21} - \lambda_{22})p(\boldsymbol{x}|\omega_2)P(\omega_2) < (\lambda_{12} - \lambda_{11})p(\boldsymbol{x}|\omega_1)P(\omega_1)$$

Minimizing the Average Risk in a 2-class classification problem.

- **Assign 0 penalty for correct decision.**

$$L = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$$

- The patterns will be assigned to class ω_2 if

$$l_2 < l_1$$

i.e.

$$p(\boldsymbol{x}|\omega_2) > p(\boldsymbol{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}}$$

2-class classification problem with 0-1 Loss Matrix

- **Assign 0 penalty for correct decision, 1 for incorrect decision (irrespective of the classes)**

$$L = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

- The patterns will be assigned to class ω_2 if

$$l_2 < l_1$$

i.e.

$$p(\boldsymbol{x}|\omega_2) > p(\boldsymbol{x}|\omega_1)$$

- **This is equivalent to minimizing the probability of error**

Numerical Example:

In a two-class problem with a single feature x the pdfs are Gaussians with variance $\sigma^2 = 1/2$ for both classes and mean values 0 and 1, respectively, that is,

$$p(x|\omega_1) = \frac{1}{\sqrt{\pi}} \exp(-x^2)$$

$$p(x|\omega_2) = \frac{1}{\sqrt{\pi}} \exp(-(x-1)^2)$$

If $P(\omega_1) = P(\omega_2) = 1/2$, compute the threshold value x_0 (a) for minimum error probability and (b) for minimum risk if the loss matrix is

$$L = \begin{bmatrix} 0 & 0.5 \\ 1.0 & 0 \end{bmatrix}$$

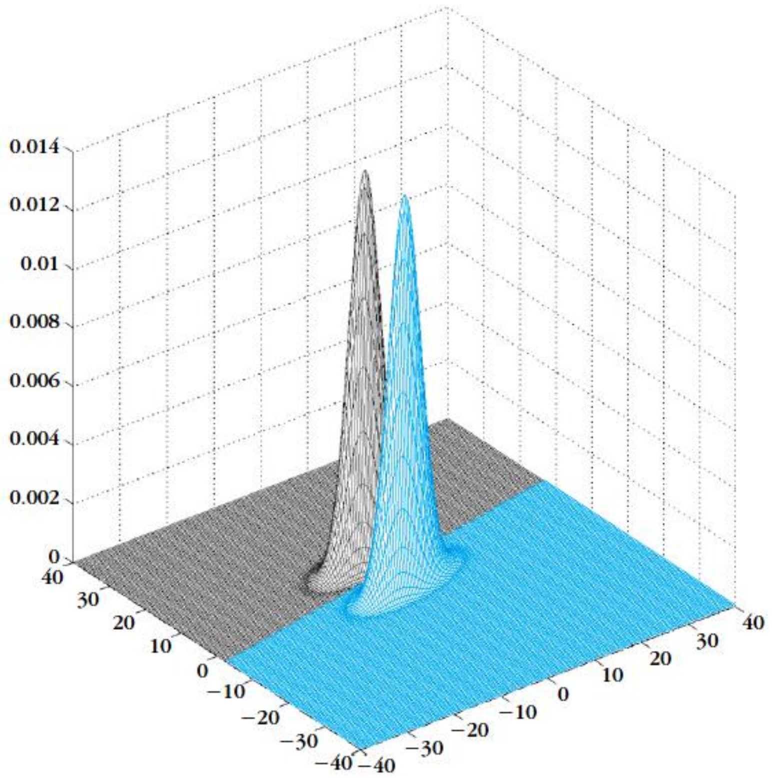
Solution (a)

Taking into account the shape of the Gaussian function graph , the threshold for the minimum probability case will be

$$x_0 : \exp(-x^2) = \exp(-(x-1)^2)$$

Taking the logarithm of both sides, we end up with $x_0 = 1/2$.

Visualization of a two-dimension equivalent case



Solution (b)

In the minimum risk case we get

$$x_0 : \exp(-x^2) = 2\exp(-(x-1)^2)$$

or $x_0 = (1 - \ln 2)/2 < 1/2$; that is, the threshold moves to the left of $1/2$. If the two classes are not equiprobable, then it is easily verified that if $P(\omega_1) > (<) P(\omega_2)$ the threshold moves to the right (left). That is, we expand the region in which we decide in favor of the most probable class, since it is better to make fewer errors for the most probable class.

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 $\boldsymbol{\omega_i}$ if:

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

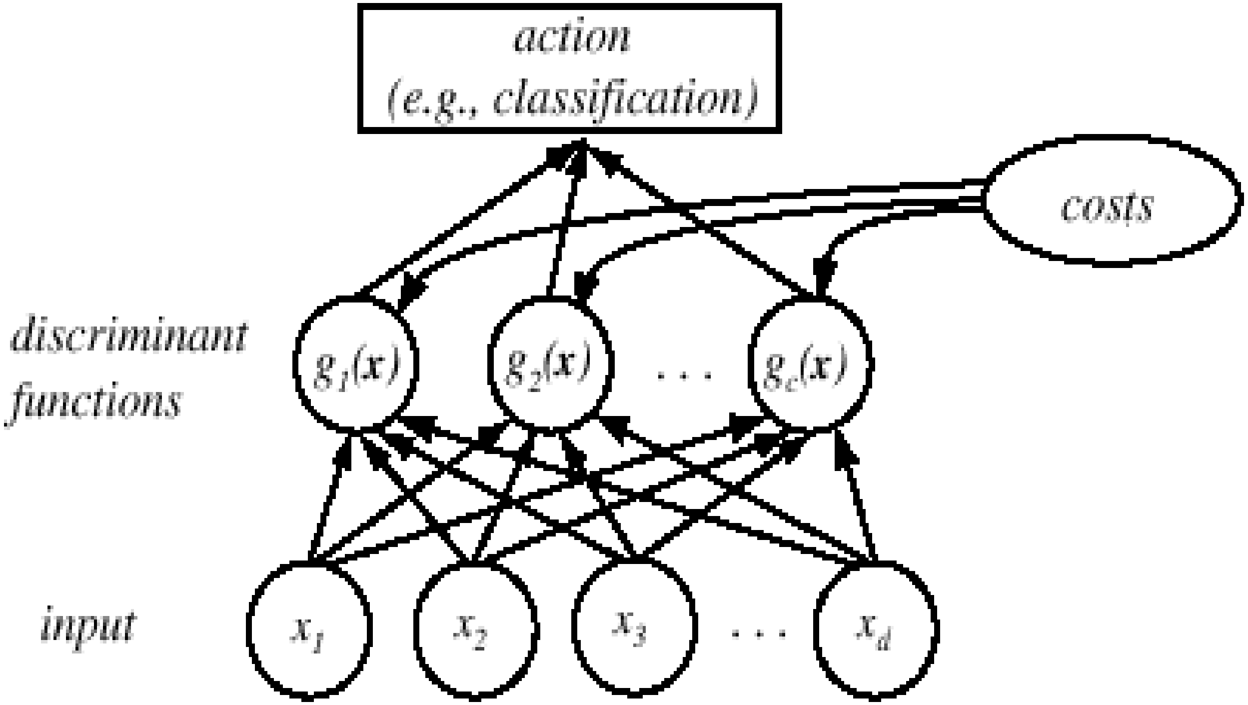


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction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

Example: Decision Regions for Binary Classifier