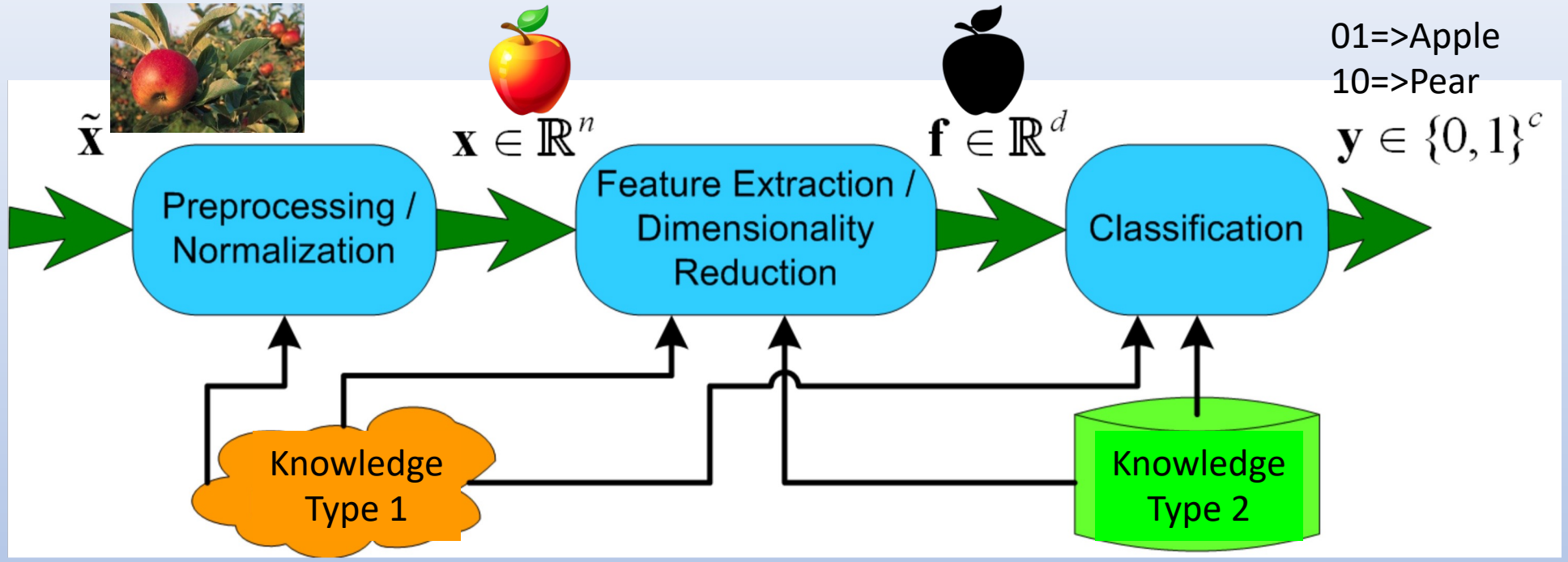
5 MAP Decision and Classifiers



- A diagram of machine vision system diagram
- Visual object recognition is very easy for human brain but very difficult for machine or computer.

5 MAP Decision and Classifiers

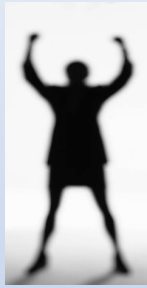
Outline

- Understand how the decision theory is developed from an intuitive process that everybody understands to an abstract mathematical theory
- The optimal decision rule: MAP and Bayesian decision rules
- Evaluation a pattern recognition system: Recognition accuracy or error rate
- Generalize the classification process to the discriminant function evaluation
- Discriminant function for multivariate Gaussian PDF
- Mahalanobis distance and Euclidean distance
- Linear classifier

How to make the best decision in uncertainty?

Let's see a simple example:

- You totally unclearly see a student in a classroom of school of EEE. You need decide/judge if this student is a male or a female. What is your decision?
- if you are not silly you will judge that the student is a male based on the fact that the student population in the school of EEE is 70% for male and 30% for female.
- Is your decision the same in a classroom of school of accounting, supposing the student population in the school of accounting is 30% for male and 70% for female?
- How do you work out your decision or describe your decision rule or justify it is a good decision scientifically or theoretically or mathematically?
- Mathematically, this is the **maximum probability** decision based on our **knowledge** that $p(\omega_1) = 0.7$ and $p(\omega_2) = 0.3$ for a male and female in the school of EEE, and that $p(\omega_1) = 0.3$ and $p(\omega_2) = 0.7$ for a male and female in the school of accounting, respectively.
- So a wise decision rule is to decide ω_1 if $p(\omega_1) > p(\omega_2)$, and vice versa.
- Is your decision definitely correct? No. But is there any other better decision?
- No. The probability of the wrong decision (error rate) is $p(e) = 0.3$, which is the minimum.
- What is the error rate for the opposite decision?



5 MAP Decision and Classifiers

- Now if you know the student in the classroom of school of accounting is of height 1.75, how do you make the best decision?
- No difference. The best decision is still based on $p(\omega_1)$ and $p(\omega_2)$ to see which is larger.
- So simple? Yes! There is no magic in our world. But what is $p(\omega_i)$, $i = 1, 2$, now?
- Same as that $p(\omega_i)$ of the school of accounting are different from those of EEE, $p(\omega_i)$ including the knowledge of height of 1.75 are different from those without this knowledge.
- For convenience, they are mathematically denoted by $p(\omega_1|x)$ and $p(\omega_2|x)$, $x = 1.75$, called a posterior probability. It means the probability of a class ω_i after knowing the data value x in a particular scenario. $p(\omega_1)$ and $p(\omega_2)$ without knowing the value of x are then renamed as the prior probability.
- The name of decision rule is modified from the **maximum probability** decision to the **maximum a posterior probability** decision, **MAP decision** in short.
- Thus, in general, we have **MAP decision rule**:

$$\text{Decide } \omega_k : \text{if } p(\omega_k \mid x) > p(\omega_i \mid x), i \neq k$$

$$\text{or more general: } \omega_k = \arg\max_{\omega_i}[p(\omega_i \mid x)]$$

5 MAP Decision and Classifiers

- As you decide ω_k, based on your observed value x, the probability that you make a correct decision is hence $p(\omega_k|x)$. Therefore, the probability that you make a wrong decision or error rate is:

$$p(e_k \mid x) = 1 - p(\omega_k \mid x)$$

- If your decision is ω_k, because $p(\omega_k|x)$ is maximum, consequently, the probability of the decision error will be minimum.
- Therefore, this decision rule is optimal in the sense of minimizing the probability of the wrong decision.
- The maximum a posterior (MAP) decision rule minimizes the probability of the decision error.

$$\text{Decide: } \omega_k = \arg\max_{\omega_i}[p(\omega_i \mid x)] = \arg\min_{\omega_i}[p(e_i \mid x)]$$

5 MAP Decision and Classifiers

- However, it is not very strait forward to get the value of $p(\omega_i|x)$. It is much easier and more convenient to get the values of $p(\omega_i)$ and $p(x|\omega_i)$, called the class-conditional probability.
- From the basic rule of probability theory:

$$p(x,\omega_i) = p(x)p(\omega_i \mid x) = p(\omega_i)p(x \mid \omega_i)$$

$$p(A,B) = p(A)p(B \mid A) = p(B)p(A \mid B)$$

We have $$p(\omega_i \mid x) = \frac{p(\omega_i)p(x \mid \omega_i)}{p(x)}$$

where $$p(x) = \sum_{i=1}^{c} p(x \mid \omega_i)p(\omega_i)$$

- $p(x)$ is called mixture PDF or PMF, the probability (density) of a person's height at x. We can remove it from the decision as it has the same value for all classes.

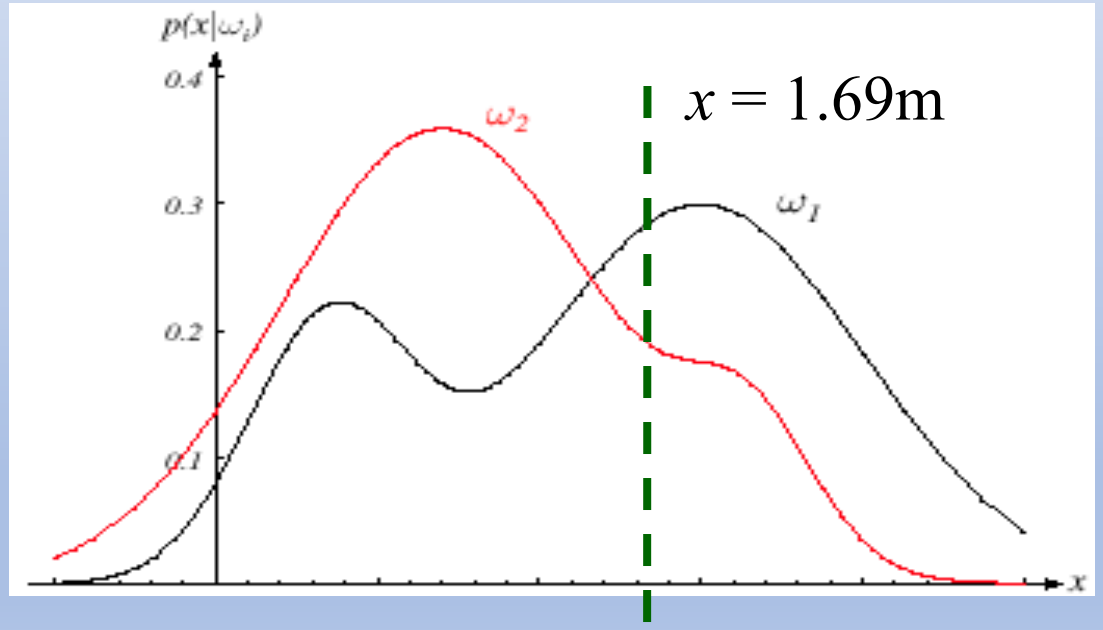
5 MAP Decision and Classifiers

- Now we know how to make the optimal decision/recognition at a specific value, x=1.75m, and compute the error rate. The concept is intuitive and easy to understand.
- However, to fully automatically recognize a person's gender for all value of x, we need the the class conditional probability $p(x|\omega_i)$, of all possible value of x. This brings big challenge.

$$\sum_{x=0}^{3} p(x \mid \omega_1) = 1, \text{ and } \sum_{x=0}^{3} p(x \mid \omega_2) = 1 \text{ for discrete } x$$

$$\int_0^3 p(x \mid \omega_1)dx = 1, \text{ and } \int_0^3 p(x \mid \omega_2)dx = 1 \text{ for continuous } x$$

$$\text{Decide: } \omega_k = \arg\max_{\omega_i}[p(\omega_i \mid x)]$$



- Is is possible to simplify the computation of the system? Convert decision into classification.

5 MAP Decision and Classifiers

- How good is the system? Or what is the performance of the system? Or how to evaluate your designed system?

$$\text{Decide: } \omega_k = \arg\max_{\omega_i}[p(\omega_i \mid x)]$$

- We have understand the formula to compute the error probability for a specific value of x.

$$p(e_k \mid x) = 1 - p(\omega_k \mid x) = \sum_{\substack{i=1 \\ i \neq k}}^{c} p(\omega_i \mid x)$$

- It is not difficult to understand that the performance of a recognition system can be measured by the average of $p(e_k|x)$ over all possible value of x.

5 MAP Decision and Classifiers

- How to average $p(e_k|x)$ over all possible value of x?

$$p(e)? = \frac{1}{n_x}\sum_{x=0}^{3} p(e_k \mid x) \text{ for discrete } x$$

$$p(e)? = \frac{1}{3}\int_0^3 p(e_k \mid x)dx, \text{ for continuous } x$$

- The above is not a proper way because x is a random variable with different probability of occurrence at different x value.
- The right way for a random variable should be:

$$p(e) = \sum_{x=-\infty}^{\infty} p(e_k \mid x)p(x) \text{ for discrete } x$$

$$p(e) = \int_{-\infty}^{\infty} p(e_k \mid x)p(x)dx, \text{ for continuous } x$$

5 MAP Decision and Classifiers

- Let's take the continuous x for further working:

$$p(e) = \int_{-\infty}^{\infty} p(e_k \mid x) p(x) dx = \int_{-\infty}^{\infty} [1 - p(\omega_k \mid x)] p(x) dx$$

$$= \int_{-\infty}^{\infty} [1 - \frac{p(\omega_k) p(x \mid \omega_k)}{p(x)}] p(x) dx = 1 - \int_{-\infty}^{\infty} p(\omega_k) p(x \mid \omega_k) dx$$

- Note that for different region of x, the pattern recognition system has different decision ω_k. Decision/classification is to partition the whole space of x into c decision regions $\mathcal{R}_i$. Therefore,

$$p(e) = 1 - \sum_{i=1}^{c} \int_{\mathfrak{R}_i} p(\omega_i) p(x \mid \omega_i) dx = 1 - \sum_{i=1}^{c} p(\omega_i) \int_{\mathfrak{R}_i} p(x \mid \omega_i) dx$$

- Obviously, the probability of the correct decision is

$$p(correct) = 1 - p(e) = \sum_{i=1}^{c} p(\omega_i) \int_{\mathfrak{R}_i} p(x \mid \omega_i) dx$$

- For the 2-class problem:

$$p(e) = 1 - p(\omega_1)\int_{\Re_1} p(x \mid \omega_1)dx - p(\omega_2)\int_{\Re_2} p(x \mid \omega_2)dx$$

$$= p(\omega_1)\int_{\Re_2} p(x \mid \omega_1)dx + p(\omega_2)\int_{\Re_1} p(x \mid \omega_2)dx$$

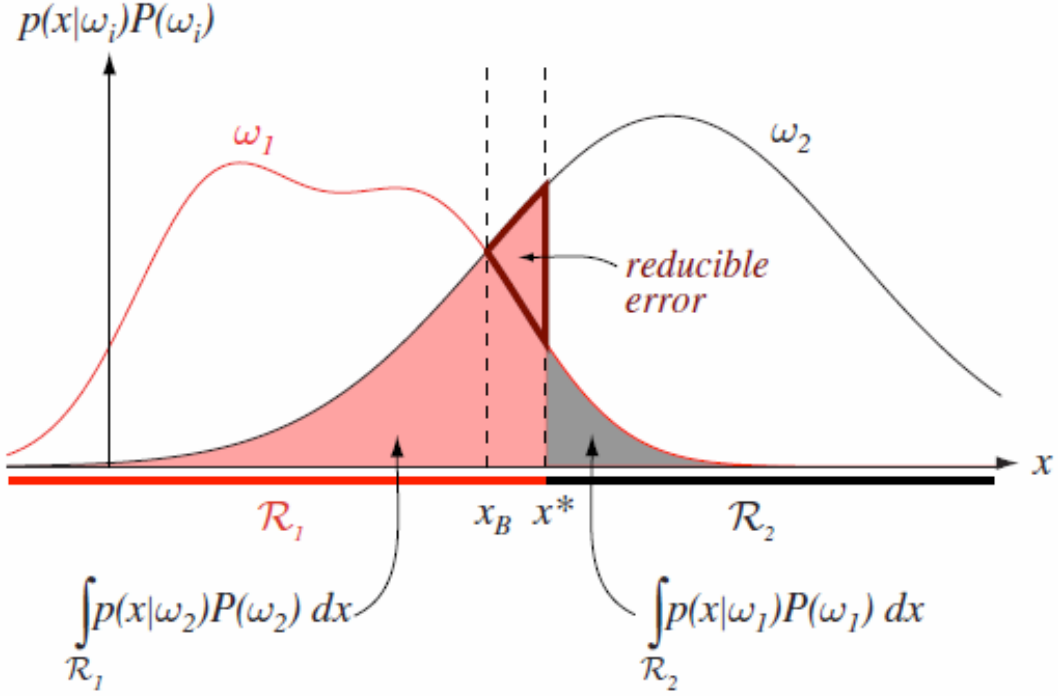
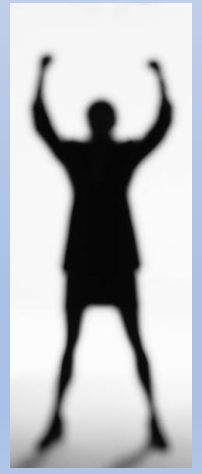
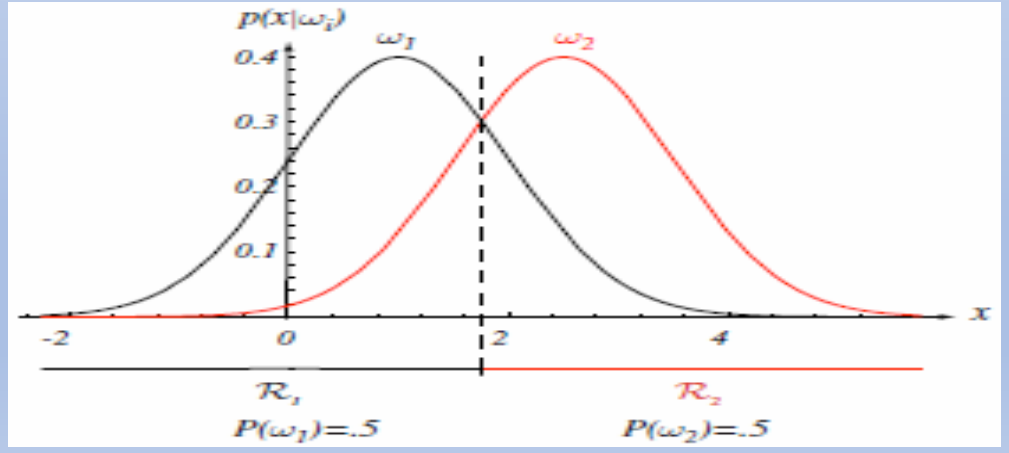


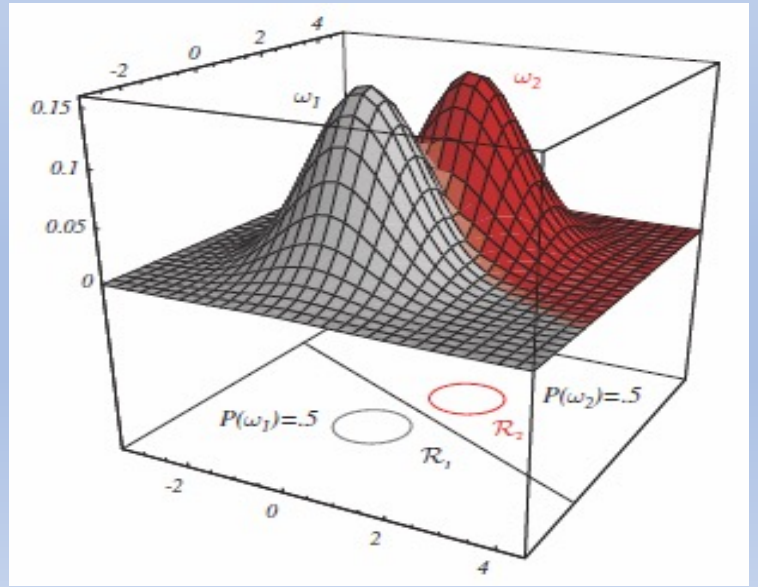
FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible; this is the Bayes decision and gives the Bayes error rate. From: Richard O.

5 MAP Decision and Classifiers

- Now if we not only know the person's height x_1, but also the length of the hair x_2, it is not difficult to image that we will make better decision, i.e. have lower error probability. The recognition principle will be the same as before as long as now you use a vector $\mathbf{x} = [x_1, x_2]^T$ to replace the scalar x before. The probability distribution is now two dimensional. A scalar threshold to separate the two classes in 1D case becomes a curve in the plane spanned by $\mathbf{x}$.

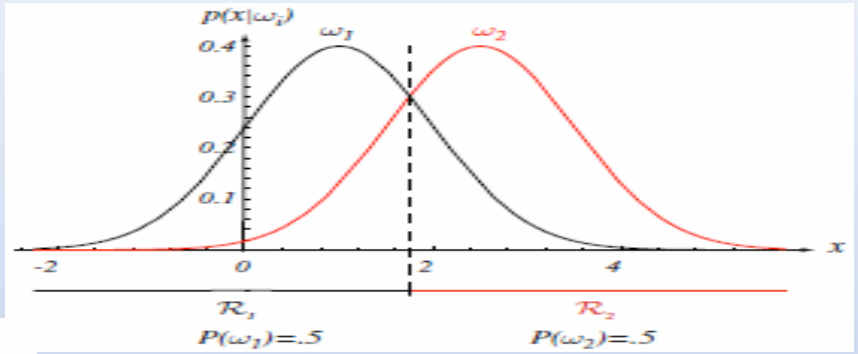


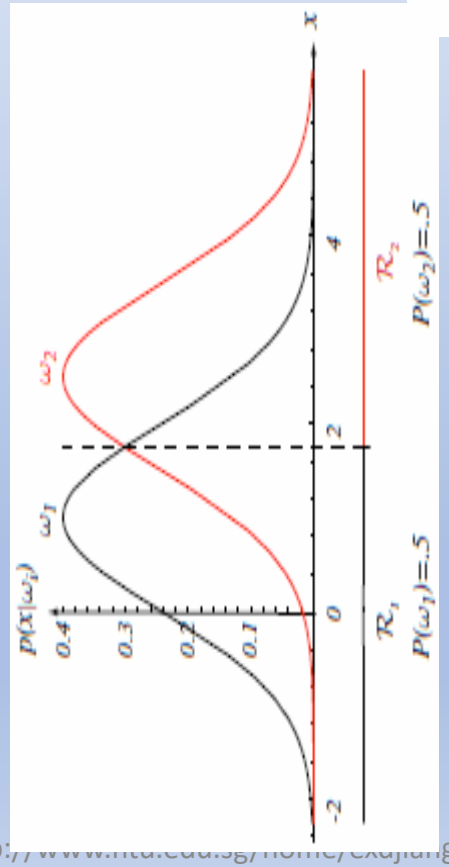


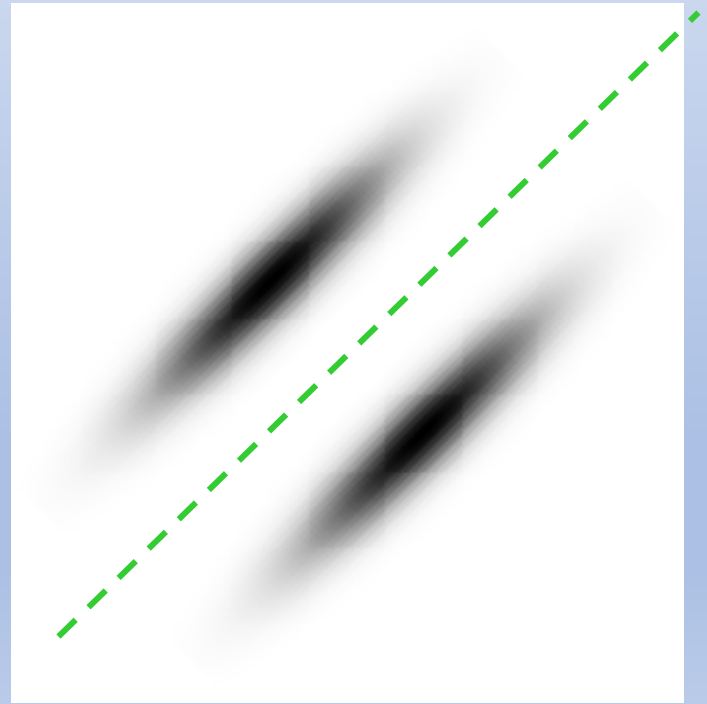


5 MAP Decision and Classifiers

- The increase of the information or data or feature dimension in general increases the probability of the correct decision or reduce the probability of the decision error.







5 MAP Decision and Classifiers

- In some application, wrongly classifying one class may not be at the same cost or same risk as wrongly classifying another class .
- Let's λ_{ki} denote the cost or risk of wrongly classifying the object of class ω_i into class ω_k. The cost or risk of the decision ω_k is then

$$R_k(\mathbf{x}) = \sum_{\substack{i=1 \\ i \neq k}}^{c} \lambda_{ki} p(\omega_i \mid \mathbf{x})$$

- Therefore the decision rule that minimizes the risk or cost becomes

$$\text{Decide: } \omega_k = \arg\min_{\omega_i}[R_i(\mathbf{x})] = \arg\min_{\omega_i}[\sum_{\substack{j=1 \\ j \neq i}}^{c} \lambda_{ij} p(\omega_j \mid \mathbf{x})]$$

- This is the famous Bayesian decision rule. Cost functions allow us to treat situations where some kinds of classification mistakes are more costly than others, although we often discuss the simplest case where all errors are equally costly.

5 MAP Decision and Classifiers

- In some applications, the correct decision for different class may have different cost $\lambda_{kk} \neq 0$. By including the cost of correct decision, the cost or risk of a decision is then

$$R_k(\mathbf{x}) = \sum_{i=1}^{c} \lambda_{ki} p(\omega_i \mid \mathbf{x})$$

- Therefore, the famous Bayesian decision rule is generalized as: $\text{Decide } \omega_k = \arg\min_{\omega_i}[R_i(\mathbf{x})]$

$$= \arg\min_{\omega_i}[\sum_{j=1}^{c} \lambda_{ij} p(\omega_j \mid \mathbf{x})]$$

$$= \arg\min_{\omega_i}[\sum_{j=1}^{c} \lambda_{ij} p(\omega_j) p(\mathbf{x} \mid \omega_j) / \ p(\mathbf{x})]$$

$$= \arg\min_{\omega_i}[\sum_{j=1}^{c} \lambda_{ij} p(\omega_j) p(\mathbf{x} \mid \omega_j)]$$

5 MAP Decision and Classifiers

$$R_k(\mathbf{x}) = \sum_{i=1}^{c} \lambda_{ki} p(\omega_i \mid \mathbf{x})$$

is the risk or cost for the decision ω_k at a specific value of $\mathbf{x}$, called the conditional risk.

- How good is a decision rule is evaluated by the average or overall cost or risk of a pattern recognition system

$$R = \int_{\Re_x} R_k(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \int_{\Re_x} \sum_{i=1}^{c} \lambda_{ki} p(\omega_i \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$= \int_{\Re_x} \sum_{i=1}^{c} \lambda_{ki} p(\omega_i) p(\mathbf{x} \mid \omega_i) d\mathbf{x}$$

$$= \sum_{k=1}^{c} \int_{\Re_{xk}} \sum_{i=1}^{c} \lambda_{ki} p(\omega_i) p(\mathbf{x} \mid \omega_i) d\mathbf{x}$$

- This overall cost is minimized by the Bayesian decision rule. It delivers the best performance that can be achieved.

5 MAP Decision and Classifiers

- Two-category classification:
  - We have two possible classes : ω_1, ω_2.
  - We also have two actions:
    - α_1 corresponds to deciding that the true class is ω_1;
    - α_2 corresponds to deciding that the true class is ω_2.
  - Let $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$, we have the conditional risks as

$$R_1(\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x})$$
$$R_2(\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x})$$

  - The Bayes decision rule : decide that the true class is ω_1 if $R_1(\mathbf{x}) < R_2(\mathbf{x})$ and ω_2 otherwise.

5 MAP Decision and Classifiers

- From the decision rule, we obtain the following

$$(\lambda_{21} - \lambda_{11})P(\omega_1|\mathrm{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathrm{x}).$$

- Normally $\lambda_{21} - \lambda_{11}$ and $\lambda_{12} - \lambda_{22}$ are positive because the loss is greater when making a mistake.
- We can also replace the posterior probability by the product of likelihood and prior , and drop the evidence term.
- Then we can write the Bayes decision rule as: Decide ω_1 if

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)}.$$

If the above likelihood ratio is greater than ratio of loss weighted priors, we take action α_1 (decide ω_1). Otherwise take action α_2 (decide ω_2)

This is the **Likelihood Ratio Test (LRT) .**

5 MAP Decision and Classifiers

- We can define the loss to be zero for correct decision and one for wrong decision for simplicity.

$$\lambda(\alpha_i|\omega_j) = \begin{matrix} 0 & \text{if} & i = j \\ 1 & \text{if} & i \neq j \end{matrix}$$

- In a multi-category setting, we can write the loss in the matrix form, whose elemnets are λ_{ij},

$$\begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 1 \\ \vdots & & \ddots & & \vdots \\ 1 & 1 & \dots & & 0 \end{bmatrix}$$

- With this 0/1 loss function, the Bayesian decision rule becomes MAP rule that minimizes the error rate.

5 MAP Decision and Classifiers

- Numerical Example 1:

In a study of two types of objects, it is found that the sizes of objects can be modeled as two Gaussian (normal) distributions. In general, there is about the same number of type A objects as there is for type B objects. For type A objects, its class conditional density is $\mathcal{N}(2,1)$ and for type B, its class conditional density is $\mathcal{N}(8,1)$. Find the decision rule that you can use to discriminate the two types of objects using the Likelihood Ratio Test if there is the same penalty for making all wrong decisions.

How would the decision rule be affected if there are twice more objects of type A than the objects of type B?

5 MAP Decision and Classifiers

Solution:

2 types of objects: $\{\omega_1$ for type A, ω_2 for type B$\}$; let size be x,

Type A: $p(x|\omega_1) \sim \mathcal{N}(2,1) \Rightarrow p(x|\omega_1) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(x-2)^2]$

Type B: $p(x|\omega_2) \sim \mathcal{N}(8,1) \Rightarrow p(x|\omega_2) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(x-8)^2]$

From the question, we know $P(\omega_2) = P(\omega_1)$.

Let penalty for making mistake be k, then $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = \lambda_{21} = k$.

Likelihood Ratio Test: $R_k(\mathbf{x}) = \sum_{i=1}^{c} \lambda_{ki} p(\omega_i \mid \mathbf{x})$

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)} \Rightarrow \frac{\exp[-\frac{1}{2}(x-2)^2]}{\exp[-\frac{1}{2}(x-8)^2]} > 1$$

$$\begin{aligned} \text{take natural log, } -\frac{1}{2}[(x-2)^2 - (x-8)^2] &> 0 \\ \text{change sign, remove } \frac{1}{2}\text{, we get} (x-2)^2 - (x-8)^2 &< 0 \\ 12x - 60 &< 0 \\ \text{Therefore, } x &< 5 \end{aligned}$$

The rule: Decide ω_1 if $x < 5$, otherwise decide ω_2

5 MAP Decision and Classifiers

Solution:

If there are twice more objects of type A than objects of type B out there,

$$\frac{P(\omega_1)}{P(\omega_2)} = 2 \Rightarrow P(\omega_1) = 2P(\omega_2) \Rightarrow \frac{P(\omega_2)}{P(\omega_1)} = \frac{1}{2}$$

Following the previous solution, we now have

Likelihood Ratio Test:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)} = \frac{1}{2} \Rightarrow \frac{\exp[-\frac{1}{2}(x-2)^2]}{\exp[-\frac{1}{2}(x-8)^2]} > \frac{1}{2}$$

$$\begin{aligned}
\text{take natural log, } -\frac{1}{2}[(x-2)^2 - (x-8)^2] &> \ln\frac{1}{2} \\
\text{change sign, } \times 2\text{, we get } (x-2)^2 - (x-8)^2 &< -2\ln\frac{1}{2} \\
12x - 60 &< -2\ln\frac{1}{2} \\
\text{Therefore, } x &< \frac{60 - 2\ln(1/2)}{12} = 5.16
\end{aligned}$$

The rule: Decide ω_1 if $x < 5.16$, otherwise decide ω_2

The decision boundary is shifted towards class ω_2.

5 MAP Decision and Classifiers

- Numerical Example 2

A set of patterns is to be classified into 2 classes based on a scalar feature. We know that the conditional density functions of the feature for the pattern classes can be represented as two Gaussians, $\mathcal{N}(2,3)$ and $\mathcal{N}(4,1)$. The patterns from the two classes are equal likely to be seen. However, if a pattern from class 1 being wrongly classified into class 2, the cost associated is $\sqrt{3}$ while the cost for a pattern from class 2 being wrongly classified into class 1 is 1 . Using the Likelihood Ratio Test, find the decision rule that minimizes the probability of classification error.

5 MAP Decision and Classifiers

Solution:

We have 2 classes: $\{\omega_1$ for class 1, ω_2 for class 2$\}$; let the feature be x,

Class 1: $p(x|\omega_1) \sim \mathcal{N}(2,3) \Rightarrow p(x|\omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}} \exp[-\frac{1}{2}\frac{(x-2)^2}{3}]$

Class 2: $p(x|\omega_2) \sim \mathcal{N}(4,1) \Rightarrow p(x|\omega_2) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(x-4)^2]$

From the question, we know $P(\omega_2) = P(\omega_1)$.

Also, $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = 1$ and $\lambda_{21} = \sqrt{3}$. ?

Likelihood Ratio Test:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\frac{P(\omega_2)}{P(\omega_1)} \Rightarrow \frac{\frac{1}{\sqrt{2\pi}\sqrt{3}} \exp[-\frac{1}{2}\frac{(x-2)^2}{3}]}{\frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(x-4)^2]} > \frac{1}{\sqrt{3}}$$

Canceling out $\frac{1}{\sqrt{2\pi}}$, $\times\sqrt{3}$, we have : $\frac{\exp[-\frac{1}{2}\frac{(x-2)^2}{3}]}{\exp[-\frac{1}{2}(x-4)^2]} > 1$

take natural log, $-\frac{1}{2}\left[\frac{(x-2)^2}{3} - (x-4)^2\right] > 0$

$\times 6$, we get : $-(x-2)^2 + 3(x-4)^2 > 0$

expand and divide by 2 : $x^2 - 10x + 22 > 0$

5 MAP Decision and Classifiers

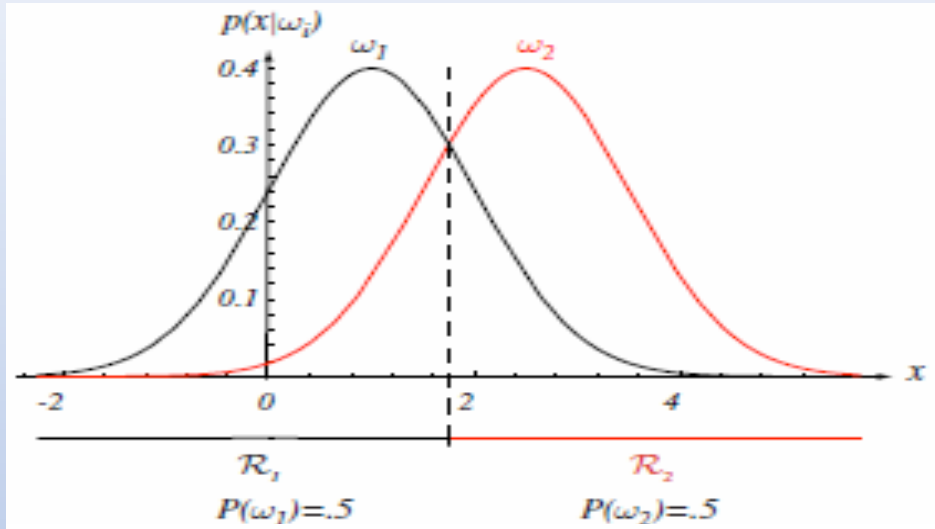
- Numerical Example 2

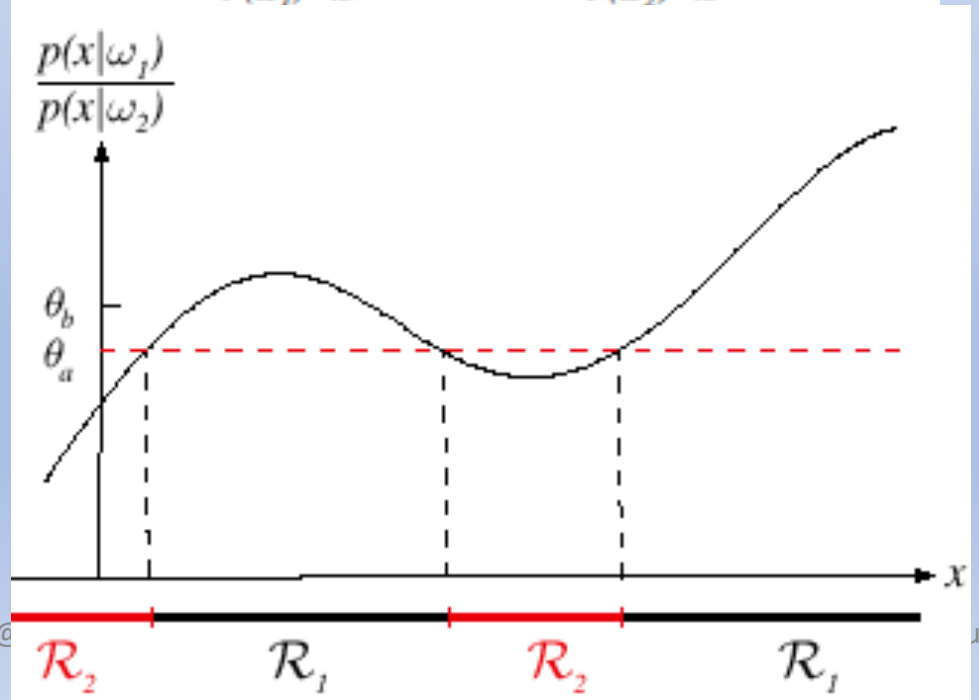
Solution (continued):
Solve the quadratic equation:
Let $f(x) = x^2 - 10x + 22$.

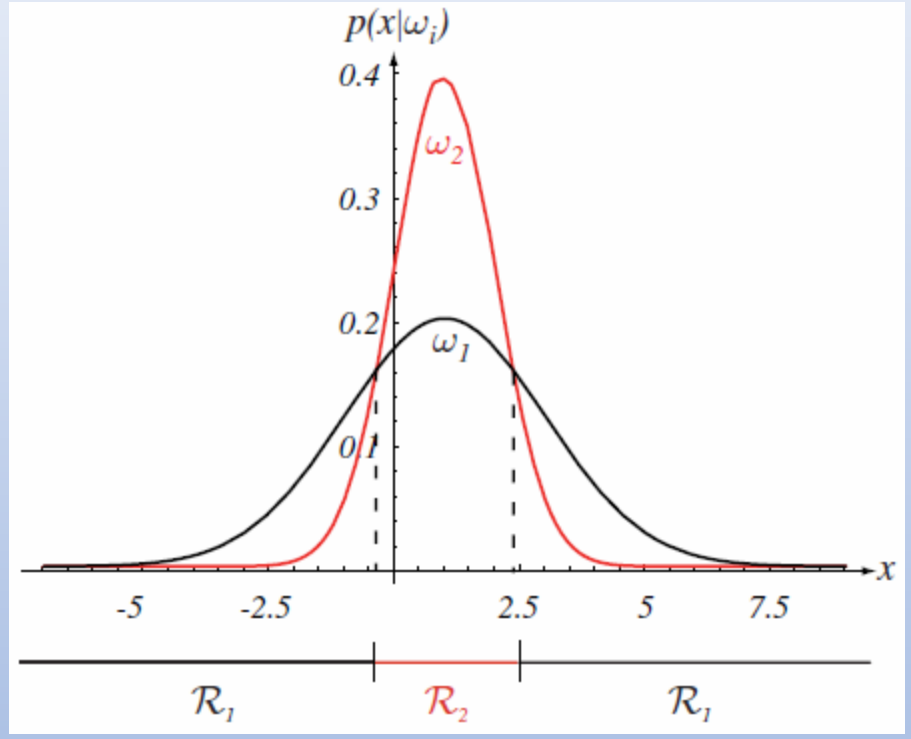
$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \Rightarrow x = \frac{10 \pm \sqrt{10^2 - 4(22)}}{2} = 3.27, 6.73$$

We have $f'(x) = 2x - 10$.
Setting $f'(x) = 0$, we have $x = 5$.
We know that $f(5) < 0$.
Hence, by testing the regions, we find that
$f(x) < 0$ for $3.27 < x < 6.73$, and $f(x) > 0$ for $x < 3.27$ and $x > 6.73$.
Therefore, we have the following rule:
Decide ω_1 if $x < 3.27$ or $x > 6.73$, otherwise decide ω_2

5 MAP Decision and Classifiers







- Connected and disconnected decision regions

5 MAP Decision and Classifiers

- We understand that the Bayesian (optimal) classification with the 0/1 loss function maximizes a posterior probability and hence minimizes the classification error by:

$$\text{Decide: } \omega_k = \arg\max_{\omega_i}[p(\omega_i \mid \mathbf{x})] = \arg\min_{\omega_i}[p(e_i \mid \mathbf{x})]$$

- Since $p(\omega_i \mid \mathbf{x}) = p(\omega_i)p(\mathbf{x} \mid \omega_i)p^{-1}(\mathbf{x})$

  and $p(\mathbf{x})$ is not a function of ω_i, the Bayesian (optimal) classification is to evaluate the called discriminant functions that can be defined as $g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln p(\omega_i)$

  and find the class ω_i that has the maximum value of the discriminant function for a given pattern $\mathbf{x}$.

- Here, a natural logarithm $\ln$ is applied as it is a monotonically increasing function that does not affect the decision result but will simplify its evaluation if $p(\mathbf{x} \mid \omega_i)$ is an exponential function.

5 MAP Decision and Classifiers

- We will derive classifiers for the case that the class conditional PDF is multivariate Gaussian of $p(\mathbf{x}|\omega_i)=N(\mu_i, \Sigma_i)$

$$p(\mathbf{x}\,|\,\omega_i) = \frac{1}{(2\pi)^{d/2}\,|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

- The discriminant function becomes a quadratic function of $\mathbf{x}$:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p(\omega_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{d}{2}\ln 2\pi$$

$$= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + b_i$$

$$= -\frac{1}{2}d_{\Sigma}(\mathbf{x},\boldsymbol{\mu}_i) + b_i$$

- where $d_{\Sigma_i}(\mathbf{x},\boldsymbol{\mu}_i) = (\mathbf{x}-\boldsymbol{\mu}_i)^T\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i), \qquad b_i = \ln p(\omega_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$

5 MAP Decision and Classifiers

$$d_{\Sigma_i}(\mathbf{x},\boldsymbol{\mu}_i)=(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

is called the square of Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_i$.

- Compare to the square of Euclidean distance

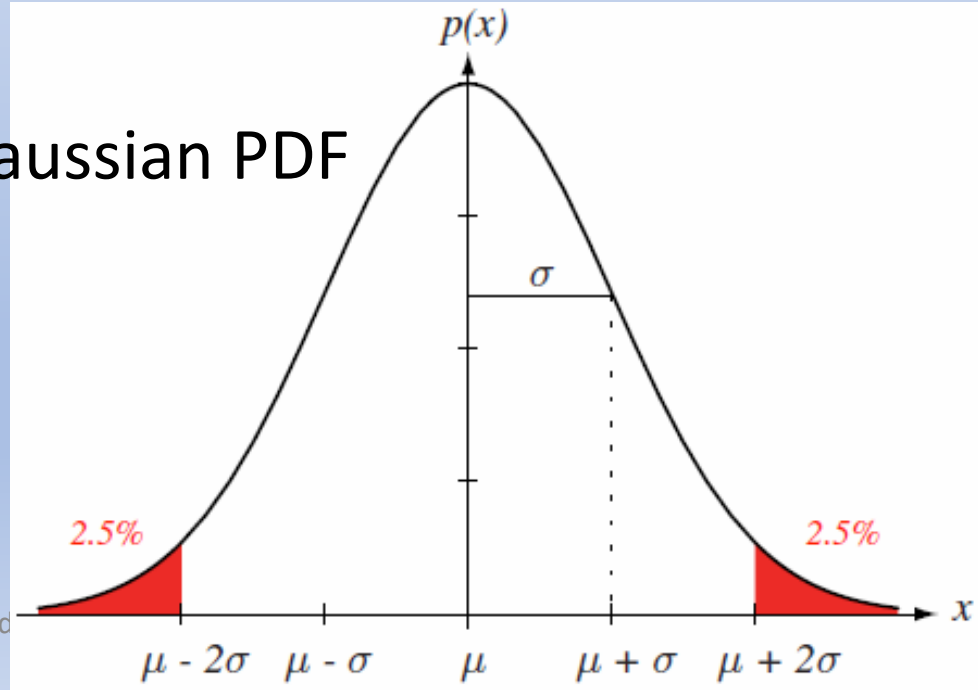
$$d_{Eu}(\mathbf{x},\boldsymbol{\mu}_i)=(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i)$$

- In1D case:

$$d_{Eu}(x,\mu_i)=(x-\mu_i)^2$$

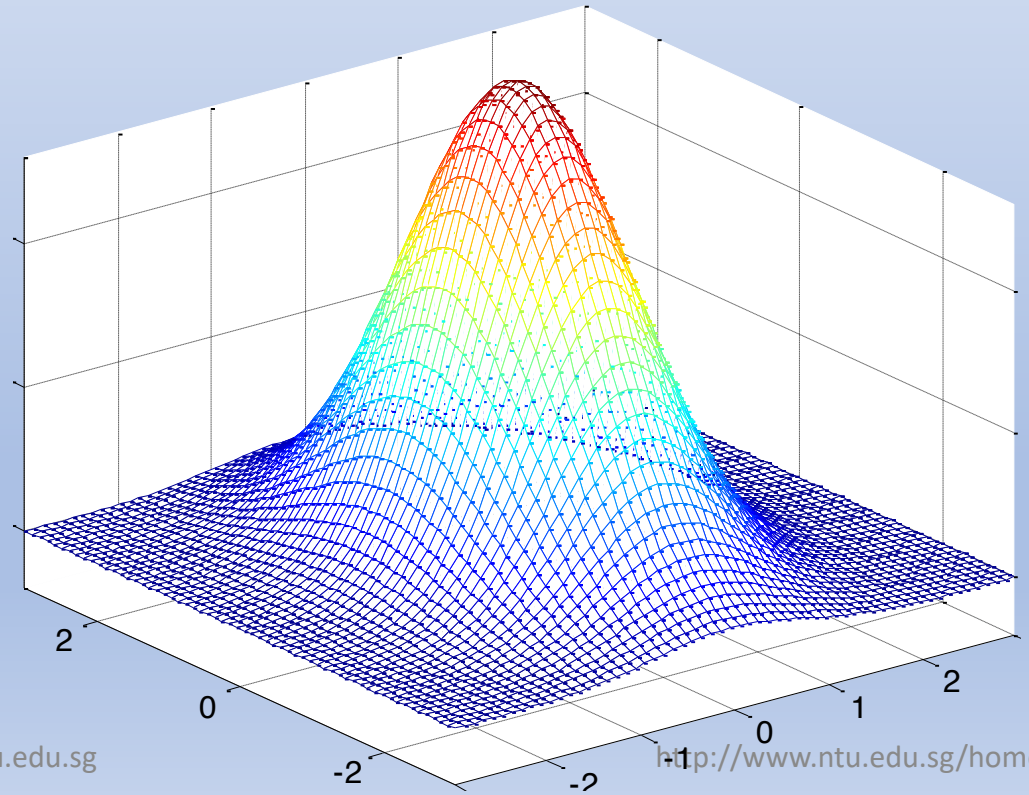
$$d_{\Sigma}(x,\mu_i)=\frac{(x-\mu_i)^2}{\sigma^2}$$

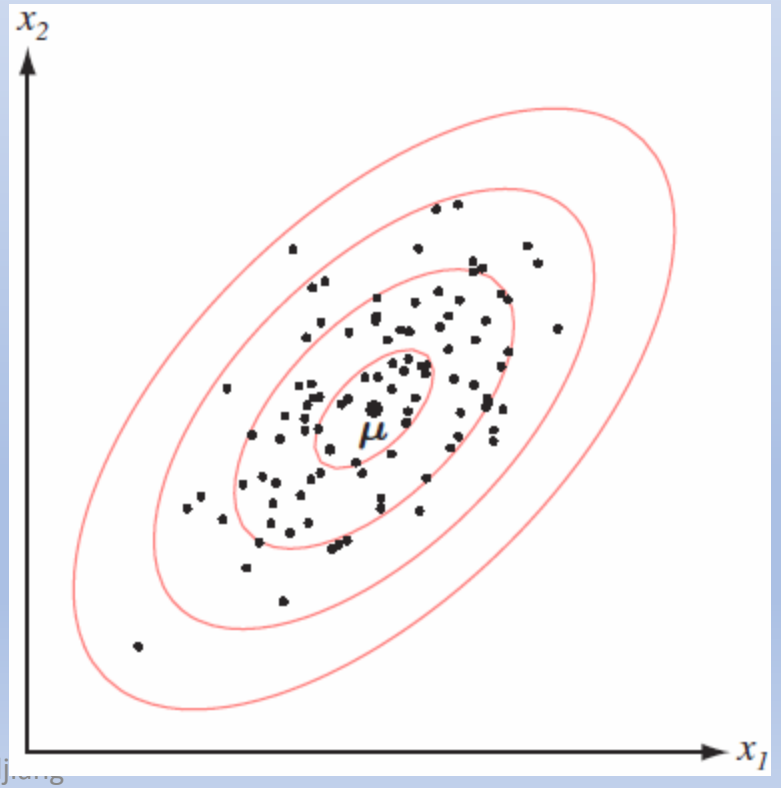
1D Gaussian PDF



5 MAP Decision and Classifiers

- For 2D Gaussian PDF Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean. The ellipses show lines of equal probability density of the Gaussian. All points on a ellipse has the same Mahalanobis distance but different Euclidean distances to the mean.





5 MAP Decision and Classifiers

- Quadratic Classifier: recall that the discriminant functions for class conditional PDF of multivariate Gaussian of $p(\mathbf{x}|\omega_i) = N(\mu_i, \Sigma_i)$ is

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p(\omega_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$$

  - This is the general multivariate normal case where the covariance matrices are different for each category. The only term that can be dropped is the $(d/2)\ln 2\pi$ term, and the resulting discriminant functions are inherently quadratic:

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + \omega_{i0}$$

    where $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$ and $\mathbf{w}_i = \Sigma_i^{-1}\mu_i$ and

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

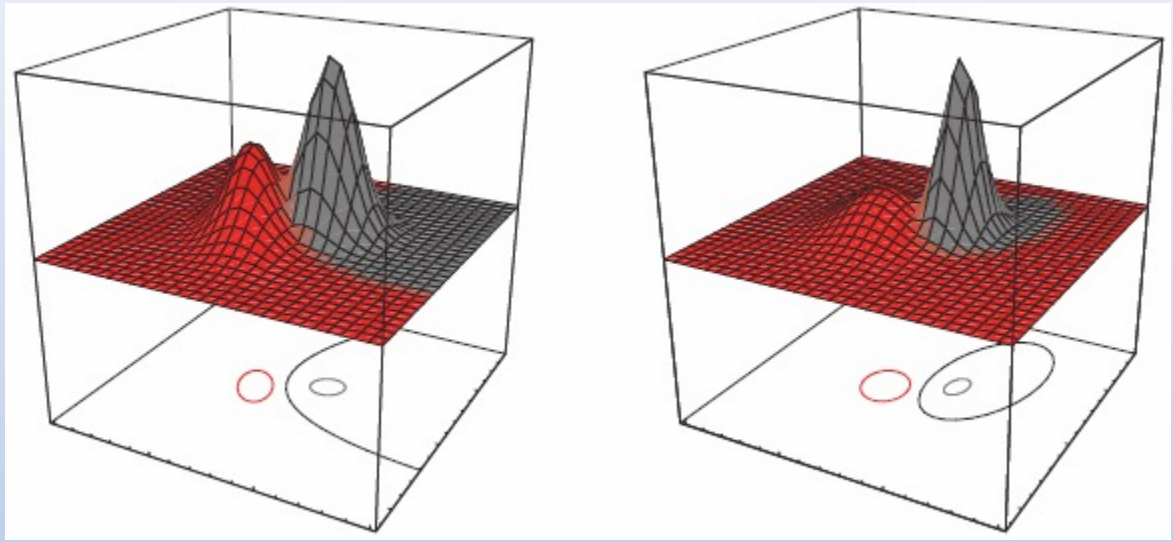
  - The decision surfaces are *hyperquadrics*, and can assume any of the general forms hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, and hyperhyperboloids of various types.

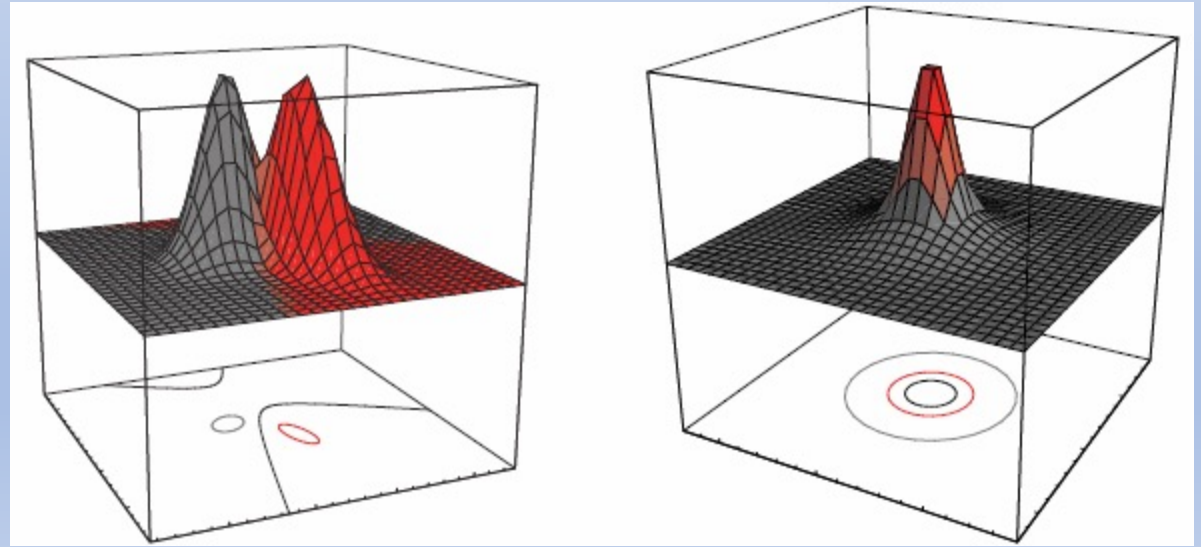
5 MAP Decision and Classifiers

- Decision boundary

  examples of quadratic classifier in

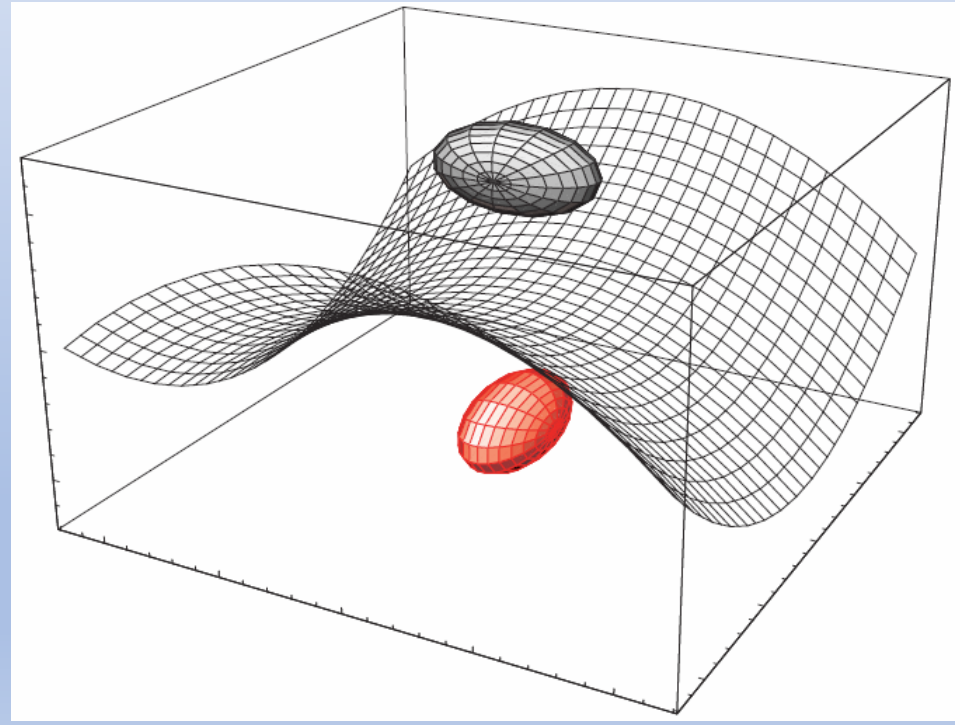
  2-dimensional, 2-class cases.

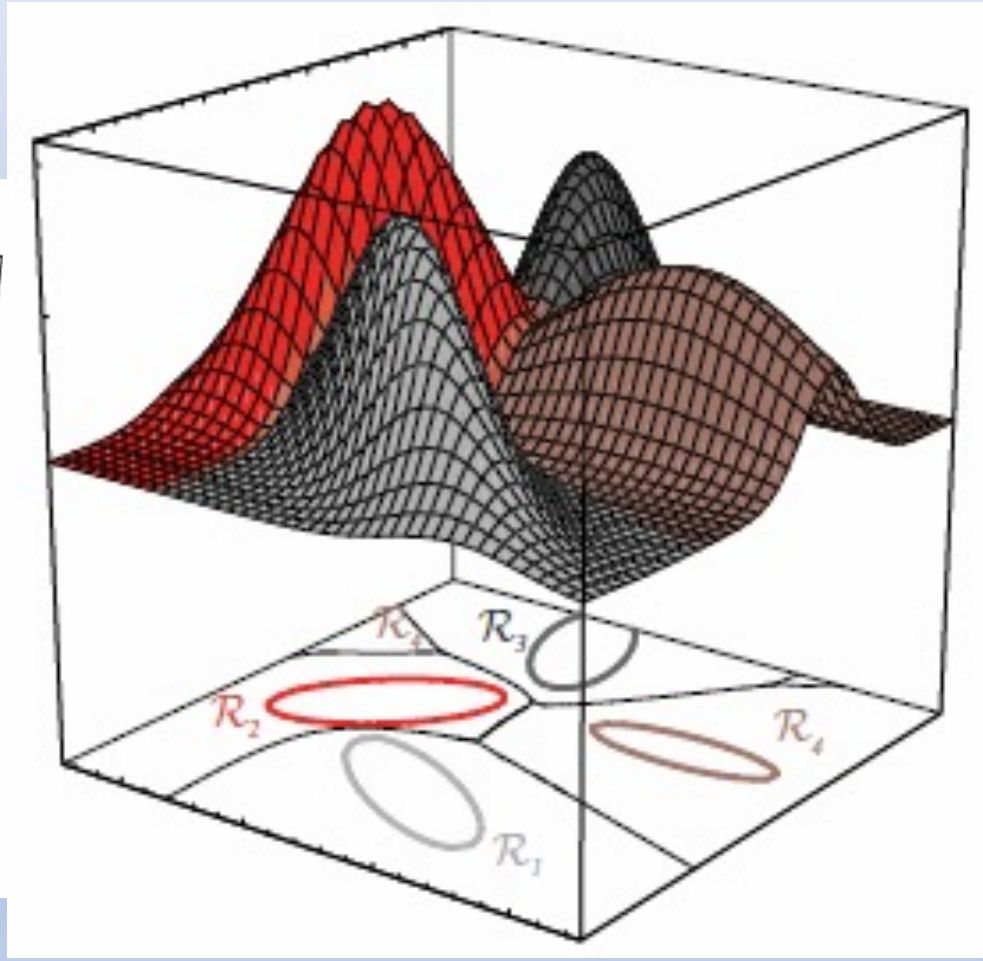




5 MAP Decision and Classifiers

- Complex decision boundary for 3-dimensional, 2-class case and 2-dimensional, multiclass problem





5 MAP Decision and Classifiers

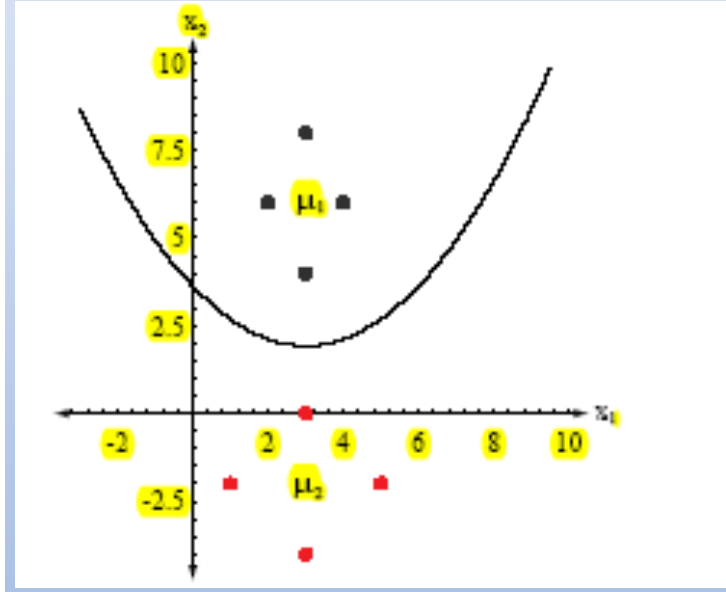
Example: Decision regions for two-dimensional Gaussian data

- Let ω_1 be the set of the four black points, and ω_2 the red points.

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} ; \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} ; \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

- The inverse matrices are then,

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \text{ and } \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$



5 MAP Decision and Classifiers

Example : Decision regions for two-dimensional Gaussian data (Cont:)

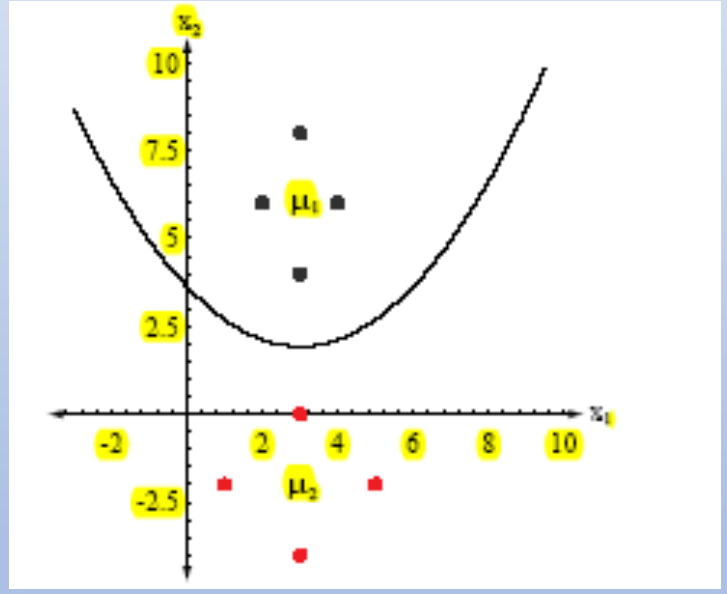
- We assume equal prior probabilities, $P(\omega_1) = P(\omega_2) = 0.5$, and substitute these into the general form for a discriminant, setting $g_1(\mathbf{x}) = g_2(\mathbf{x})$ to obtain the decision boundary:

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

- This equation describes a parabola with vertex at $\begin{pmatrix} 3 \\ 1.83 \end{pmatrix}$ and the decision boundary does not pass through the point $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$

5 MAP Decision and Classifiers

Example : Decision regions for two-dimensional Gaussian data (Cont:)



The computed Bayes decision boundary for two Gaussian distributions, each based on four data points.

5 MAP Decision and Classifiers

- Special case 1: all classes own the same covariance matrix $\Sigma_i=\Sigma$.
- The discriminant functions are simplified as

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p(\omega_i)$$

$$= -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p(\omega_i)$$

- and drop terms that are independent to the class label i we have:

$$g_i(\mathbf{x}) = \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p(\omega_i)$$

$$= \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

- The discriminant function is a linear function of $\mathbf{x}$. This results in a linear classifier as we will show that the decision or classification boundary between any two classes is a hyperplane. We call w_{i0} the threshold or bias for the class ω_i.

5 MAP Decision and Classifiers

- The decision or classification boundary between any two classes is the solution of the equation.

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \Rightarrow (\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} + w_{i0} - w_{j0} = 0$$

$$\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j\right) + \ln[p(\omega_i)/p(\omega_j)] = 0$$

- It is a straight line in 2D space $\mathbf{x}$, a plane in 3D space $\mathbf{x}$ and a hyperplane in higher dimensional space $\mathbf{x}$.
- We see that the decision hyper plane that separates two classes is generally not orthogonal to the line between the means. (in what case is it orthogonal?)

5 MAP Decision and Classifiers

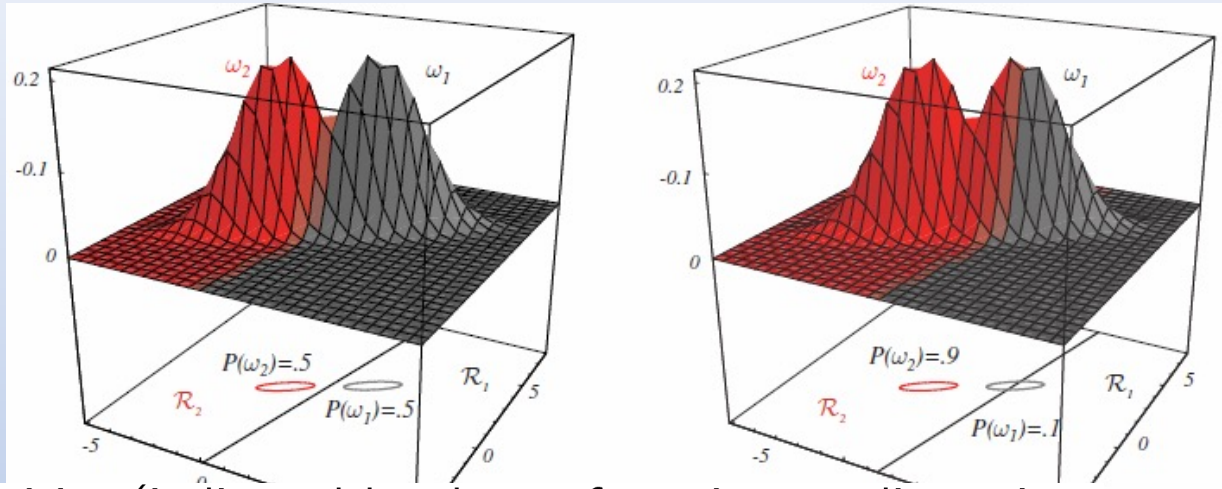
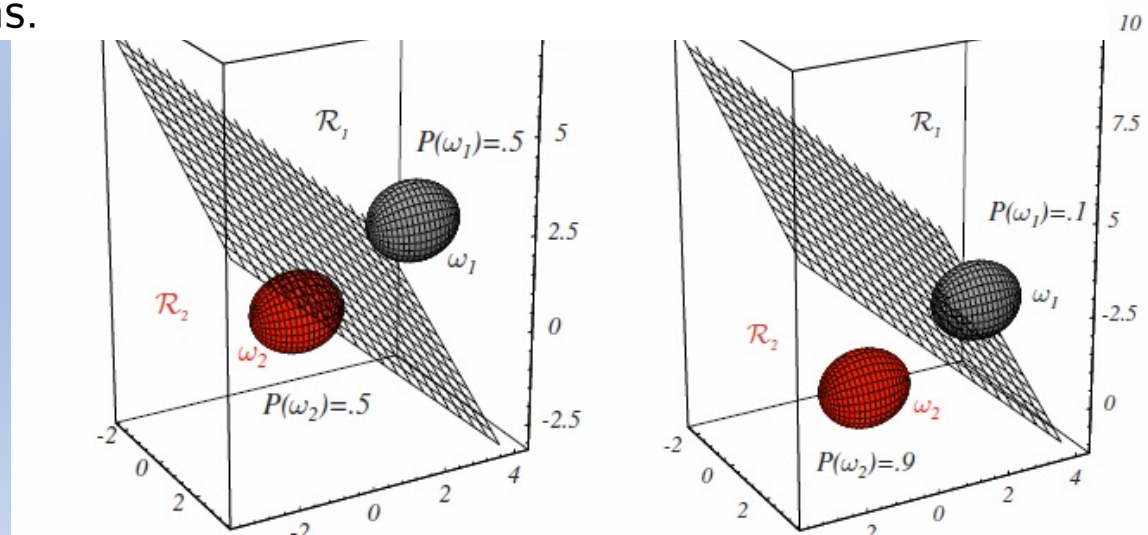


Figure 2.12: Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means.



5 MAP Decision and Classifiers

Example:

Given pattern vectors $(1,2)^t, (2,2)^t, (3,1)^t, (3,2)^t$, and $(2,3)^t$, are known to be in class ω_1. Another set of vectors, $(7,9)^t, (8,9)^t, (9,8)^t, (9,9)^t$, and $(8,10)^t$, are known to be in class ω_2. Find the Bayes classifier and the decision boundary. (assume all classes are equally likely to occur.)

Solution:

$c = 2, j = 1, 2.$

$$\mu_1 = \frac{1}{5}\left[\begin{pmatrix}1\\2\end{pmatrix} + \begin{pmatrix}2\\2\end{pmatrix} + \begin{pmatrix}3\\1\end{pmatrix} + \begin{pmatrix}3\\2\end{pmatrix} + \begin{pmatrix}2\\3\end{pmatrix}\right] = \frac{1}{5}\begin{bmatrix}11\\10\end{bmatrix}$$

$$\mu_2 = \frac{1}{5}\left[\begin{pmatrix}7\\9\end{pmatrix} + \begin{pmatrix}8\\9\end{pmatrix} + \begin{pmatrix}9\\8\end{pmatrix} + \begin{pmatrix}9\\9\end{pmatrix} + \begin{pmatrix}8\\10\end{pmatrix}\right] = \frac{1}{5}\begin{bmatrix}41\\45\end{bmatrix}$$

5 MAP Decision and Classifiers

Solution: (cont)

$$\Sigma_j = \left(\frac{1}{N_j} \sum_{\mathbf{x} \in \omega_j} \mathbf{x}\mathbf{x}^t \right) - \mu_j \mu_j^t$$

$$\Sigma_1 = \frac{1}{5} \left[\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 2 \end{pmatrix} \begin{pmatrix} 2 & 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 3 & 1 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} \begin{pmatrix} 3 & 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \end{pmatrix} \begin{pmatrix} 2 & 3 \end{pmatrix} \right]$$

$$-\frac{1}{25} \begin{pmatrix} 11 \\ 10 \end{pmatrix} \begin{pmatrix} 11 & 10 \end{pmatrix}$$

$$\Sigma_1 = \frac{1}{25} \begin{pmatrix} 14 & -5 \\ -5 & 10 \end{pmatrix}$$

Similarly, we find $\Sigma_2 = \Sigma_1 = \Sigma$

$$\Sigma^{-1} = \frac{5}{23} \begin{pmatrix} 10 & 5 \\ 5 & 14 \end{pmatrix}$$

5 MAP Decision and Classifiers

Solution: (cont)

All classes are equally likely to occur means $P(\omega_1) = P(\omega_2)$. Also, $\Sigma_1 = \Sigma_2 = \Sigma$, the decision function is simplified to

$$d_j(\mathbf{x}) = \mathbf{x}^t \Sigma^{-1} \mu_j - \frac{1}{2} \mu_j^t \Sigma^{-1} \mu_j$$

$\Sigma^{-1}\mu_1 = \frac{1}{23}\begin{pmatrix} 160 \\ 195 \end{pmatrix}$ and $\mu_1^t \Sigma^{-1} \mu_1 = \frac{742}{23}$ $\quad g_i(\mathbf{x}) = -(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$

$$d_1(\mathbf{x}) = \frac{160}{23} x_1 + \frac{195}{23} x_2 - 16.13$$

Similarly, $\Sigma^{-1}\mu_2 = \frac{1}{23}\begin{pmatrix} 635 \\ 835 \end{pmatrix}$ and $\mu_2^t \Sigma^{-1} \mu_2 = 553.12$

$$d_2(\mathbf{x}) = \frac{635}{23} x_1 + \frac{835}{23} x_2 - 276.56$$

5 MAP Decision and Classifiers

Solution: (cont) The decision matrix is

$$\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} \frac{160}{23} & \frac{195}{23} & -16.13 \\ \frac{635}{23} & \frac{835}{23} & -276.56 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

- To find the decision boundary, $d_{12}(\mathbf{x}) = d_1(\mathbf{x}) - d_2(\mathbf{x})$
- The decision rule is : decide ω_1 if $d_{12} > 0$, otherwise, decide ω_2.

5 MAP Decision and Classifiers

- Special case 2: all classes own the same diagonal, scalar covariance matrix $\Sigma_i=\sigma^2\mathbf{I}$. $\mathbf{I}$ is the identity matrix.
- This case occurs when all features (components of $\mathbf{x}$) are statistically uncorrelated and each feature has the same variance, σ^2.
- Note that $$\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I} \;\Rightarrow\; \boldsymbol{\Sigma}_i^{-1} = \mathbf{I}/\sigma^2, \quad |\boldsymbol{\Sigma}_i| = \sigma^{2d}$$
- The discriminant functions are simplified as

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)+\ln p(\omega_i)$$

$$= -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i)+\ln p(\omega_i)$$

$$= -\frac{1}{2\sigma^2}(\mathbf{x}^T\mathbf{x}-2\boldsymbol{\mu}_i{}^T\mathbf{x}+\boldsymbol{\mu}_i{}^T\boldsymbol{\mu}_i)+\ln p(\omega_i)$$

5 MAP Decision and Classifiers

- The quadratic term $\mathbf{x}^T\mathbf{x}$ is the same for all class i and hence can be dropped. Multiplying the constant σ^2, the discriminant functions are simplified as

$$g_i(\mathbf{x}) = \boldsymbol{\mu}_i^T \mathbf{x} - \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i / 2 + \sigma^2 \ln p(\omega_i)$$

$$= \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

- We call w_{i0} the threshold or bias for the class ω_i. it is a linear function of $\mathbf{x}$. This results a linear classifier.
- The decision or classification boundary between any two classes is the solution of the equation.

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \Rightarrow (\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} + w_{i0} - w_{j0} = 0$$

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \mathbf{x} - (\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j) / 2 + \sigma^2 \ln[p(\omega_i) / p(\omega_j)] = 0$$

- This equation defines a hyperplane orthogonal to the line linking the two means.

5 MAP Decision and Classifiers

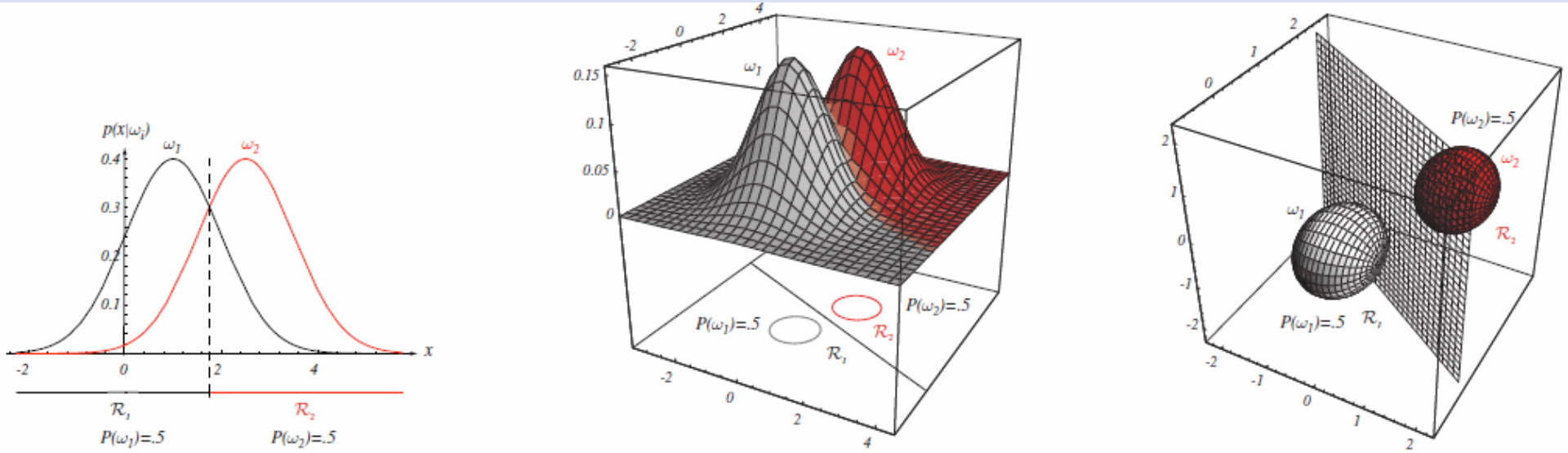
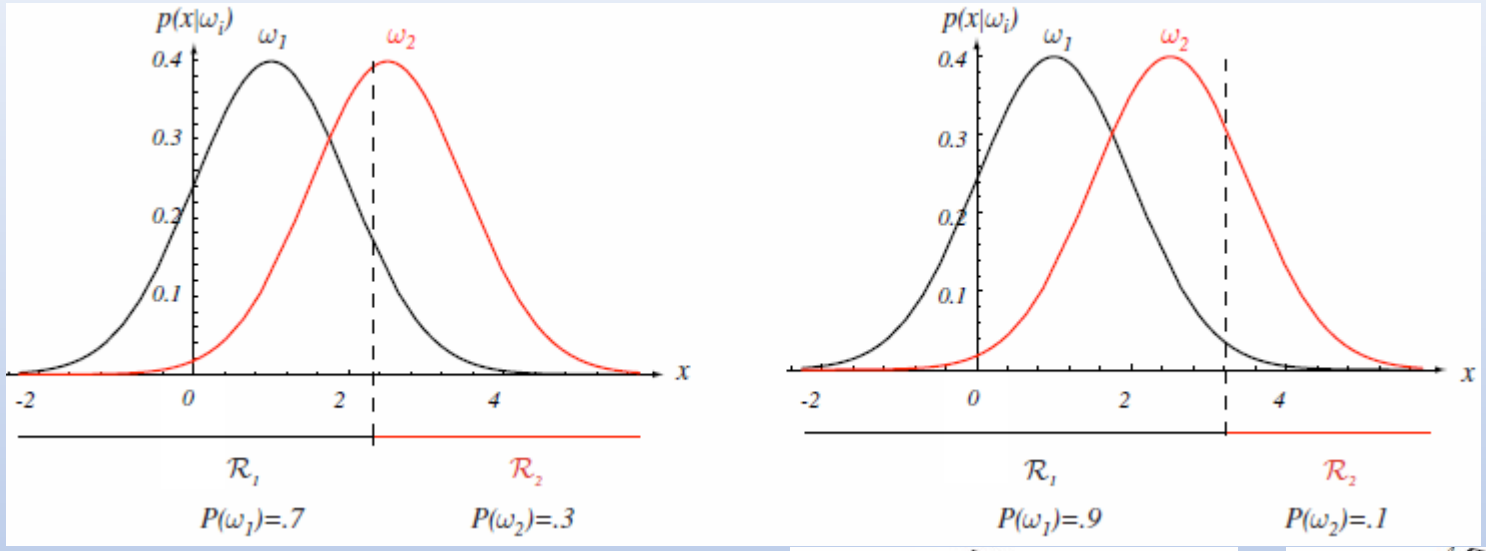


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern*

5 MAP Decision and Classifiers



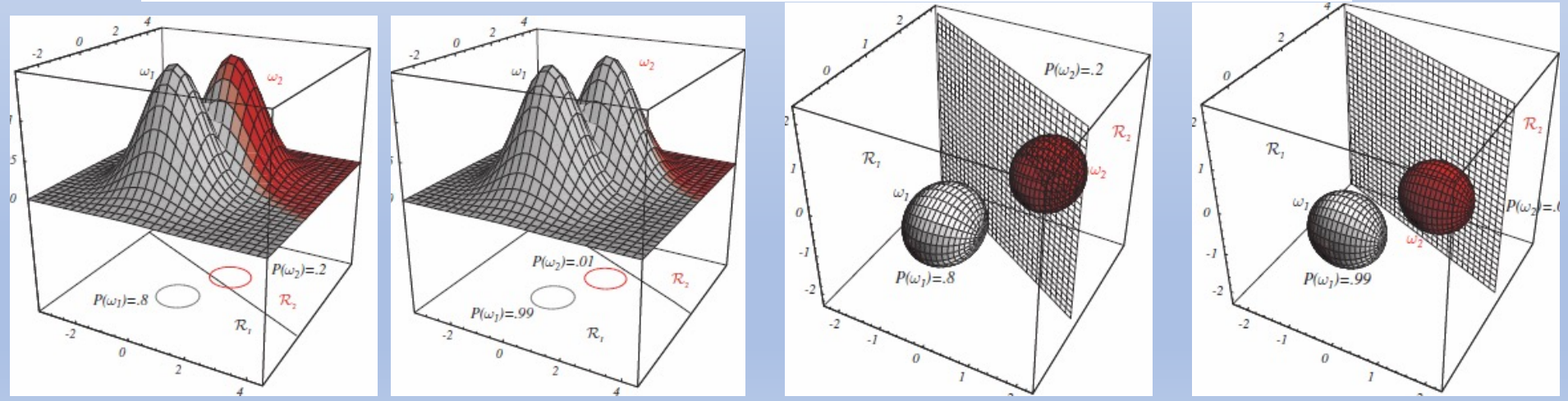


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E.

5 MAP Decision and Classifiers

- Recall the discriminant function

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p(\omega_i)$$

- If the prior probabilities are the same for all classes, then

$$g_i(\mathbf{x}) = -(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i) = -d_{Eu}(\mathbf{x},\boldsymbol{\mu}_i) = \| \mathbf{x}-\boldsymbol{\mu}_i \|$$

- When this happens, the optimum decision rule can be stated very simply: to classify a feature vector $\mathbf{x}$, measure the Euclidean distance from $\mathbf{x}$ to each of the c mean vectors, and assign $\mathbf{x}$ to the category of the nearest mean. Such a classifier is called a minimum distance classifier. If each mean vector is thought of as being an ideal prototype or template for patterns in its class, then this is essentially a template matching procedure.

5 MAP Decision and Classifiers

- We have derived several classifiers under Gaussian assumption.
- Gaussian PDF is the most natural and most common distribution.
- For Non-Gaussian distribution, it is very difficult if not impossible to derive a theoretical optimal classifier.
- One way to solve non-Gaussian data distribution is to perform some proper feature transform to convert it into Gaussian PDF. Example:

J. Ren, X.D. Jiang and J. Yuan, "A Chi-Squared-Transformed Subspace of LBP Histogram for Visual Recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1893-1904, June, 2015.