

# ***LECTURE 4: Bayesian Decision Theory***

---

- **The Likelihood Ratio Test**
- **The Probability of Error**
- **The Bayes Risk**
- **Bayes, MAP and ML Criteria**
- **Multi-class problems**
- **Discriminant Functions**



# Likelihood Ratio Test (LRT)

- Assume we are to classify an object based on the evidence provided by a measurement (or feature vector)  $x$
- Would you agree that a reasonable decision rule would be the following?
  - "Choose the class that is most 'probable' given the observed feature vector  $x$ "
    - More formally: Evaluate the posterior probability of each class  $P(\omega_i|x)$  and choose the class with largest  $P(\omega_i|x)$
- Let us examine the implications of this decision rule for a 2-class problem
  - In this case the decision rule becomes

if  $P(\omega_1 | x) > P(\omega_2 | x)$  choose  $\omega_1$   
else choose  $\omega_2$

- Or, in a more compact form

$$P(\omega_1 | x) \underset{\omega_2}{\overset{\omega_1}{\geq}} P(\omega_2 | x)$$

- Applying Bayes Rule

$$\frac{P(x | \omega_1)P(\omega_1)}{P(x)} \underset{\omega_2}{\overset{\omega_1}{\geq}} \frac{P(x | \omega_2)P(\omega_2)}{P(x)}$$

- $P(x)$  does not affect the decision rule so it can be eliminated\*. Rearranging the previous expression

$$\Lambda(x) = \frac{P(x | \omega_1)}{P(x | \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\geq}} \frac{P(\omega_2)}{P(\omega_1)}$$

- The term  $\Lambda(x)$  is called the **likelihood ratio**, and the decision rule is known as the **likelihood ratio test**

*\* $P(x)$  can be disregarded in the decision rule since it is constant regardless of class  $\omega$ . However,  $P(x)$  will be needed if we want to estimate the posterior  $P(\omega|x)$  which, unlike  $P(x|\omega)P(x)$ , is a true probability value and, therefore, gives us an estimate of the "goodness" of our decision.*



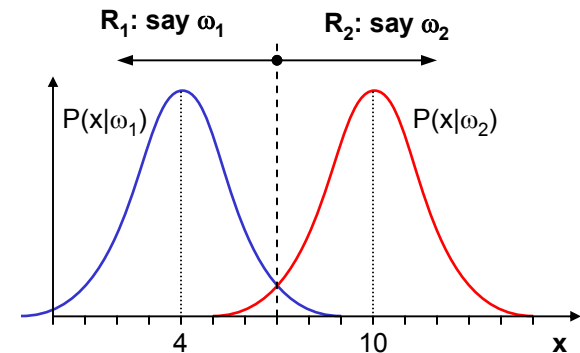
# Likelihood Ratio Test: an example

- Given a classification problem with the following class conditional densities, derive a decision rule based on the Likelihood Ratio Test (assume equal priors)

$$P(x | \omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2} \quad P(x | \omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}$$

## Solution

- Substituting the given likelihoods and priors into the LRT expression:  $\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}} > \frac{\omega_1}{\omega_2}$
- Simplifying the LRT expression:  $\Lambda(x) = \frac{e^{-\frac{1}{2}(x-4)^2}}{e^{-\frac{1}{2}(x-10)^2}} > 1$
- Changing signs and taking logs:  $(x-4)^2 - (x-10)^2 < 0$
- Which yields:  $x < 7$
- This LRT result makes sense from an intuitive point of view since the likelihoods are identical and differ only in their mean value



- How would the LRT decision rule change if, say, the priors were such that  $P(\omega_1)=2P(\omega_2)$  ?



# The probability of error (1)

- The performance of any decision rule can be measured by its probability of error  $P[\text{error}]$  which, making use of the Theorem of total probability (Lecture 2), can be broken up into

$$P[\text{error}] = \sum_{i=1}^C P[\text{error} | \omega_i] P[\omega_i]$$

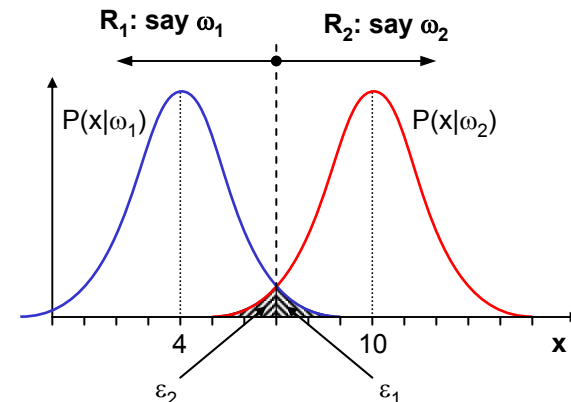
- The class conditional probability of error  $P[\text{error} | \omega_i]$  can be expressed as

$$P[\text{error} | \omega_i] = P[\text{choose } \omega_j | \omega_i] = \int_{R_j} P(x | \omega_i) dx$$

- So, for our 2-class problem, the probability of error becomes

$$P[\text{error}] = P[\omega_1] \underbrace{\int_{R_2} P(x | \omega_1) dx}_{\varepsilon_1} + P[\omega_2] \underbrace{\int_{R_1} P(x | \omega_2) dx}_{\varepsilon_2}$$

- where  $\varepsilon_i$  is the integral of the likelihood  $P(x | \omega_i)$  over the region  $R_j$  where we choose  $\omega_j$
- For the decision rule of the previous example, the integrals  $\varepsilon_1$  and  $\varepsilon_2$  are depicted below
  - Since we assumed equal priors, then  $P[\text{error}] = (\varepsilon_1 + \varepsilon_2)/2$



- Compute the probability for the example above



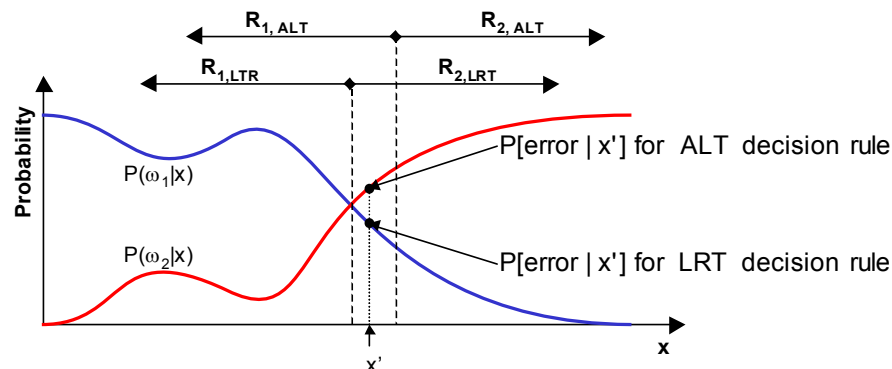
# The probability of error (2)

- Now that we can measure the performance of a decision rule we can ask the following question: How good is the Likelihood Ratio Test decision rule?

- For this purpose it is convenient to express  $P[\text{error}]$  in terms of the posterior  $P[\text{error}|x]$

$$P[\text{error}] = \int_{-\infty}^{+\infty} P[\text{error} | x] P(x) dx$$

- The optimal decision rule will minimize  $P[\text{error}|x]$  for every value of  $x$ , so that the integral above is minimized
- At each point  $x'$ ,  $P[\text{error}|x']$  is equal to  $P[\omega_i|x']$  when we choose the other class  $\omega_j$ 
  - This is depicted in the following figure:



- From the figure it becomes clear that, for any value of  $x'$ , the Likelihood Ratio Test decision rule will always have a lower  $P[\text{error}|x']$ 
  - Therefore, when we integrate over the real line, the LRT decision rule will yield a lower  $P[\text{error}]$

For any given problem, the minimum probability of error is achieved by the Likelihood Ratio Test decision rule. This probability of error is called the **Bayes Error Rate** and is the **BEST** any classifier can do.



# The Bayes Risk (1)

- So far we have assumed that the penalty of misclassifying a class  $\omega_1$  example as class  $\omega_2$  is the same as the reciprocal. In general, this is not the case:
  - For example, misclassifying a cancer sufferer as a healthy patient is a much more serious problem than the other way around
- This concept can be formalized in terms of a cost function  $C_{ij}$ 
  - $C_{ij}$  represents the cost of choosing class  $\omega_i$  when class  $\omega_j$  is the true class
- We define the Bayes Risk as the expected value of the cost

$$\mathfrak{R} = E[C] = \sum_{i=1}^2 \sum_{j=1}^2 C_{ij} \cdot P[\text{choose } \omega_i \text{ and } x \in \omega_j] = \sum_{i=1}^2 \sum_{j=1}^2 C_{ij} \cdot P[x \in R_i | \omega_j] \cdot P[\omega_j]$$

- What is the decision rule that minimizes the Bayes Risk?

- First notice that

$$P[x \in R_i | \omega_j] = \int_{R_i} P(x | \omega_j) dx$$

- We can express the Bayes Risk as

$$\begin{aligned} \mathfrak{R} = & \int_{R_1} [C_{11} \cdot P[\omega_1] \cdot P(x | \omega_1) + C_{12} \cdot P[\omega_2] \cdot P(x | \omega_2)] dx + \\ & \int_{R_2} [C_{21} \cdot P[\omega_1] \cdot P(x | \omega_1) + C_{22} \cdot P[\omega_2] \cdot P(x | \omega_2)] dx \end{aligned}$$

- Then we note that, for either likelihood, one can write:

$$\int_{R_1} P(x | \omega_i) dx + \int_{R_2} P(x | \omega_i) dx = \int_{R_1 \cup R_2} P(x | \omega_i) dx = 1$$



# The Bayes Risk (2)

- Merging the last equation into the Bayes Risk expression yields

$$\mathfrak{R} = \begin{array}{|l} \boxed{C_{11}P[\omega_1] \int_{R_1} P(x | \omega_1) dx} + \boxed{C_{12}P[\omega_2] \int_{R_1} P(x | \omega_2) dx} + \\ \boxed{C_{21}P[\omega_1] \int_{R_2} P(x | \omega_1) dx} + \boxed{C_{22}P[\omega_2] \int_{R_2} P(x | \omega_2) dx} + \\ \boxed{C_{21}P[\omega_1] \int_{R_1} P(x | \omega_1) dx} + \boxed{C_{22}P[\omega_2] \int_{R_1} P(x | \omega_2) dx} + \\ \boxed{-C_{21}P[\omega_1] \int_{R_1} P(x | \omega_1) dx} - \boxed{C_{22}P[\omega_2] \int_{R_1} P(x | \omega_2) dx} \end{array}$$

- Now we cancel out all the integrals over  $R_2$

$$\mathfrak{R} = \boxed{C_{21}P[\omega_1]} + \boxed{C_{22}P[\omega_2]} + \underbrace{+ (C_{12} - C_{22})P[\omega_2] \int_{R_1} P(x | \omega_2) dx}_{\underset{0}{\downarrow}} - \underbrace{(C_{21} - C_{11})P[\omega_1] \int_{R_1} P(x | \omega_1) dx}_{\underset{0}{\downarrow}}$$

- The first two terms are constant as far as our minimization is concerned since they do not depend on  $R_1$ , so we will be seeking a decision region  $R_1$  that minimizes:

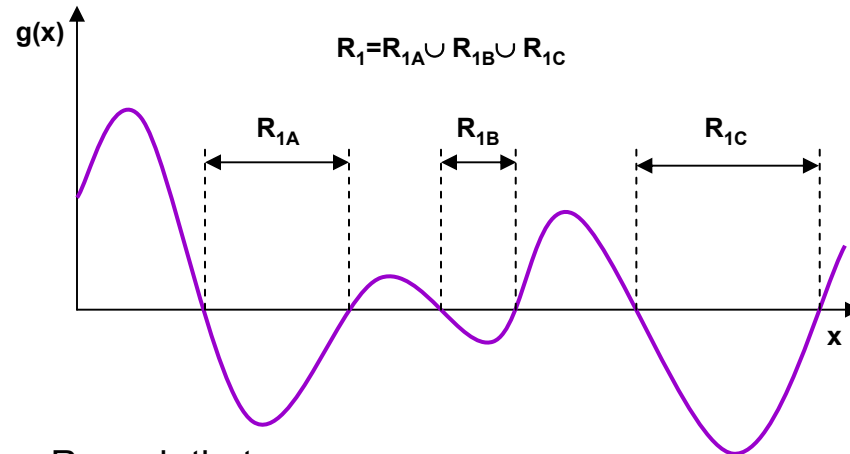
$$\begin{aligned} R_1 &= \operatorname{argmin} \left\{ \int_{R_1} [(C_{12} - C_{22})P[\omega_2]P(x | \omega_2) - (C_{21} - C_{11})P[\omega_1]P(x | \omega_1)] dx \right\} \\ &= \operatorname{argmin} \left\{ \int_{R_1} g(x) dx \right\} \end{aligned}$$



# The Bayes Risk (3)

- Let's forget about the actual expression of  $g(x)$  to develop some intuition for what kind of decision region  $R_1$  we are looking for

- Intuitively, we will select for  $R_1$  those regions that minimize the integral  $\int_{R_1} g(x) dx$ 
  - In other words, those regions where  $g(x) < 0$



- So we will choose  $R_1$  such that

$$(C_{21} - C_{11})P[\omega_1]P(x | \omega_1) > (C_{12} - C_{22})P[\omega_2]P(x | \omega_2)$$

- And rearranging

$$\frac{P(x | \omega_1)}{P(x | \omega_2)} > \frac{(C_{12} - C_{22}) P[\omega_2]}{(C_{21} - C_{11}) P[\omega_1]}$$

- Therefore, minimization of the Bayes Risk also leads to a **Likelihood Ratio Test**



# The Bayes Risk: an example

- Consider a classification problem with two classes defined by the following likelihood functions

$$P(x | \omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2} \frac{x^2}{3}}$$

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

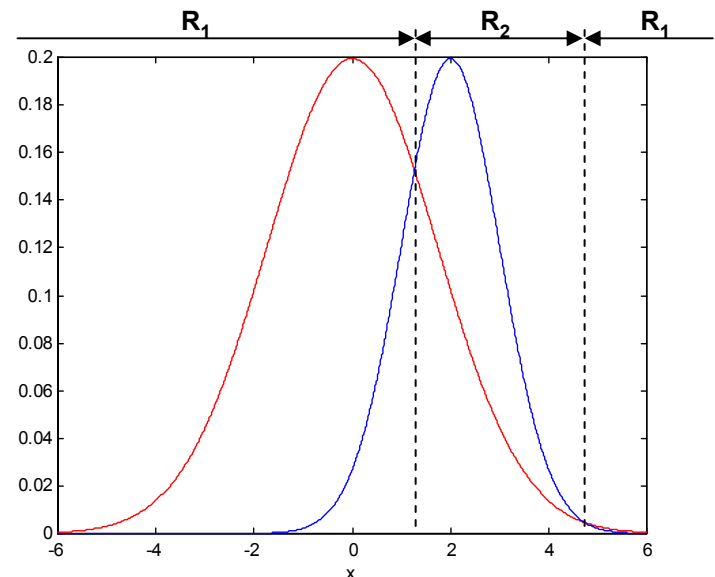
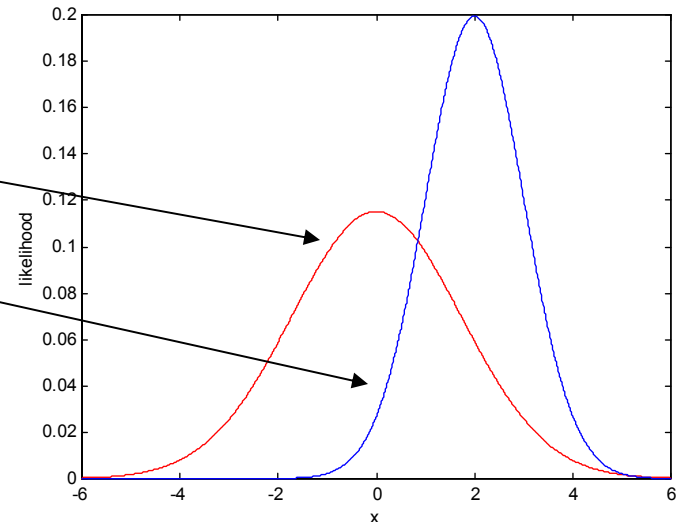
- Sketch the two densities
- What is the likelihood ratio?
- Assume  $P[\omega_1]=P[\omega_2]=0.5$ ,  $C_{11}=C_{22}=0$ ,  $C_{12}=1$  and  $C_{21}=3^{1/2}$ . Determine a decision rule that minimizes the probability of error

$$\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2} \frac{x^2}{3}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}} > \frac{1}{\sqrt{3}}$$

$$\frac{e^{-\frac{1}{2} \frac{x^2}{3}}}{e^{-\frac{1}{2}(x-2)^2}} > 1$$

$$-\frac{1}{2} \frac{x^2}{3} + \frac{1}{2}(x-2)^2 > 0$$

$$2x^2 - 12x + 12 > 0 \Rightarrow x = 4.73, 1.27$$



# Variations of the Likelihood Ratio Test (1)

- The LRT decision rule that minimizes the Bayes Risk is commonly called the Bayes Criterion

$$\Lambda(x) = \frac{P(x | \omega_1)}{P(x | \omega_2)} \underset{\omega_2}{>} \frac{(C_{12} - C_{22}) P[\omega_2]}{(C_{21} - C_{11}) P[\omega_1]} \underset{\omega_1}{<} \quad \text{Bayes criterion}$$

- Many times we will simply be interested in minimizing the probability of error, which is a special case of the Bayes Criterion that uses the so-called symmetrical or zero-one cost function. This version of the LRT decision rule is referred to as the Maximum A Posteriori Criterion, since it seeks to maximize the posterior  $P(\omega_i|x)$

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \Rightarrow \Lambda(x) = \frac{P(x | \omega_1)}{P(x | \omega_2)} \underset{\omega_2}{>} \frac{P(\omega_2)}{P(\omega_1)} \Leftrightarrow \frac{P(\omega_1 | x)}{P(\omega_2 | x)} \underset{\omega_2}{>} 1 \quad \text{Maximum A Posteriori (MAP) Criterion}$$

- Finally, for the case of equal priors  $P[\omega_i]=1/2$ , and the zero-one cost function the LTR decision rule is called the Maximum Likelihood Criterion, since it will minimize the likelihood  $P(x|\omega_i)$

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \Rightarrow \Lambda(x) = \frac{P(x | \omega_1)}{P(x | \omega_2)} \underset{\omega_2}{>} 1 \quad \text{Maximum Likelihood (ML) Criterion}$$

$$P(\omega_i) = \frac{1}{C} \quad \forall i$$



# Variations of the Likelihood Ratio Test (2)

---

## ■ Two more decision rules are commonly cited in the related literature

- The **Neyman-Pearson Criterion**, used in Detection and Estimation Theory, which also leads to an LRT decision rule, fixes one class error probabilities, say  $\varepsilon_1 < \alpha$ , and seeks to minimize the other
  - For instance, for the sea-bass/salmon classification problem of Lecture 1, there may be some kind of government regulation that we must not misclassify more than 1% of salmon as sea bass
  - The Neyman-Pearson Criterion is very attractive since it does not require knowledge of priors and cost function
- The **Minimax Criterion**, used in Game Theory, is derived from the Bayes criterion, and seeks to minimize the maximum Bayes Risk
  - The Minimax Criterion does not require knowledge of the priors, but it needs a cost function
- For more information on these methods, the reader is referred to “Detection, Estimation and Modulation Theory”, by H.L. van Trees, the classical reference in this field



# Minimum $P[\text{error}]$ rule for multi-class problems

- **The decision rule that minimizes  $P[\text{error}]$  generalizes very easily to multi-class problems**

- For clarity in the derivation, the probability of error is better expressed in terms of the probability of making a correct assignment

$$P[\text{error}] = 1 - P[\text{correct}]$$

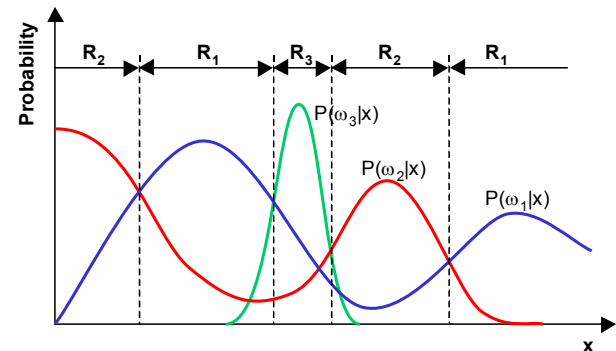
- The probability of making a correct assignment is

$$P[\text{correct}] = \sum_{i=1}^C P(\omega_i) \int_{R_i} P(x | \omega_i) dx$$

- The problem of minimizing  $P[\text{error}]$  is equivalent to that of maximizing  $P[\text{correct}]$ . Expressing  $P[\text{correct}]$  in terms of the posteriors:

$$P[\text{correct}] = \sum_{i=1}^C P(\omega_i) \int_{R_i} P(x | \omega_i) dx = \sum_{i=1}^C \int_{R_i} P(x | \omega_i) P(\omega_i) dx = \sum_{i=1}^C \underbrace{\int_{R_i} P(\omega_i | x) P(x) dx}_{\mathfrak{I}_i}$$

- In order to maximize  $P[\text{correct}]$ , we will have to maximize each of the integrals  $\mathfrak{I}_i$ . In turn, each integral  $\mathfrak{I}_i$  will be maximized by choosing the class  $\omega_i$  that yields the maximum  $P[\omega_i|x]$   
 $\Rightarrow$  we will define  $R_i$  to be the region where  $P[\omega_i|x]$  is maximum



- **Therefore, the decision rule that minimizes  $P[\text{error}]$  is the MAP Criterion**



# Minimum Bayes Risk for multi-class problems

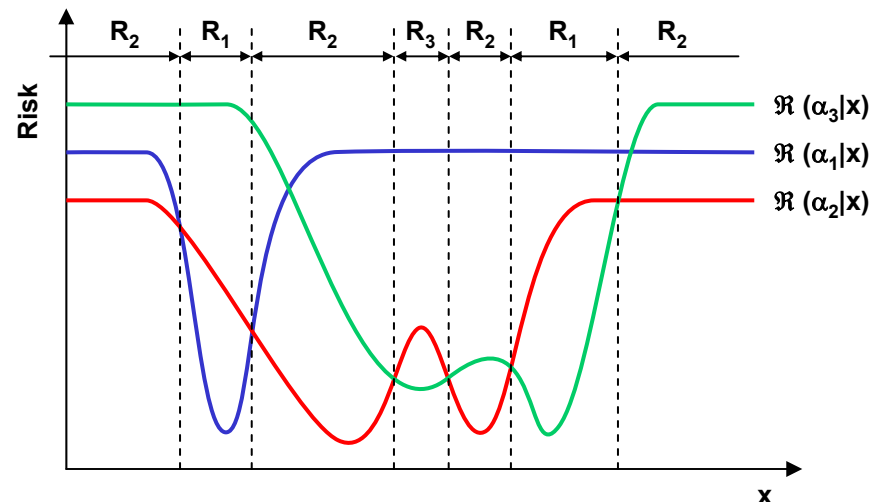
- To determine which decision rule yields the minimum Bayes Risk for the multi-class problem we will use a slightly different formulation
  - We will denote by  $\alpha_i$  the decision to choose class  $\omega_i$ ,
  - We will denote by  $\alpha(x)$  the overall decision rule that maps features  $x$  into classes  $\omega_i$ :  $\alpha(x) \rightarrow \{\alpha_1, \alpha_2, \dots, \alpha_C\}$
- The (conditional) risk  $\mathcal{R}(\alpha_i|x)$  of assigning a feature  $x$  to class  $\omega_i$  is

$$\mathcal{R}(\alpha(x) \rightarrow \alpha_i) = \mathcal{R}(\alpha_i | x) = \sum_{j=1}^C C_{ij} P(\omega_j | x)$$

- And the Bayes Risk associated with the decision rule  $\alpha(x)$  is

$$\mathcal{R}(\alpha(x)) = \int \mathcal{R}(\alpha(x) | x) P(x) dx$$

- In order to minimize this expression, we will have to minimize the conditional risk  $\mathcal{R}(\alpha(x)|x)$  at each point  $x$  in the feature space, which in turn is equivalent to choosing  $\omega_i$  such that  $\mathcal{R}(\alpha_i|x)$  is minimum

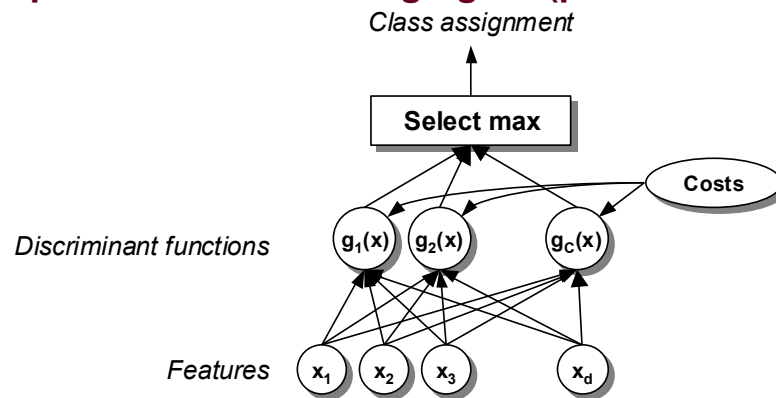


# Discriminant functions

- All the decision rules we have presented in this lecture have the same structure
  - At each point  $x$  in feature space choose class  $\omega_i$  which maximizes (or minimizes) some measure  $g_i(x)$
- This structure can be formalized with a set of discriminant functions  $g_i(x)$ ,  $i=1..C$ , and the following decision rule

"assign  $x$  to class  $\omega_i$  if  $g_i(x) > g_j(x) \quad \forall j \neq i$ "

- Therefore, we can visualize the decision rule as a network or machine that computes  $C$  discriminant functions and selects the category corresponding to the largest discriminant. Such network is depicted in the following figure (presented already in Lecture 1)



- Finally, we express the three basic decision rules: Bayes, MAP and ML in terms of Discriminant Functions to show the generality of this formulation

Criterion	Discriminant Function
Bayes	$g_i(x) = -\mathcal{R}(\alpha_i x)$
MAP	$g_i(x) = P(\omega_i x)$
ML	$g_i(x) = P(x \omega_i)$

