Lecture 1

Chapter 2

PATTERN RECOGNITION

- ❖ Typical application areas
  - ➢ Machine vision
  - ➢ Character recognition (OCR)
  - ➢ Computer aided diagnosis
  - ➢ Speech recognition
  - ➢ Face recognition
  - ➢ Biometrics
  - ➢ Image Data Base retrieval
  - ➢ Data mining
  - ➢ Bionformatics

- ❖ The task: Assign unknown objects – patterns – into the correct class. This is known as classification.

❖ Features: These are measurable quantities obtained from the patterns, and the classification task is based on their respective values.

❖Feature vectors: A number of features

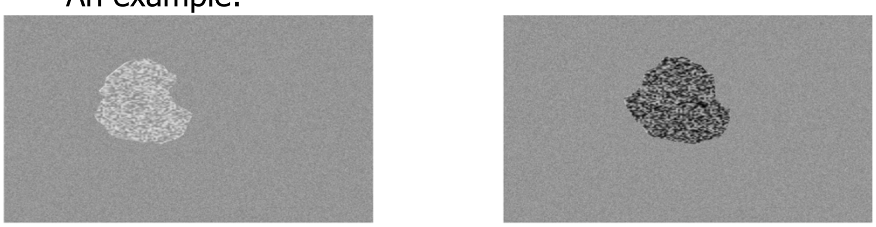
$$x_1, \ldots, x_l,$$

constitute the feature vector

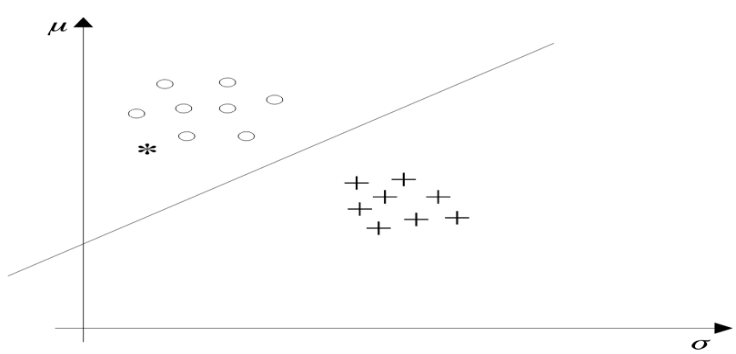
$$\underline{x} = [x_1, \ldots, x_l]^T \in R^l$$

Feature vectors are treated as random vectors.

An example:

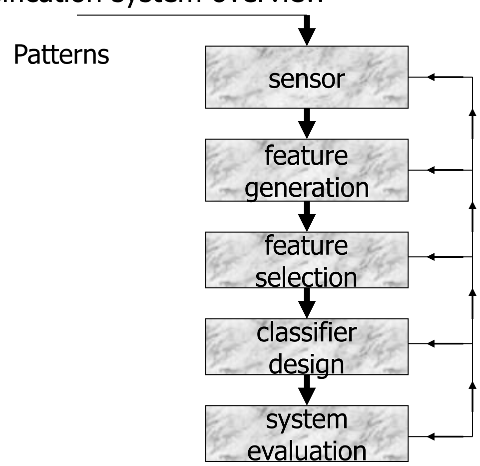


(a) (b)



❖ The classifier consists of a set of functions, whose values, computed at $\underline{x}$, determine the class to which the corresponding pattern belongs

❖ Classification system overview



5

❖ Supervised – unsupervised – semisupervised pattern recognition:
The major directions of learning are:

- **Supervised**: Patterns whose class is known a-priori are used for training.
- **Unsupervised**: The number of classes/groups is (in general) unknown and no training patterns are available.
- **Semisupervised:** A mixed type of patterns is available. For some of them, their corresponding class is known and for the rest is not.

6

CLASSIFIERS BASED ON BAYES DECISION THEORY

❖ Statistical nature of feature vectors

$$\underline{x} = [x_1, x_2, \ldots, x_l]^T$$

❖ Assign the pattern represented by feature vector $\underline{x}$ to the most probable of the available classes

$$\omega_1, \omega_2, \ldots, \omega_M$$

That is $\underline{x} \longrightarrow \omega_i : P(\omega_i | \underline{x})$ maximum

7

❖ Computation of a-posteriori probabilities

- Assume known
  - a-priori probabilities

    $$P(\omega_1), P(\omega_2) \ldots, P(\omega_M)$$

  - $p(\underline{x}|\omega_i), i = 1,2,\ldots,M$

This is also known as the likelihood of

$$\underline{x} \text{ w.r. to } \omega_i.$$

8

➢ The Bayes rule (M=2)

$$p(\underline{x})P(\omega_i|\underline{x}) = p(\underline{x}|\omega_i)P(\omega_i) \Rightarrow$$

$$P(\omega_i|\underline{x}) = \frac{p(\underline{x}|\omega_i)P(\omega_i)}{p(\underline{x})}$$

where

$$p(\underline{x}) = \sum_{i=1}^{2} p(\underline{x}|\omega_i)P(\omega_i)$$

9

❖ The Bayes classification rule (for two classes M=2)

➢ Given $\underline{x}$ classify it according to the rule

$$\boxed{\begin{array}{l} If\ P(\omega_1|\underline{x}) > P(\omega_2|\underline{x})\ \ \underline{x} \rightarrow \omega_1 \\ If\ P(\omega_2|\underline{x}) > P(\omega_1|\underline{x})\ \ \underline{x} \rightarrow \omega_2 \end{array}}$$

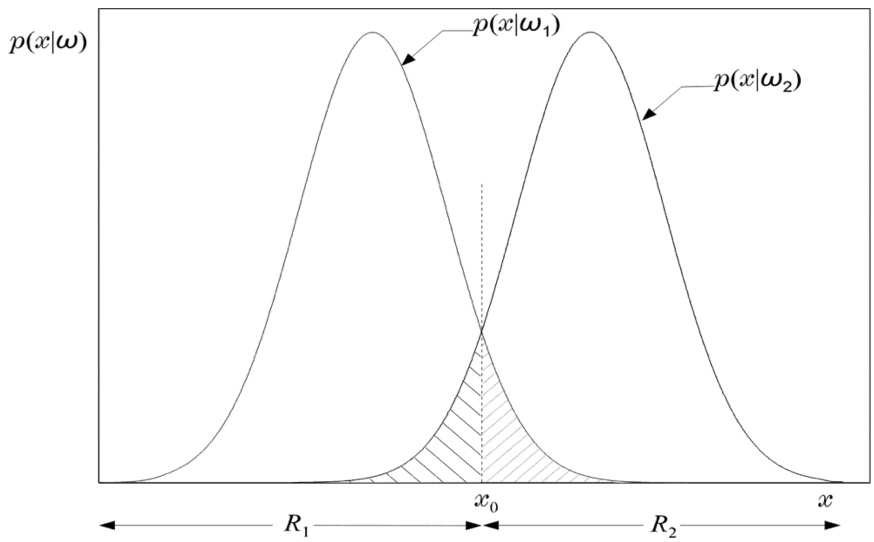
➢ Equivalently: classify $\underline{x}$ according to the rule

$$\boxed{p(\underline{x}|\omega_1)P(\omega_1)(><)p(\underline{x}|\omega_2)P(\omega_2)}$$

➢ For equiprobable classes the test becomes

$$\boxed{p(\underline{x}|\omega_1)(><)p(\underline{x}|\omega_2)}$$

10



$$R_1(\rightarrow \omega_1) \ and \ R_2(\rightarrow \omega_2)$$

11

❖ Equivalently in words: Divide space in two regions

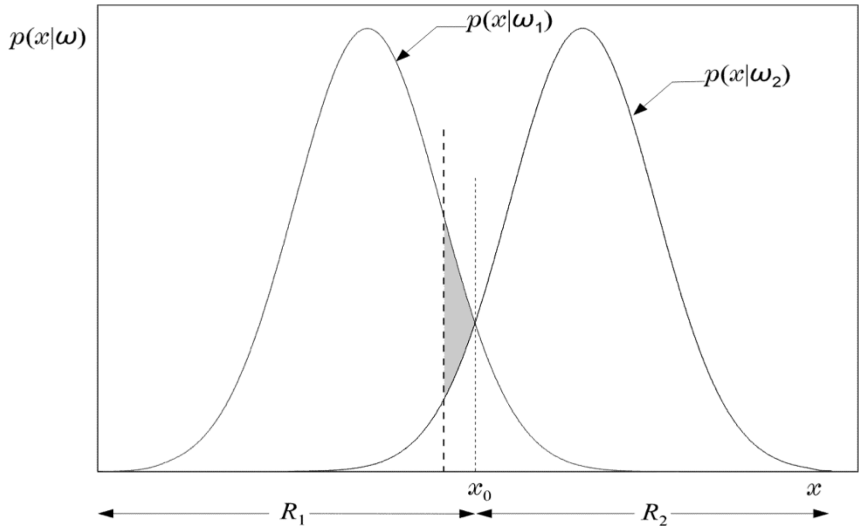
$$\boxed{\begin{array}{l} \text{If } \underline{x} \in R_1 \Rightarrow \underline{x} \text{ in } \omega_1 \\ \text{If } \underline{x} \in R_2 \Rightarrow \underline{x} \text{ in } \omega_2 \end{array}}$$

❖ Probability of error

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

❖ Bayesian classifier is OPTIMAL with respect to minimising the classification error probability!!!!

12



- Indeed: Moving the threshold the total shaded area INCREASES by the extra "grey" area.

13

- ❖ The Bayes classification rule for many (M>2) classes:
  - Given $\underline{x}$ classify it to ω_i if:

$$\boxed{P(\omega_i|\underline{x}) > P(\omega_j|\underline{x}) \quad \forall j \neq i}$$

  - Such a choice also minimizes the classification error probability

- ❖ Minimizing the average risk
  - For each wrong decision, a penalty term is assigned since some decisions are more sensitive than others

14

- For M=2
  - Define the loss matrix

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

11 - cost of deciding class 1 when the true class is indeed 1.

- λ_{12} penalty term for deciding class ω_2 , although the pattern belongs to ω_1 , etc.

- Risk with respect to ω_1

$$r_1 = \lambda_{11} \int_{R_1} p(\underline{x}|\omega_1) d\underline{x} + \boxed{\lambda_{12} \int_{R_2} p(\underline{x}|\omega_1) d\underline{x}}$$

15

- Risk with respect to ω_2

$$r_2 \boxed{= \lambda_{21} \int_{R_1} p(\underline{x}|\omega_2) d\underline{x}} + \lambda_{22} \int_{R_2} p(\underline{x}|\omega_2) d\underline{x}$$

- $\boxed{} \Rightarrow$ Probabilities of wrong decisions, weighted by the penalty terms

- Average risk

$$r = r_1 P(\omega_1) + r_2 P(\omega_2)$$

16

❖ Choose R_1 and R_2 so that r is minimized

❖ Then assign $\underline{x}$ to ω_i if

$$\ell_1 \equiv \lambda_{11} p(\underline{x}|\omega_1)P(\omega_1) + \lambda_{21} p(\underline{x}|\omega_2)P(\omega_2) \quad <$$

$$\ell_2 \equiv \lambda_{12} p(\underline{x}|\omega_1)P(\omega_1) + \lambda_{22} p(\underline{x}|\omega_2)P(\omega_2)$$

❖ Equivalently:
assign $\underline{x}$ in $\omega_1(\omega_2)$ if

$$\boxed{\ell_{12} \equiv \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} > (<) \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}}$$

ℓ_{12} : likelihood ratio

17

❖ If $P(\omega_1) = P(\omega_2) = \dfrac{1}{2}$ and $\lambda_{11} = \lambda_{22} = 0$

$$\boxed{\begin{array}{l} \underline{x} \rightarrow \omega_1 \text{ if } P(\underline{x}|\omega_1) > P(\underline{x}|\omega_2)\dfrac{\lambda_{21}}{\lambda_{12}} \\ \underline{x} \rightarrow \omega_2 \text{ if } P(\underline{x}|\omega_2) > P(\underline{x}|\omega_1)\dfrac{\lambda_{12}}{\lambda_{21}} \end{array}}$$

if $\lambda_{21} = \lambda_{12} \Rightarrow$ Minimum classification error probability

18

❖ An example:

- $p(x|\omega_1) = \dfrac{1}{\sqrt{\pi}} \exp(-x^2)$
- $p(x|\omega_2) = \dfrac{1}{\sqrt{\pi}} \exp(-(x-1)^2)$
- $P(\omega_1) = P(\omega_2) = \dfrac{1}{2}$
- $L = \begin{pmatrix} 0 & 0.5 \\ 1.0 & 0 \end{pmatrix}$

19

➢ Then the threshold value is:

$$x_0 \text{ for minimum } P_e :$$

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

$$x_0 = \frac{1}{2}$$

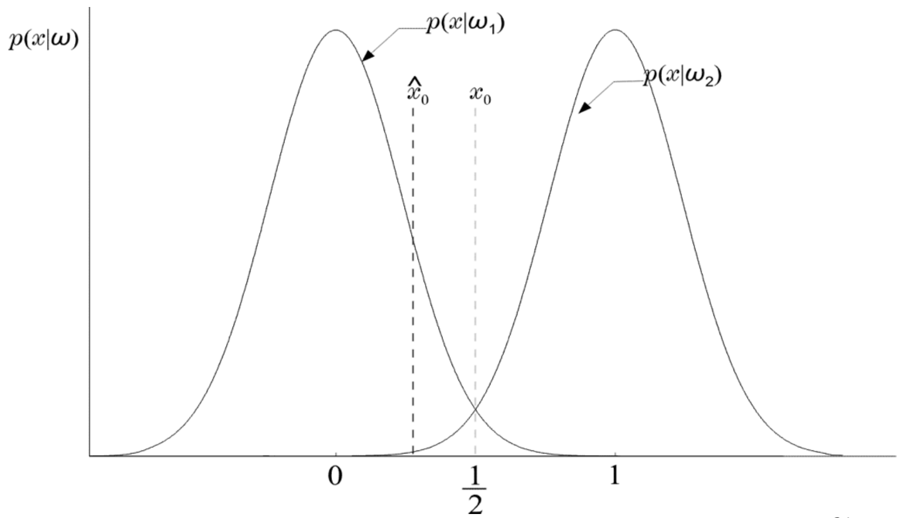
➢ Threshold $\hat{x}_0$ for minimum r

$$\hat{x}_0 : \exp(-x^2) = 2\exp(-(x-1)^2) \Rightarrow$$

$$\hat{x}_0 = \frac{(1-\ell n 2)}{2} < \frac{1}{2}$$

20

Thus $\hat{x}_0$ moves to the left of $\frac{1}{2} = x_0$
(WHY?)



21

DISCRIMINANT FUNCTIONS
DECISION SURFACES

❖ If R_i, R_j are contiguous: $g(\underline{x}) \equiv P(\omega_i|\underline{x}) - P(\omega_j|\underline{x}) = 0$

$$R_i: \ P(\omega_i|\underline{x}) > P(\omega_j|\underline{x})$$

+

\-

$$g(\underline{x}) = 0$$

$$R_j: \ P(\omega_j|\underline{x}) > P(\omega_i|\underline{x})$$

is the surface separating the regions. On the one side is positive (+), on the other is negative (-). It is known as Decision Surface.

22

- If $f(.)$ monotonically increasing, the rule remains the same if we use:

  $$\underline{x} \rightarrow \omega_i \text{ if}: \ f(P(\omega_i|\underline{x})) > f(P(\omega_j|\underline{x})) \ \forall i \neq j$$

- $g_i(\underline{x}) \equiv f(P(\omega_i|\underline{x}))$ is a **discriminant function.**

- In general, discriminant functions may be defined independently of the Bayesian rule ➡ They lead to suboptimal solutions
- If chosen appropriately, they can be computationally more tractable.
- Moreover, in practice, they may also lead to better solutions.
- Sometimes the case when the underlying pdf's are unknown.

23

THE GAUSSIAN DISTRIBUTION

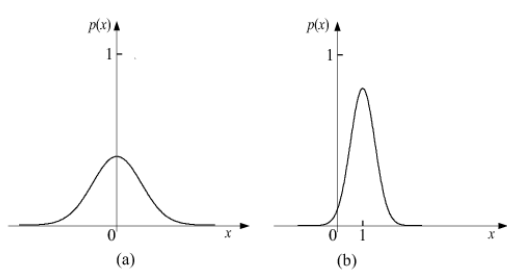
- The one-dimensional case

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

  where

  μ is the mean value, i.e.: $\mu = E[x] = \int_{-\infty}^{+\infty} x p(x) dx$

  σ^2 is the variance, $\sigma^2 = E\left[(x - E[x])^2\right] = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx$



24

❖ The Multivariate (Multidimensional) case:

$$p(\underline{x}) = \frac{1}{(2\pi)^{\frac{\ell}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right)$$

where $\underline{\mu}$ is the mean value, $\underline{\mu} = E[\underline{x}]$

and Σ is known s the covariance matrix and it is defined as:

$$\Sigma = E[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T]$$

❖ **An example**: The two-dimensional case:

$$p(\underline{x}) = p(x_1, x_2) = \frac{1}{(2\pi)|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} [x_1 - \mu_1, x_2 - \mu_2] \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} E[x_1] \\ E[x_2] \end{bmatrix}, \quad \Sigma = E\left[\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} [x_1 - \mu_1, x_2 - \mu_2] \right] = \begin{bmatrix} \sigma_1^2 & \sigma \\ \sigma & \sigma_2^2 \end{bmatrix}$$

where $\sigma = E[(x_1 - \mu_1)(x_2 - \mu_2)]$

25

BAYESIAN CLASSIFIER FOR NORMAL DISTRIBUTIONS

❖ Multivariate Gaussian pdf

$$p(\underline{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{\ell}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\underline{x} - \underline{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i) \right)$$

$\underline{\mu}_i = E[\underline{x}]$ is an $\ell \times 1$ vector, for $\underline{x} \in \omega_i$

$$\Sigma_i = E\left[(\underline{x} - \underline{\mu}_i)(\underline{x} - \underline{\mu}_i)^{\mathrm{T}} \right]$$

is the $\ell \times \ell$ covariance matrix.

26

❖ $\ln(\cdot)$ is monotonic. Define:

- $g_i(\underline{x}) = \ln(p(\underline{x}|\omega_i)P(\omega_i)) =$
  $\ln p(\underline{x}|\omega_i) + \ln P(\omega_i)$
- $g_i(\underline{x}) = -\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x}-\underline{\mu}_i) + \ln P(\omega_i) + C_i$

  $C_i = -(\frac{\ell}{2})\ln 2\pi - (\frac{1}{2})\ln|\Sigma_i|$
- **Example:** $\Sigma_i = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$

27

- $g_i(\underline{x}) = -\frac{1}{2\sigma^2}(x_1^2 + x_2^2) + \frac{1}{\sigma^2}(\mu_{i1}x_1 + \mu_{i2}x_2)$
  $-\frac{1}{2\sigma^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln(P\omega_i) + C_i$

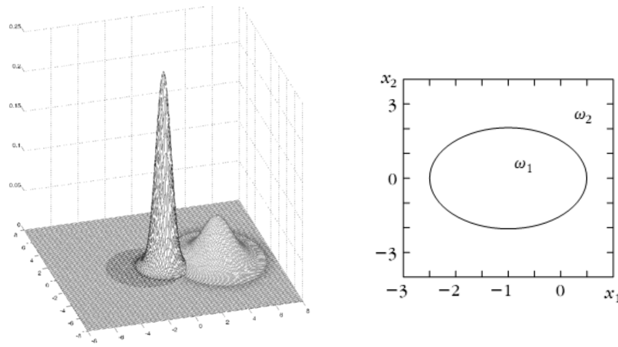
  That is, $g_i(x)$ is quadratic and the surfaces

  $$g_i(\underline{x}) - g_j(\underline{x}) = 0$$

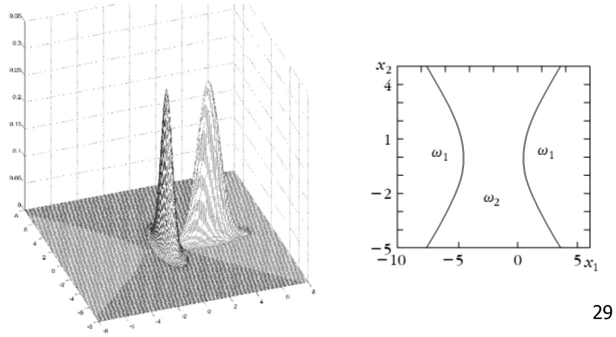
  quadrics, ellipsoids, parabolas, hyperbolas, pairs of lines.

28

❖ **Example 1:**



❖ **Example 2:**



❖ Decision Hyperplanes

➢ Quadratic terms: $\underline{x}^T \Sigma_i^{-1} \underline{x}$

If ALL $\Sigma_i = \Sigma$ (the same) the quadratic terms are not of interest. They are not involved in comparisons. Then, equivalently, we can write:

$$\boxed{\begin{aligned} g_i(\underline{x}) &= \underline{w}_i^T \underline{x} + w_{io} \\ \underline{w}_i &= \Sigma^{-1} \underline{\mu}_i \\ w_{i0} &= \ln P(\omega_i) - \frac{1}{2} \underline{\mu}^T{}_i \Sigma^{-1} \underline{\mu}_i \end{aligned}}$$

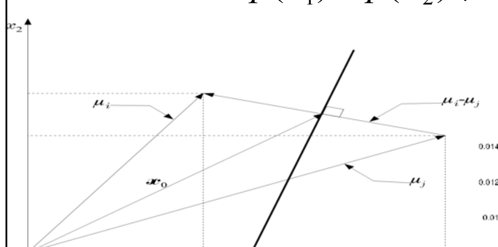
Discriminant functions are LINEAR.

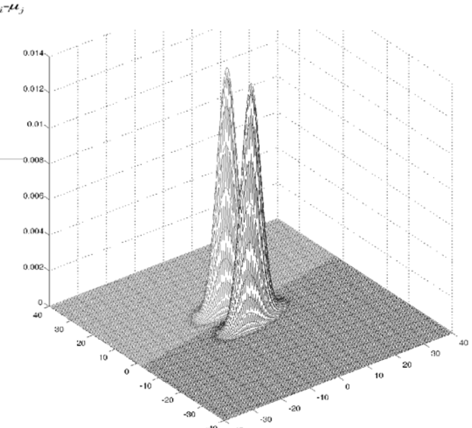
- Let in addition:
  - $\Sigma = \sigma^2 I$. Then

    $$g_i(\underline{x}) = \frac{1}{\sigma^2}\underline{\mu}_i^T\,\underline{x} + w_{i0}$$
  - $$g_{ij}(\underline{x}) = g_i(\underline{x}) - g_j(\underline{x}) = 0$$
    $$= \underline{w}^T(\underline{x} - \underline{x}_o)$$
  - $$\underline{w} = \underline{\mu}_i - \underline{\mu}_j,$$
  - $$\underline{x}_o = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \sigma^2 \ln\frac{P(\omega_i)}{P(\omega_j)}\frac{\underline{\mu}_i - \underline{\mu}_j}{\left\|\underline{\mu}_i - \underline{\mu}_j\right\|^2}$$

31

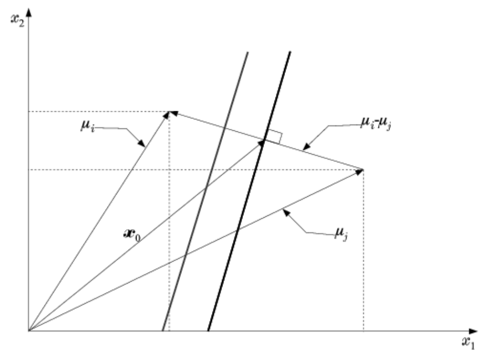
- Remark :
  - If $p(\omega_1) = p(\omega_2)$, then $\underline{x}_0 = \frac{1}{2}\left(\underline{\mu}_1 + \underline{\mu}_2\right)$





32

- If $p(\omega_1) \neq p(\omega_2)$, the linear classifier moves towards the class with the smaller probability



33

➢ Nondiagonal: $\Sigma \neq \sigma^2 I$

- $g_{ij}(\underline{x}) = \underline{w}^T(\underline{x} - \underline{x}_0) = 0$
- $\underline{w} = \Sigma^{-1}(\underline{\mu}_i - \underline{\mu}_j)$
- $\underline{x}_0 = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \ell\mathrm{n}(\frac{P(\omega_i)}{P(\omega_j)}) \frac{\underline{\mu}_i - \underline{\mu}_j}{\left\| \underline{\mu}_i - \underline{\mu}_j \right\|^2_{\Sigma^{-1}}}$

where

$\left\| \underline{x} \right\|_{\Sigma^{-1}} \equiv (\underline{x}^T \Sigma^{-1} \underline{x})^{\frac{1}{2}}$

➢ Decision hyperplane

not normal to $\underline{\mu}_i - \underline{\mu}_j$

normal to $\Sigma^{-1}(\underline{\mu}_i - \underline{\mu}_j)$

34

❖ Minimum Distance Classifiers

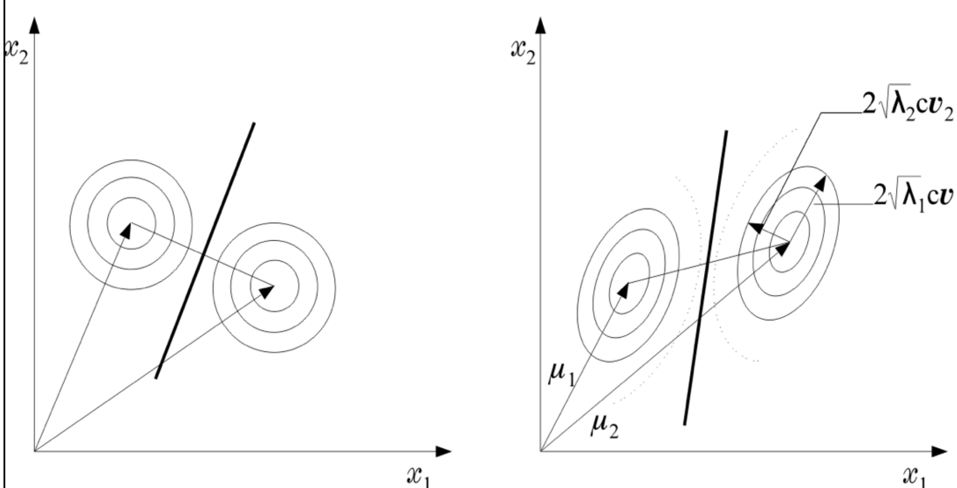
- $P(\omega_i) = \dfrac{1}{M}$ equiprobable
- $g_i(\underline{x}) = -\dfrac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_i)$
- $\Sigma = \sigma^2 I:$ Assign $\underline{x} \rightarrow \omega_i:$

  Euclidean Distance: $d_E \equiv \|\underline{x} - \underline{\mu}_i\|$
  smaller

- $\Sigma \neq \sigma^2 I:$ Assign $\underline{x} \rightarrow \omega_i:$

  Mahalanobis Distance: $d_m = ((\underline{x} - \underline{\mu}_i)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_i))^{\frac{1}{2}}$
  smaller

35



(a) (b)

36

❖ Example:

Given ω_1, ω_2 : $P(\omega_1) = P(\omega_2)$ and $p(\underline{x}|\omega_1) = N(\underline{\mu}_1, \Sigma)$,

$p(\underline{x}|\omega_2) = N(\underline{\mu}_2, \Sigma)$, $\underline{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\underline{\mu}_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$

classify the vector $\underline{x} = \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$ using Bayesian classification :

- $\Sigma^{-1} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$
  - Compute Mahalanobis d_m from μ_1, μ_2 : $d^2{}_{m,1} = [1.0, \quad 2.2]$ $\Sigma^{-1}\begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952$, $d^2{}_{m,2} = [-2.0, \quad -0.8]\, \Sigma^{-1}\begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$
- Classify $\underline{x} \rightarrow \omega_1$. Observe that $d_{E,2} < d_{E,1}$

37