

A Course on PATTERN RECOGNITION

Sergios Theodoridis
Konstantinos Koutroumbas

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

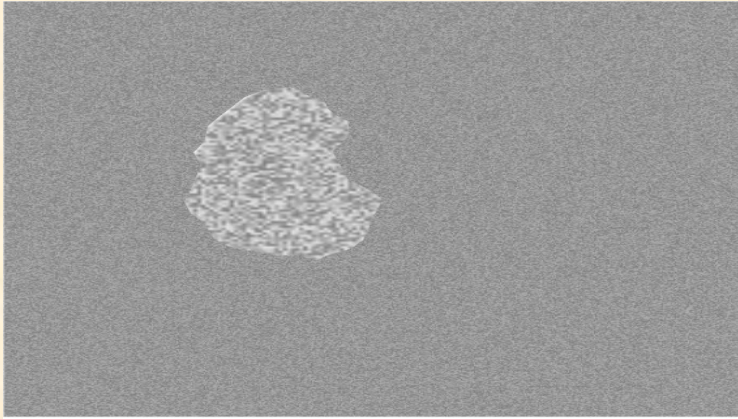
$$x_1, \dots, x_l,$$

constitute the feature vector

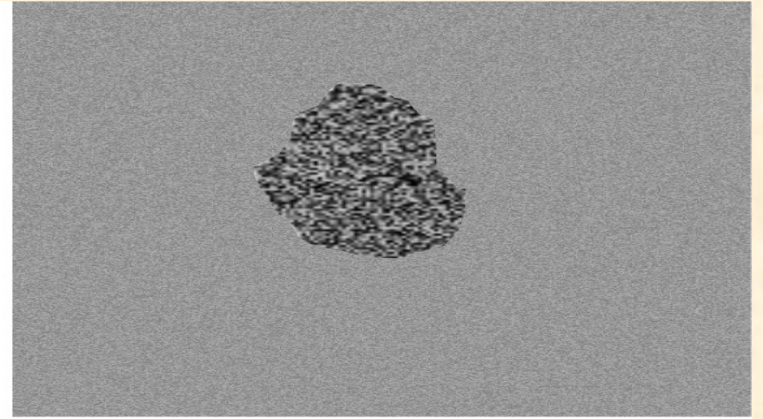
$$\underline{x} = [x_1, \dots, x_l]^T \in R^l$$

Feature vectors are treated as **random vectors**.

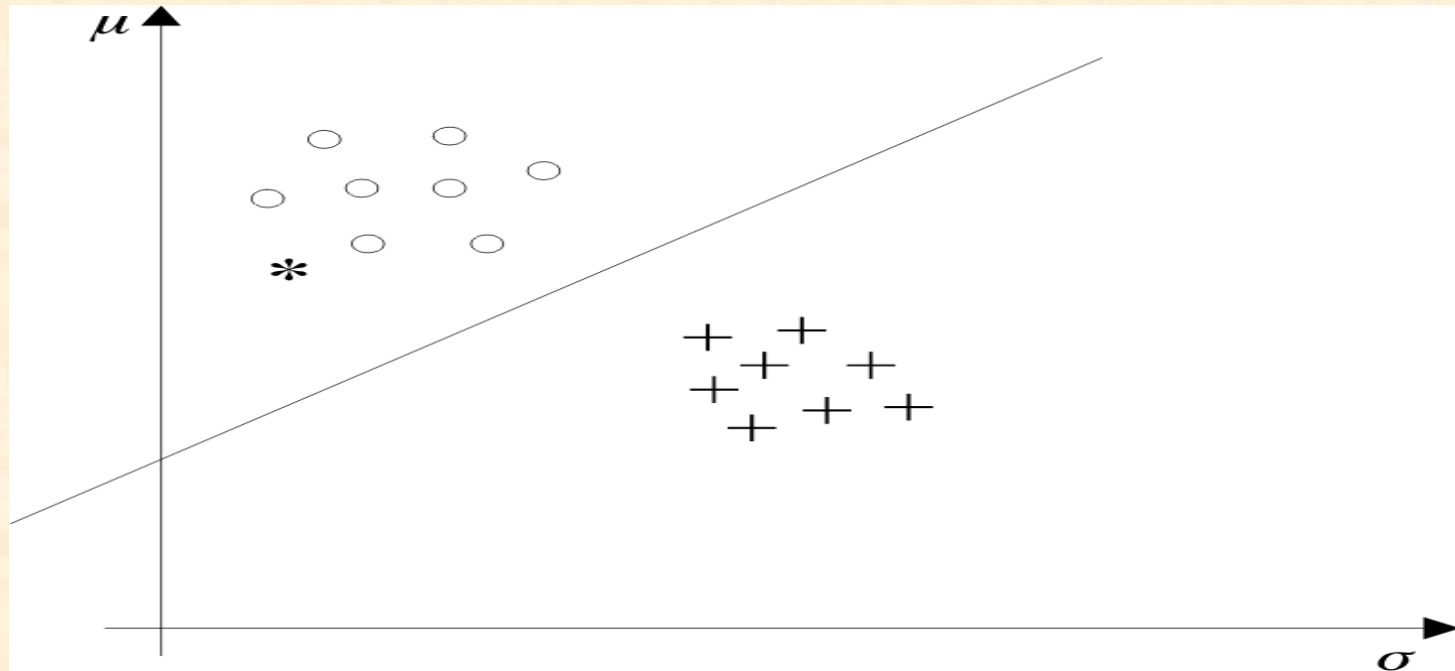
An example:



(a)

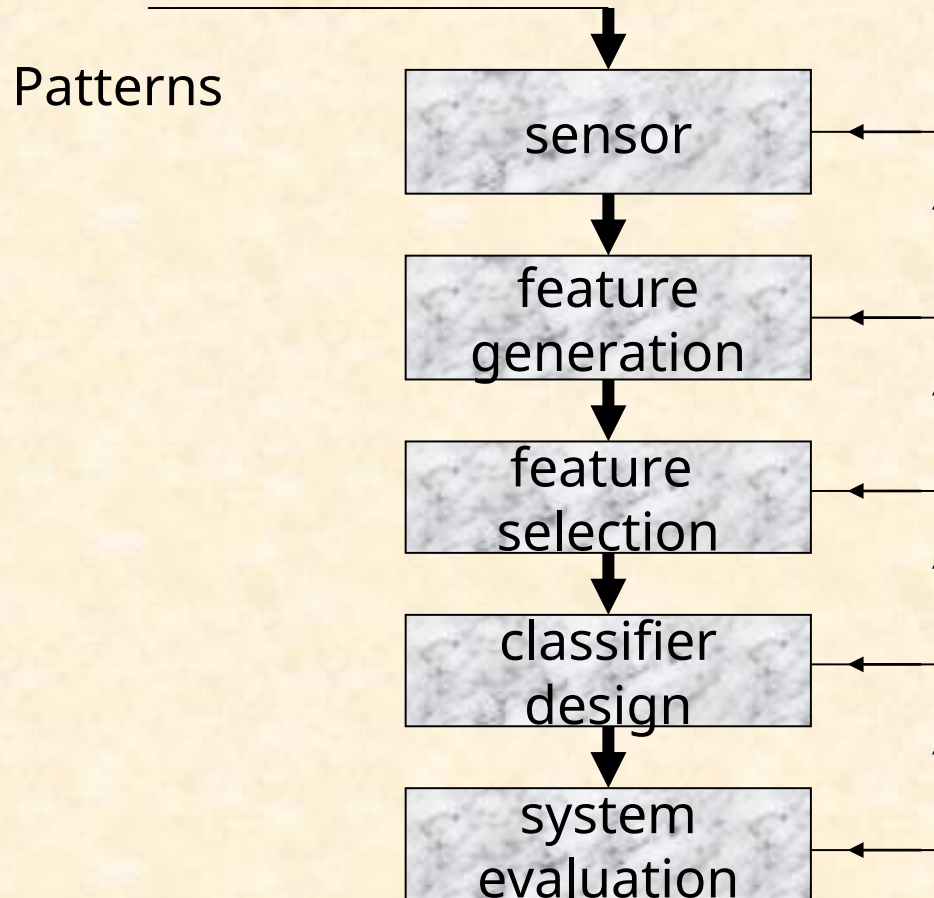


(b)



- ❖ The **classifier** consists of a **set of functions**, whose values, computed at x , determine the class to which the corresponding pattern belongs

- ❖ Classification system overview



❖ Supervised – unsupervised pattern recognition:

The two major directions

- **Supervised:** Patterns whose class is known a-priori are used for training.
- **Unsupervised:** The number of classes is (in general) unknown and no training patterns are available.

CLASSIFIERS BASED ON BAYES DECISION THEORY

- ❖ Statistical nature of feature vectors

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

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

$$\omega_1, \omega_2, \dots, \omega_M$$

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

❖ Computation of **a-posteriori** probabilities

➤ Assume known

- **a-priori** probabilities

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

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

This is also known as the **likelihood of**

$\underline{x}$ *w.r. to* ω_i .

- 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

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

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

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

$$\text{If } P(\omega_1|\underline{x}) > P(\omega_2|\underline{x}) \quad \underline{x} \rightarrow \omega_1$$

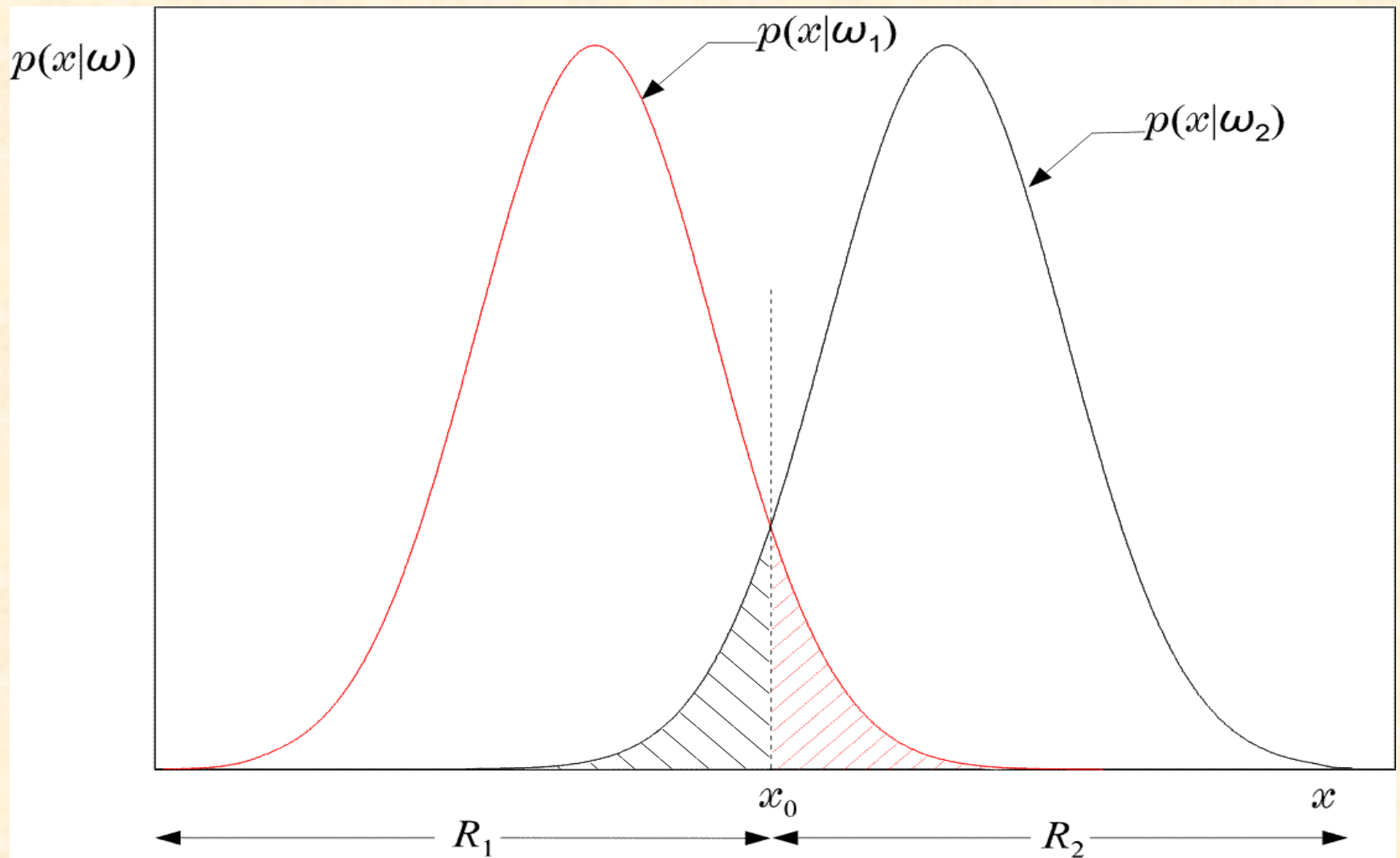
$$\text{If } P(\omega_2|\underline{x}) > P(\omega_1|\underline{x}) \quad \underline{x} \rightarrow \omega_2$$

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

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

- For equiprobable classes the test becomes

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



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

❖ Equivalently in words: Divide space in two regions

If $\underline{x} \in R_1 \Rightarrow \underline{x} \text{ in } \omega_1$

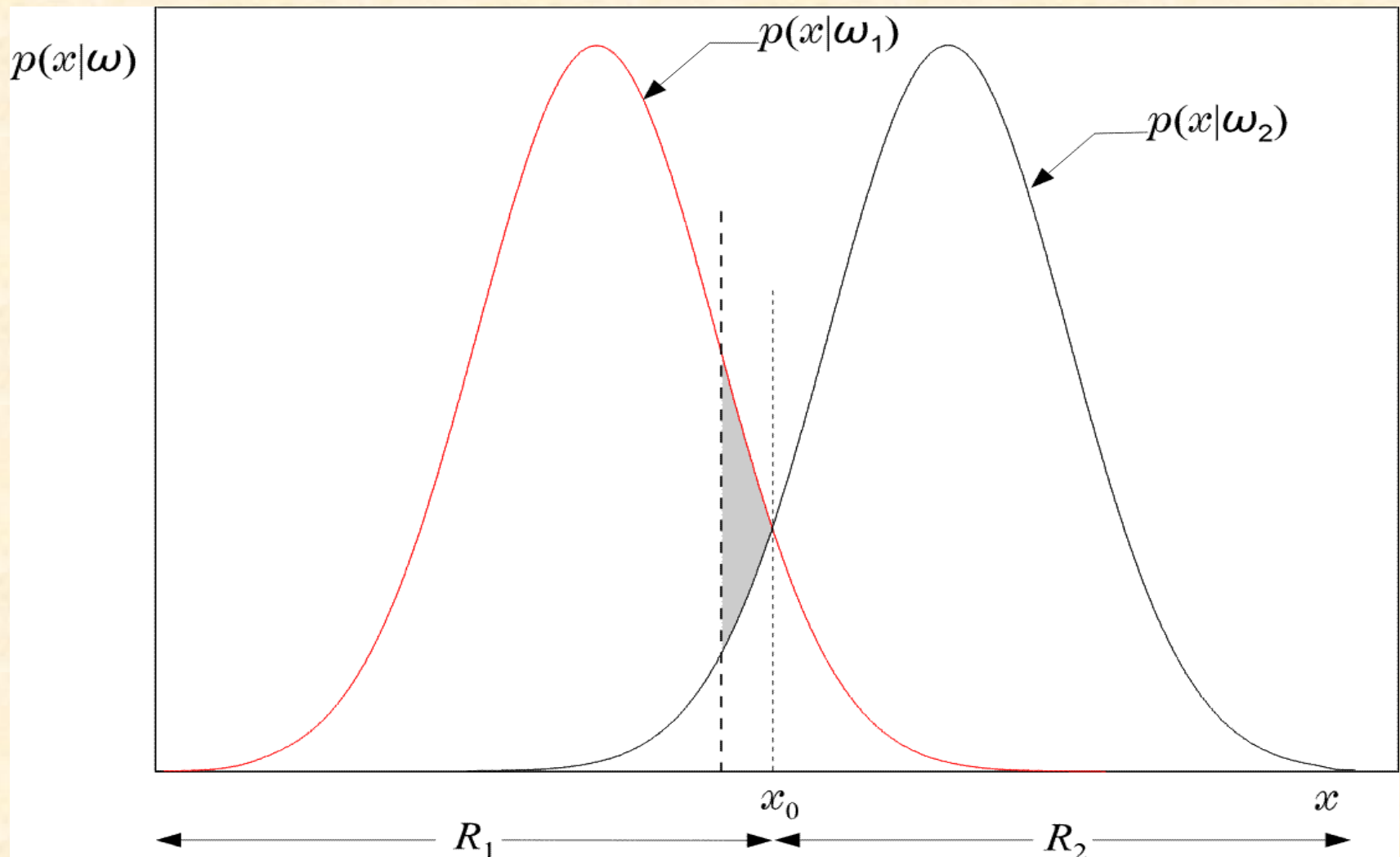
If $\underline{x} \in R_2 \Rightarrow \underline{x} \text{ in } \omega_2$

❖ Probability of error

➤ Total shaded area

$$\text{➤ } P_e = \int_{-\infty}^{x_0} p(x|\omega_2)dx + \int_{x_0}^{+\infty} p(x|\omega_1)dx$$

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



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

❖ The Bayes classification rule for many ($M > 2$) classes:

- Given $\underline{x}$ classify it to ω_i if:

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

- 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

➤ For $M=2$

- Define the **loss matrix**

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

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

➤ Risk with respect to ω_1

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

- Risk with respect to ω_2

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

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

- Average risk

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

❖ Choose R_1 and R_2 so that r is minimized

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

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

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

❖ Equivalently:

assign $\underline{x}$ in ω_1 (if ω_2)

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

: likelihood ratio

ℓ_{12}

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

$$\underline{x} \rightarrow \omega_1 \text{ if } P(\underline{x}|\omega_1) > P(\underline{x}|\omega_2) \frac{\lambda_{21}}{\lambda_{12}}$$

$$\underline{x} \rightarrow \omega_2 \text{ if } P(\underline{x}|\omega_2) > P(\underline{x}|\omega_1) \frac{\lambda_{12}}{\lambda_{21}}$$

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

❖ An example:

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

➤ Then the threshold value is:

x_0 for minimum P_e :

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

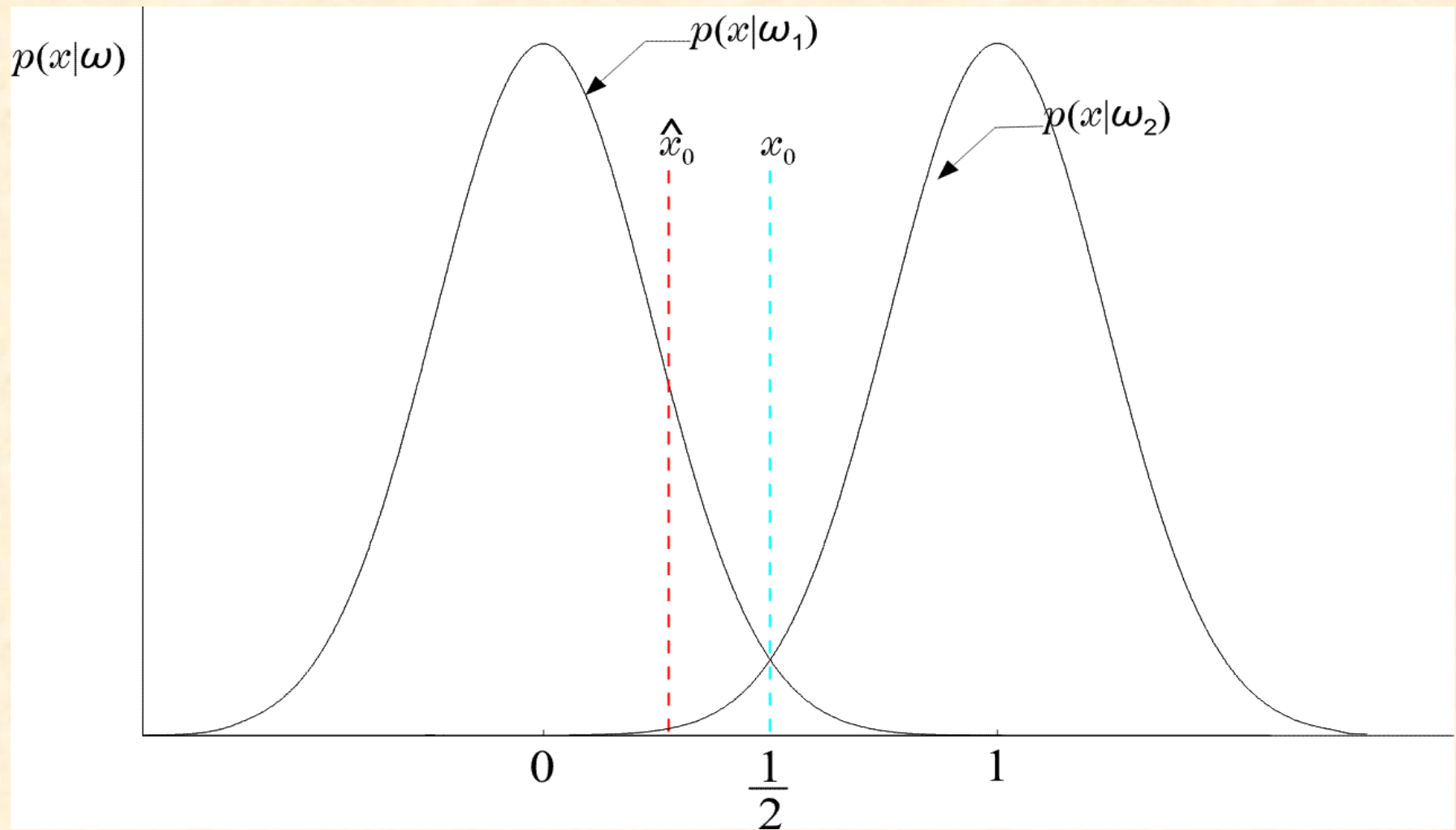
$$x_0 = \frac{1}{2}$$

➤ Threshold $\hat{x}_0$ for minimum r

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

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

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



DISCRIMINANT FUNCTIONS

DECISION SURFACES

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

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

$$\begin{array}{r} + \\ \hline - \end{array} g(\underline{x}) = 0$$

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

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

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

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

- ❖ $g_i(\underline{x}) \equiv f(P(\omega_i|\underline{x}))$ is a **discriminant function**
- ❖ In general, discriminant functions can be defined **independent** of the Bayesian rule. They lead to **suboptimal** solutions, yet if chosen appropriately, can be computationally more tractable.

BAYESIAN CLASSIFIER FOR NORMAL DISTRIBUTIONS

❖ Multivariate Gaussian pdf

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

$$\underline{\mu}_i = E[\underline{x}] \text{ } \ell \times \ell \text{ matrix in } \omega_i$$

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

called **covariance matrix**

❖ $\ln(\cdot)$ monotonic. Define:

➤ $g_i(\underline{x}) = \ln(p(\underline{x}|\omega_i)P(\omega_i)) =$

➤ $\ln p(\underline{x}|\omega_i) + \ln P(\omega_i)$

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

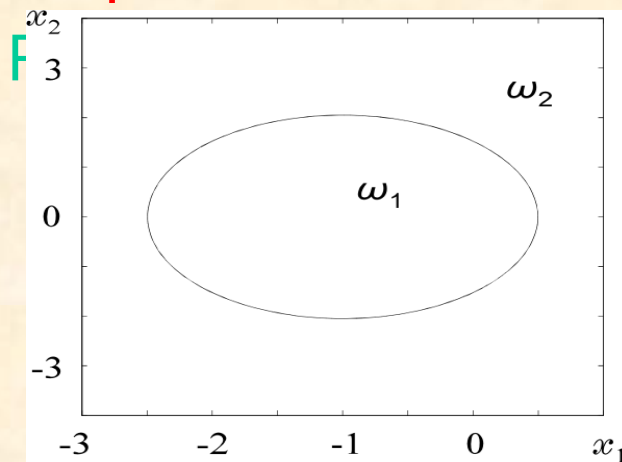
➤ Example: $C_i = -\left(\frac{\ell}{2}\right) \ln 2\pi - \left(\frac{1}{2}\right) \ln |\Sigma_i|$

$$\Sigma_i = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

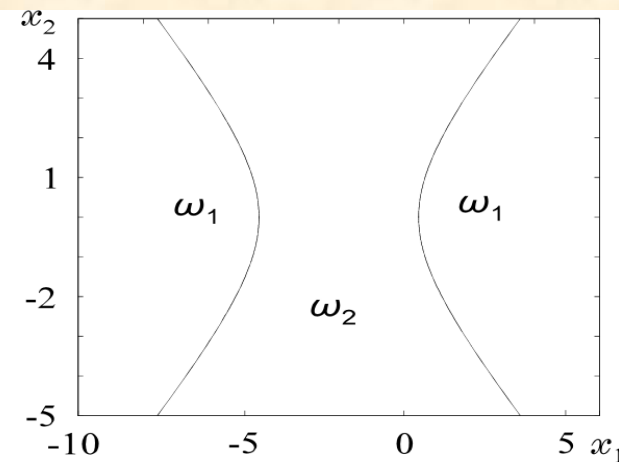
$$\begin{aligned} \triangleright \quad g_i(\underline{x}) = & -\frac{1}{2\sigma^2}(x_1^2 + x_2^2) + \frac{1}{\sigma^2}(\mu_{i1}x_1 + \mu_{i2}x_2) \\ & - \frac{1}{2\sigma^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln(P\omega_i) + C_i \end{aligned}$$

That $g_i(\underline{x})$ is **quadratic** and the surfaces $g_i(\underline{x}) - g_j(\underline{x}) = 0$

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



(a)



(b)

❖ 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:

$$g_i(\underline{x}) = \underline{w}_i^T \underline{x} + w_{i0}$$

$$\underline{w}_i = \Sigma^{-1} \underline{\mu}_i$$

$$w_{i0} = \ln P(\omega_i) - \frac{1}{2} \underline{\mu}_i^T \Sigma^{-1} \underline{\mu}_i$$

Discriminant functions are **LINEAR**

➤ Let in addition:

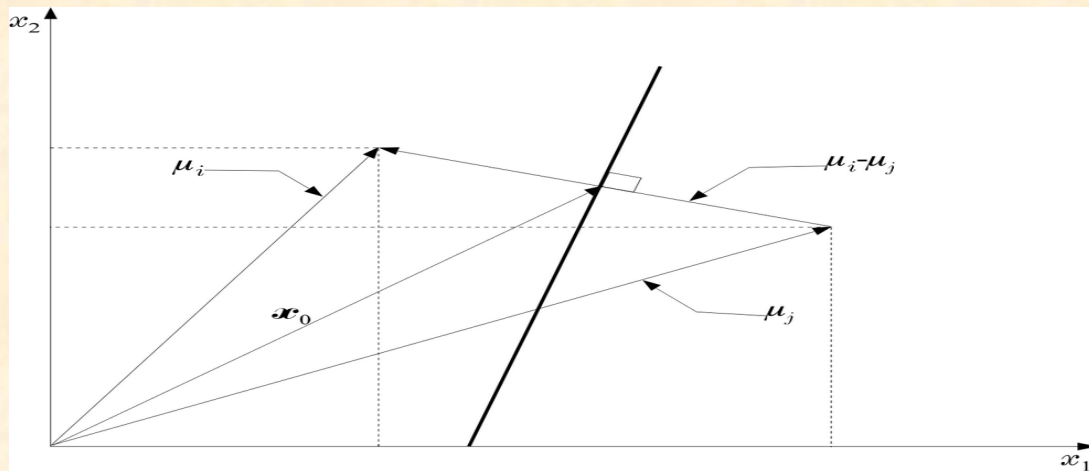
- $\Sigma = \sigma^2 I$. Then

$$g_i(\underline{x}) = \frac{1}{\sigma^2} \underline{\mu}_i^T \underline{x} + w_{i0}$$

- $g_{ij}(\underline{x}) = g_i(\underline{x}) - g_j(\underline{x}) = 0$
 $= \underline{w}^T (\underline{x} - \underline{x}_o)$

- $\underline{w} = \underline{\mu}_i - \underline{\mu}_j,$

$$\underline{x}_o = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \sigma^2 \ln \frac{P(\omega_i)}{P(\omega_j)} \frac{\underline{\mu}_i - \underline{\mu}_j}{\|\underline{\mu}_i - \underline{\mu}_j\|^2}$$



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

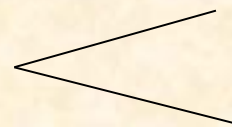
- $g_{ij}(\underline{x}) = \underline{w}^T (\underline{x} - \underline{x}_0) = 0$

- $\underline{w} = \Sigma^{-1}(\underline{\mu}_i - \underline{\mu}_j)$

- $\underline{x}_0 = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\underline{\mu}_i - \underline{\mu}_j}{\left\| \underline{\mu}_i - \underline{\mu}_j \right\|_{\Sigma^{-1}}^2}$

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

➤ Decision hyperplane



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

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

❖ Minimum Distance Classifiers

- equiprobable

- $P(\omega_i) = \frac{1}{M}$

- $g_i(\underline{x}) = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma^{-1}(\underline{x} - \underline{\mu}_i)$

Euclidean Distance:

smaller
 $\Sigma = \sigma^2 I$: Assign $\underline{x} \rightarrow \omega_i$:

-

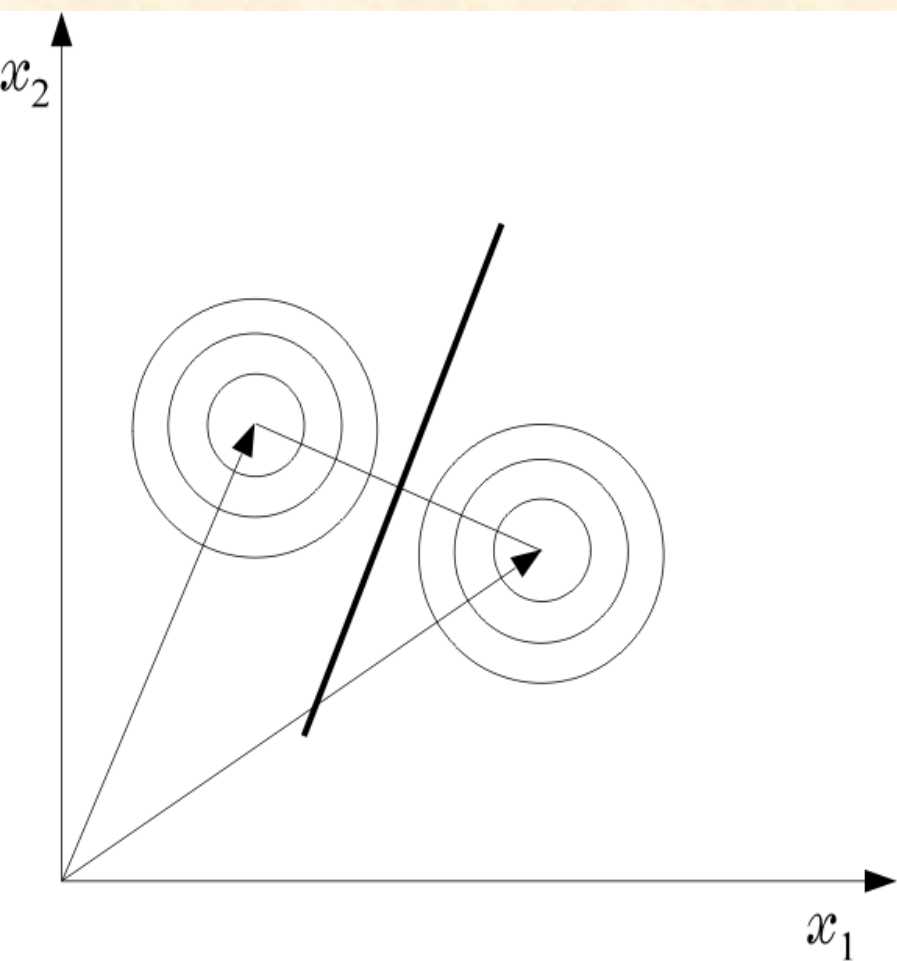
$$d_E \equiv \left\| \underline{x} - \underline{\mu}_i \right\|$$

Mahalanobis Distance:

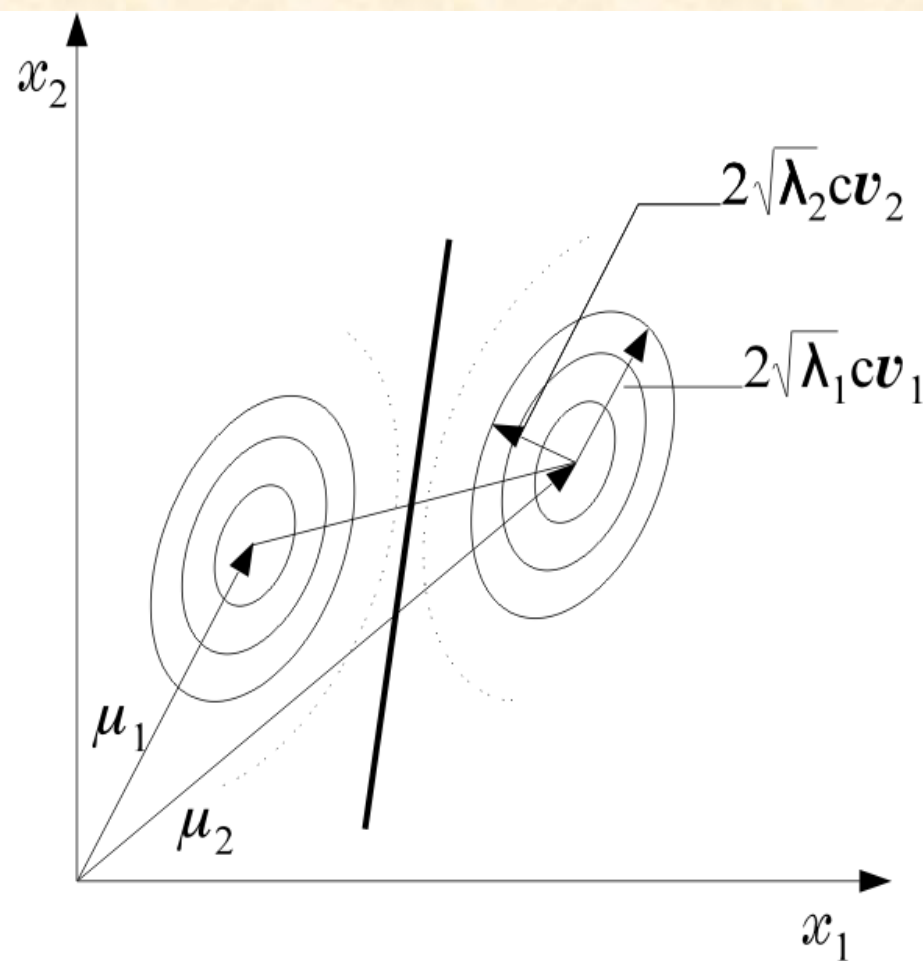
smaller

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

$$d_m = \left((\underline{x} - \underline{\mu}_i)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_i) \right)^{\frac{1}{2}}$$



(a)



(b)

❖ Example:

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

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

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 : $d_{m,1}^2 = [1.0, 2.2]$
 $\Sigma^{-1} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952, d_{m,2}^2 = [-2.0, -0.8] \Sigma^{-1} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$
- Classify $\underline{x} \rightarrow \omega_1$. Observe that $d_{E,2} < d_{E,1}$

ESTIMATION OF UNKNOWN PROBABILITY DENSITY FUNCTIONS

❖ Maximum Likelihood

- Let $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N$ known and independent
- Let $p(\underline{x})$ known within an unknown vector parameter $\underline{\theta} : p(\underline{x}) \equiv p(\underline{x}; \underline{\theta})$
- $X = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\}$
- $p(X; \underline{\theta}) \equiv p(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N; \underline{\theta})$
$$= \prod_{k=1}^N p(\underline{x}_k; \underline{\theta})$$

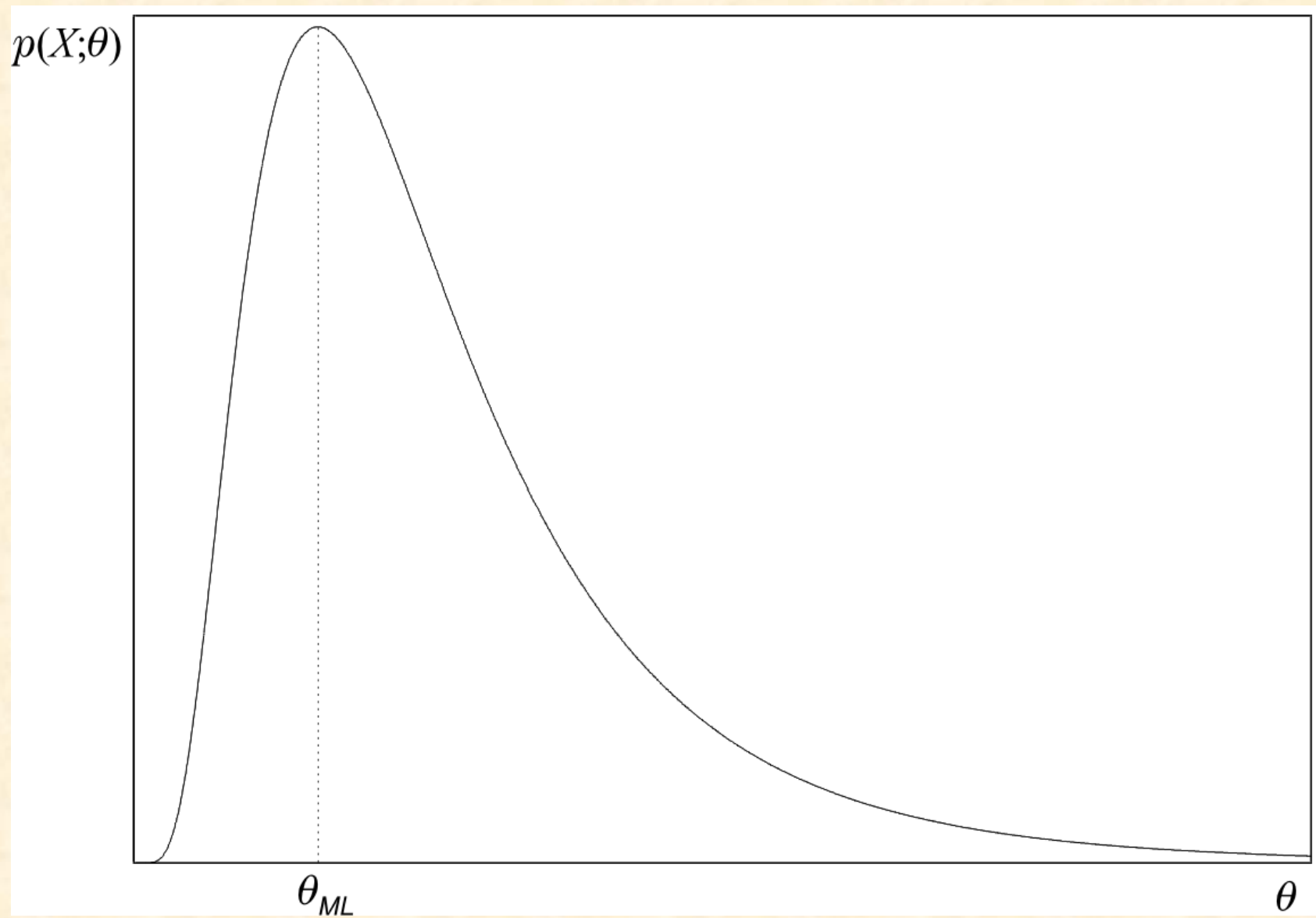
which is known as the Likelihood of $\underline{\theta}$ w.r. to X

The method :

$$\triangleright \hat{\underline{\theta}}_{\text{ML}} : \arg \max_{\underline{\theta}} \prod_{k=1}^N p(\underline{x}_k; \underline{\theta})$$

$$\triangleright L(\underline{\theta}) \equiv \ln p(X; \underline{\theta}) = \sum_{k=1}^N \ln p(\underline{x}_k; \underline{\theta})$$

$$\triangleright \hat{\underline{\theta}}_{ML} : \frac{\partial L(\underline{\theta})}{\partial(\underline{\theta})} = \sum_{k=1}^N \frac{1}{p(\underline{x}_k; \underline{\theta})} \frac{\partial p(\underline{x}_k; \underline{\theta})}{\partial(\underline{\theta})} = \underline{0}$$



If, indeed, there is a $\underline{\theta}_0$ such that

$p(\underline{x}) = p(\underline{x}; \underline{\theta}_0)$, then

$$\lim_{N \rightarrow \infty} E[\underline{\theta}_{ML}] = \underline{\theta}_0$$

$$\lim_{N \rightarrow \infty} E\left\|\hat{\underline{\theta}}_{ML} - \underline{\theta}_0\right\|^2 = 0$$

Asymptotically unbiased and consistent

❖ Example:

$$p(\underline{x}) : N(\underline{\mu}, \underline{\Sigma}) : \underline{\mu} \text{ unknown, } \underline{x}_1, \underline{x}_2, \dots, \underline{x}_N \quad p(\underline{x}_k) \equiv p(\underline{x}_k; \underline{\mu})$$

$$L(\underline{\mu}) = \ln \prod_{k=1}^N p(\underline{x}_k; \underline{\mu}) = C - \frac{1}{2} \sum_{k=1}^N (\underline{x}_k - \underline{\mu})^T \underline{\Sigma}^{-1} (\underline{x}_k - \underline{\mu})$$

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

$$\frac{\partial L(\underline{\mu})}{\partial(\underline{\mu})} \equiv \begin{bmatrix} \frac{\partial L}{\partial \mu_1} \\ \vdots \\ \frac{\partial L}{\partial \mu_l} \end{bmatrix} = \sum_{k=1}^N \underline{\Sigma}^{-1} (\underline{x}_k - \underline{\mu}) = \underline{0} \Rightarrow \underline{\mu}_{ML} = \frac{1}{N} \sum_{k=1}^N \underline{x}_k$$

$$\text{Remember: if } A = A^T \Rightarrow \frac{\partial(\underline{\alpha}^T A \underline{\alpha})}{\partial \underline{\alpha}} = 2A\underline{\alpha}$$

❖ Maximum A Posteriori Probability Estimation

- In ML method, $\underline{\theta}$ was considered as a **parameter**
- Here we shall look at $\underline{\theta}$ as a **random vector** described by a pdf $p(\underline{\theta})$, **assumed to be known**
- Given $X = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\}$

Compute the maximum of
 $p(\underline{\theta}|X)$

- From Bayes theorem $p(\underline{\theta})p(X|\underline{\theta}) = p(X)p(\underline{\theta}|X)$ or

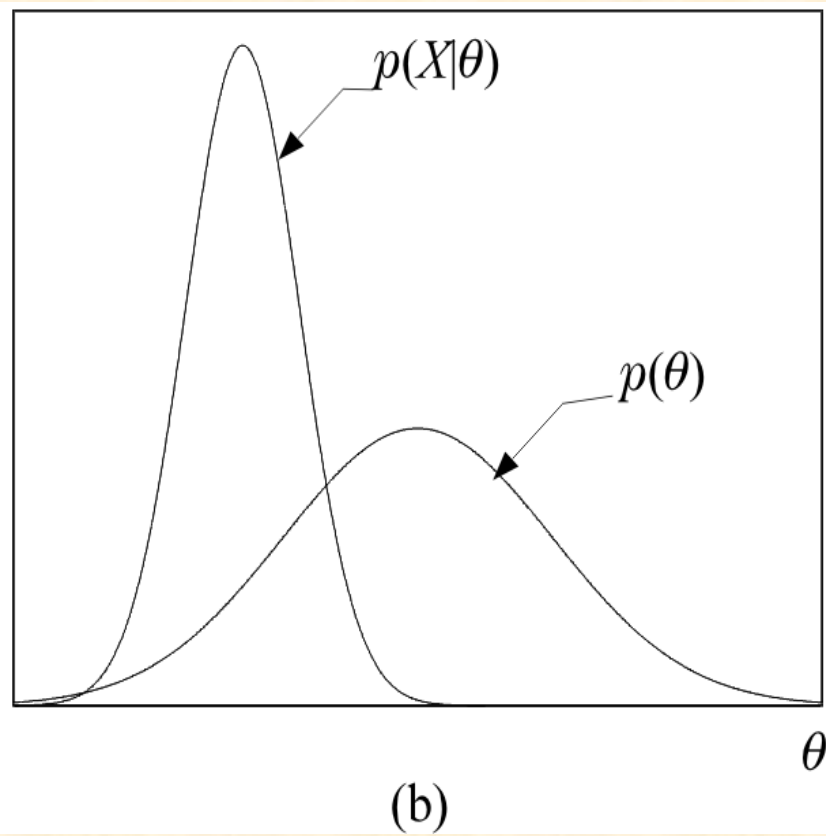
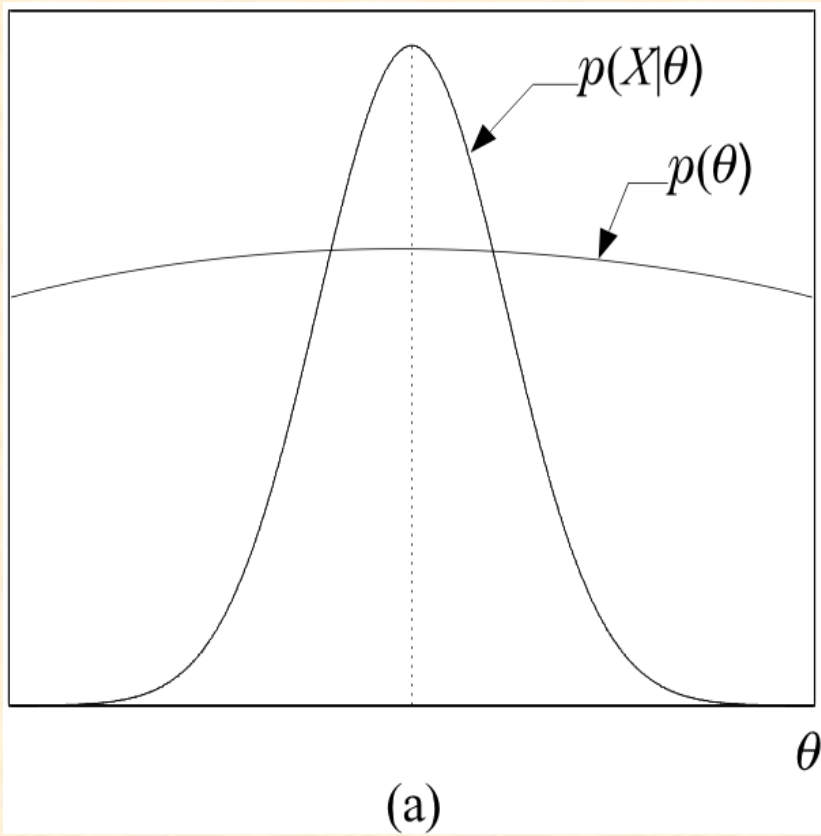
$$p(\underline{\theta}|X) = \frac{p(\underline{\theta})p(X|\underline{\theta})}{p(X)}$$

➤ The method:

$$\hat{\underline{\theta}}_{MAP} = \arg \max_{\underline{\theta}} p(\underline{\theta}|X) \text{ or}$$

$$\hat{\underline{\theta}}_{MAP} : \frac{\partial}{\partial \underline{\theta}} (P(\underline{\theta}) p(X|\underline{\theta}))$$

If $p(\underline{\theta})$ is uniform or broad enough $\hat{\underline{\theta}}_{MAP} \cong \underline{\theta}_{ML}$



❖ Example:

$$p(\underline{x}) : N(\underline{\mu}, \Sigma), \quad \underline{\mu} \text{ unknown}, \quad X = \{\underline{x}_1, \dots, \underline{x}_N\}$$

$$p(\underline{\mu}) = \frac{1}{(2\pi)^{\frac{l}{2}} \sigma_{\mu}^l} \exp\left(-\frac{\|\underline{\mu} - \underline{\mu}_0\|^2}{2\sigma_{\mu}^2}\right)$$

$$\underline{\hat{\theta}}_{MAP} : \frac{\partial}{\partial \underline{\mu}} \ln\left(\prod_{k=1}^N p(\underline{x}_k | \underline{\mu}) p(\underline{\mu})\right) = \underline{0} \quad \text{or} \quad \sum_{k=1}^N \frac{1}{\sigma^2} (\underline{x}_k - \underline{\mu}) - \frac{1}{\sigma_{\mu}^2} (\hat{\underline{\mu}} - \underline{\mu}_0) = \underline{0} \Rightarrow$$

$$\hat{\underline{\mu}}_{MAP} = \frac{\underline{\mu}_0 + \frac{\sigma_{\mu}^2}{\sigma^2} \sum_{k=1}^N \underline{x}_k}{1 + \frac{\sigma_{\mu}^2}{\sigma^2} N} \quad \text{For } \frac{\sigma_{\mu}^2}{\sigma^2} \gg 1, \text{ or for } N \rightarrow \infty$$

$$\hat{\underline{\mu}}_{MAP} \cong \hat{\underline{\mu}}_{ML} = \frac{1}{N} \sum_{k=1}^N \underline{x}_k$$

❖ Bayesian Inference

- ML, MAP $\Rightarrow$ a single estimate for $\underline{\theta}$.

Here a different root is followed.

Given: $X = \{\underline{x}_1, \dots, \underline{x}_N\}$, $p(\underline{x}|\underline{\theta})$ and $p(\underline{\theta})$

The goal: estimate $p(\underline{x}|X)$

How??

$$p(\underline{x}|X) = \int p(\underline{x}|\underline{\theta})p(\underline{\theta}|X)d\underline{\theta}$$

$$p(\underline{\theta}|X) = \frac{p(X|\underline{\theta})p(\underline{\theta})}{p(X)} = \frac{p(X|\underline{\theta})p(\underline{\theta})}{\int p(X|\underline{\theta})p(\underline{\theta})d\underline{\theta}}$$

$$p(X|\underline{\theta}) = \prod_{k=1}^N p(x_k|\underline{\theta})$$

A bit more insight via an example

- Let $p(x|\mu) \rightarrow N(\mu, \sigma^2)$
- $p(\mu) \rightarrow N(\mu_0, \sigma_0^2)$
- It turns out that: $p(\mu|X) \rightarrow N(\mu_N, \sigma_N^2)$

$$\mu_N = \frac{N\sigma_0^2 \bar{x} + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2},$$

$$\sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2},$$

$$\bar{x} = \frac{1}{N} \sum_{k=1}^N x_k$$

➤ The above is a sequence of Gaussians as $N \rightarrow \infty$



❖ Maximum Entropy

➤ Entropy $H = - \int p(\underline{x}) \ln p(\underline{x}) d\underline{x}$



$\hat{p}(x)$: maximum H subject to the
available constraints

➤ **Example:** x is nonzero in the interval $x_1 \leq x \leq x_2$ and zero otherwise. Compute the ME pdf

- The constraint:

$$\int_{x_1}^{x_2} p(x) dx = 1$$

- Lagrange Multipliers

$$H_L = H + \lambda \left(\int_{x_1}^{x_2} p(x) dx - 1 \right)$$

•

$$\hat{p}(x) = \exp(\lambda - 1)$$

$$\hat{p}(x) = \begin{cases} \frac{1}{x_2 - x_1} & x_1 \leq x \leq x_2 \\ 0 & \text{otherwise} \end{cases}$$

❖ Mixture Models

$$\triangleright p(\underline{x}) = \sum_{j=1}^J p(\underline{x}|j)P_j$$

$$\sum_{j=1}^M P_j = 1, \int_{\underline{x}} p(\underline{x}|j) d\underline{x} = 1$$

$$\triangleright \text{Assume parametric modeling, i.e., } p(\underline{x}|j; \underline{\theta})$$

$$\triangleright \text{The goal is to estimate } \underline{\theta} \text{ and } P_1, P_2, \dots, P_j \\ \text{given a set } X = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\}$$

$\triangleright$ Why not ML? As before?

$$\max_{\underline{\theta}, P_1, \dots, P_j} \prod_{k=1}^N P(\underline{x}_k; \underline{\theta}, P_1, \dots, P_j)$$

- This is a **nonlinear problem** due to the missing label information. This is a typical problem with **an incomplete data set**.
- The Expectation-Maximisation (EM) algorithm.
 - General formulation
 - $\underline{y}$ the complete data set $\underline{y} \in Y \subseteq R^m$, with $p_{\underline{y}}(\underline{y}; \underline{\theta})$, which are **not observed directly**.

We observe

$$\underline{x} = g(\underline{y}) \in X_{ob} \subseteq R^l, l < m \text{ with } P_x(\underline{x}; \underline{\theta}),$$

a many to one transformation

- Let $Y(\underline{x}) \subseteq Y$ all $\underline{y}$'s $\rightarrow$ to a specific $\underline{x}$

$$p_{\underline{x}}(\underline{x}; \underline{\theta}) = \int_{Y(\underline{x})} p_{\underline{y}}(\underline{y}; \underline{\theta}) d\underline{y}$$

- What we need is to compute

$$\hat{\theta}_{ML} : \sum_k \frac{\partial \ln(p_{\underline{y}}(\underline{y}_k; \underline{\theta}))}{\partial \underline{\theta}} = \underline{0}$$

- But $\underline{y}_k$'s are not observed. Here comes the EM.
Maximize the **expectation** of the loglikelihood
conditioned on the observed samples and the current
iteration estimate of

$$\underline{\theta}.$$

➤ The algorithm:

- E-step: $Q(\underline{\theta}; \underline{\theta}(t)) = E\left[\sum_k \ln(p_{\underline{y}}(\underline{y}_k; \underline{\theta} | X; \underline{\theta}(t)))\right]$

- M-step: $\underline{\theta}(t+1) : \frac{\partial Q(\underline{\theta}; \underline{\theta}(t))}{\partial \underline{\theta}} = \underline{0}$

➤ Application to the mixture modeling problem

- Complete data $(\underline{x}_k, j_k), k = 1, 2, \dots, N$

- Observed data $\underline{x}_k, k = 1, 2, \dots, N$

- $p(\underline{x}_k, j_k; \underline{\theta}) = p(\underline{x}_k | j_k; \underline{\theta}) P_{jk}$

- Assuming mutual independence

$$L(\underline{\theta}) = \sum_{k=1}^N \ln(p(\underline{x}_k | j_k; \underline{\theta}) P_{jk})$$

- Unknown parameters

$$\underline{\Theta}^T = [\underline{\theta}^T, \underline{P}^T]^T, \quad \underline{P} = [P_1, P_2, \dots, P_J]^T$$

- E-step

$$Q(\underline{\Theta}; \underline{\Theta}(t)) = E \left[\sum_{k=1}^N \ln(p(\underline{x}_k | j_k; \underline{\theta}) P_{j_k}) \right] = \sum_{k=1}^N E[\quad] =$$

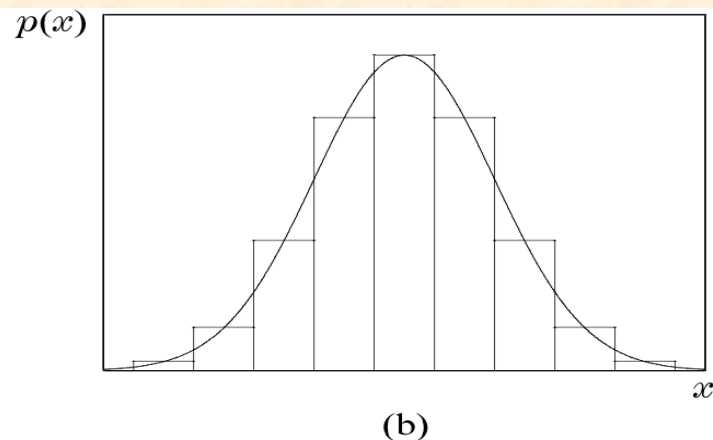
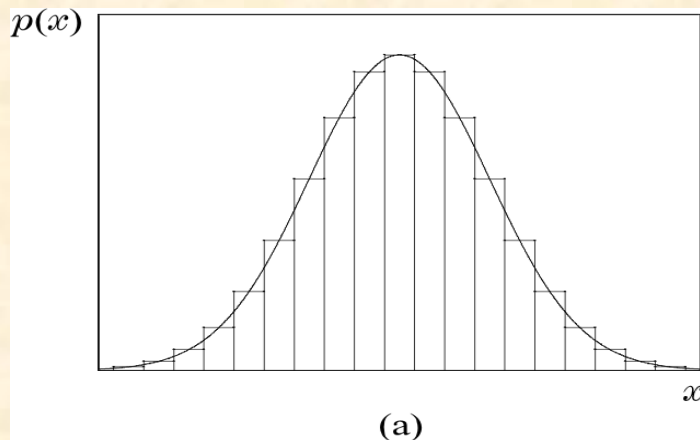
$$\sum_{k=1}^N \sum_{j_k=1}^J \boxed{P(j_k | \underline{x}_k; \underline{\Theta}(t))} \ln(p(\underline{x}_k | j_k; \underline{\theta}) P_{j_k})$$

- M-step $\frac{\partial Q}{\partial \underline{\theta}} = \underline{0} \quad \frac{\partial Q}{\partial P_{j_k}} = 0, \quad j_k = 1, 2, \dots, J$

$$P(j | \underline{x}_k; \underline{\Theta}(t)) = \frac{p(\underline{x}_k | j; \underline{\Theta}(t)) P_j}{P(\underline{x}_k; \underline{\Theta}(t))}$$

$$p(\underline{x}_k; \underline{\Theta}(t)) = \sum_{j=1}^J p(\underline{x}_k | j; \underline{\Theta}(t)) P_j$$

❖ Nonparametric Estimation



$$P \approx \frac{k_N}{N} \begin{matrix} \nearrow k_N \text{ in } h \\ \searrow N \text{ total} \end{matrix}$$



$$\hat{p}(x) \equiv \hat{p}(\hat{x}) = \frac{1}{h} \frac{k_N}{N}, |x - \hat{x}| \leq \frac{h}{2}$$

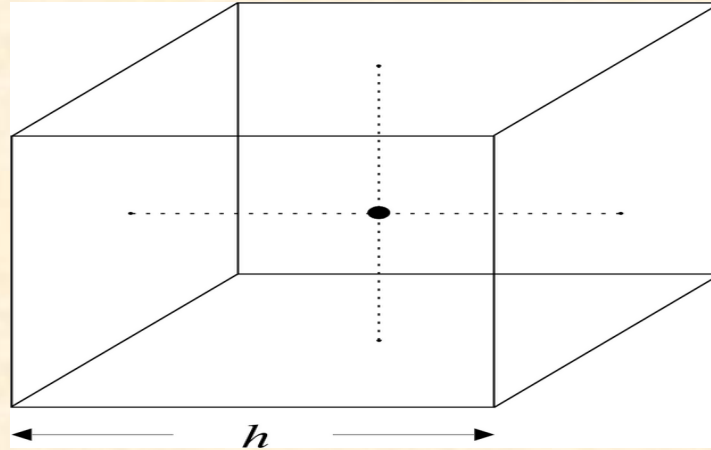
$$\hat{x} - \frac{h}{2} \quad \hat{x} \quad \hat{x} + \frac{h}{2}$$

If $p(x)$ continuous , $\hat{p}(x) \rightarrow p(x)$ as $N \rightarrow \infty$, if

$$h_N \rightarrow 0, \quad k_N \rightarrow \infty, \quad \frac{k_N}{N} \rightarrow 0$$

❖ Parzen Windows

- Divide the multidimensional space in hypercubes



➤ Define

$$\varphi(\underline{x}_i) = \begin{cases} 1 & |x_{ij}| \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

- That is, it is 1 inside a unit side hypercube centered at 0

•

$$\hat{p}(\underline{x}) = \frac{1}{h^l} \left(\frac{1}{N} \sum_{i=1}^N \varphi\left(\frac{\underline{x}_i - \underline{x}}{h}\right) \right)$$

- $\frac{1}{\text{volume}} * \frac{1}{N} * \text{number of points inside an } h\text{-side hypercube centered at } \underline{x}$

- Parzen windows-kernels-potential functions
 $p(\underline{x})$ continuous
 $\varphi(\cdot)$ discontinuous

$\varphi(\underline{x})$ is smooth

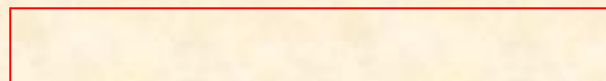
$$\varphi(\underline{x}) \geq 0, \quad \int_{\underline{x}} \varphi(\underline{x}) d\underline{x} = 1$$

➤ Mean value

$$E[\hat{p}(\underline{x})] = \frac{1}{h^l} \left(\frac{1}{N} \sum_{i=1}^N E[\varphi(\frac{\underline{x}_i - \underline{x}}{h})] \right) = \int_{\underline{x}'} \frac{1}{h^l} \varphi(\frac{\underline{x}' - \underline{x}}{h}) p(\underline{x}') d\underline{x}'$$

- $h \rightarrow 0, \frac{1}{h^l} \rightarrow \infty$
- $h \rightarrow 0$ the width of $\varphi(\frac{\underline{x}' - \underline{x}}{h}) \rightarrow 0$
- $\int \frac{1}{h^l} \varphi(\frac{\underline{x}' - \underline{x}}{h}) d\underline{x} = 1$
- $h \rightarrow 0 \quad \frac{1}{h^l} \varphi(\frac{\underline{x}}{h}) \rightarrow \delta(\underline{x})$

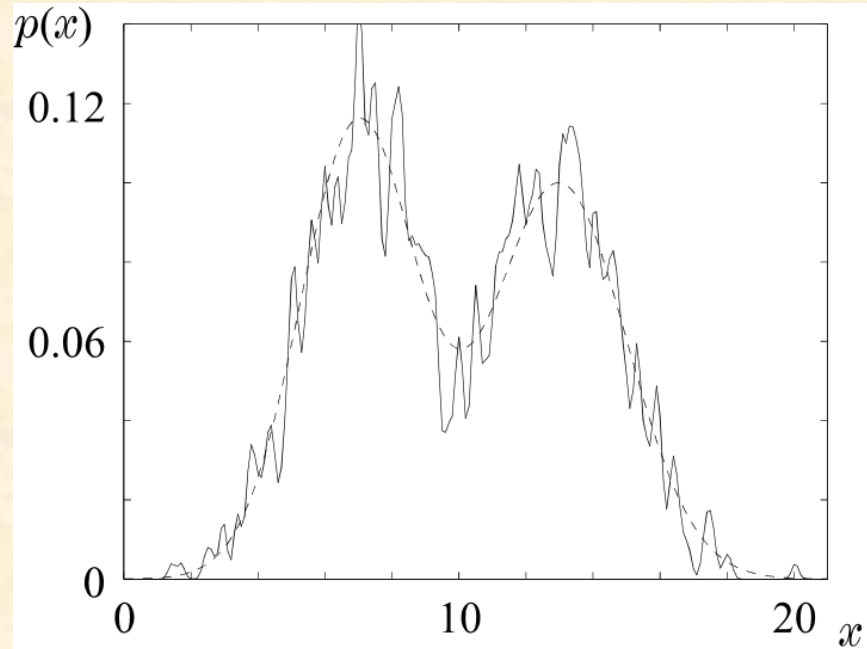
Hence $E[\hat{p}(\underline{x})] = \int_{\underline{x}'} \delta(\underline{x}' - \underline{x}) p(\underline{x}') d\underline{x}' = p(\underline{x})$
 unbiased in the limit



➤ Variance

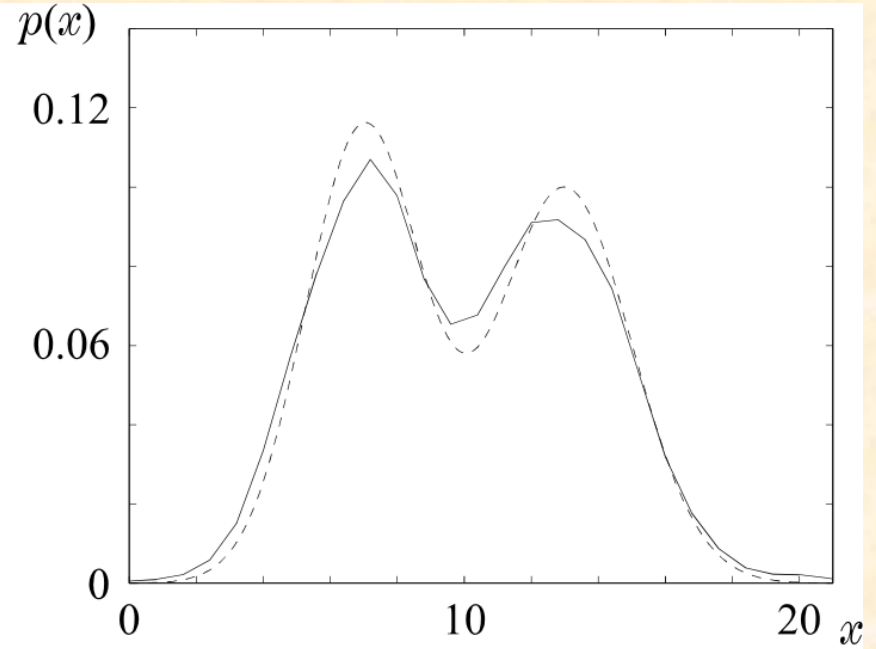
- The **smaller** the h the **higher** the variance

$h=0.1, N=1000$



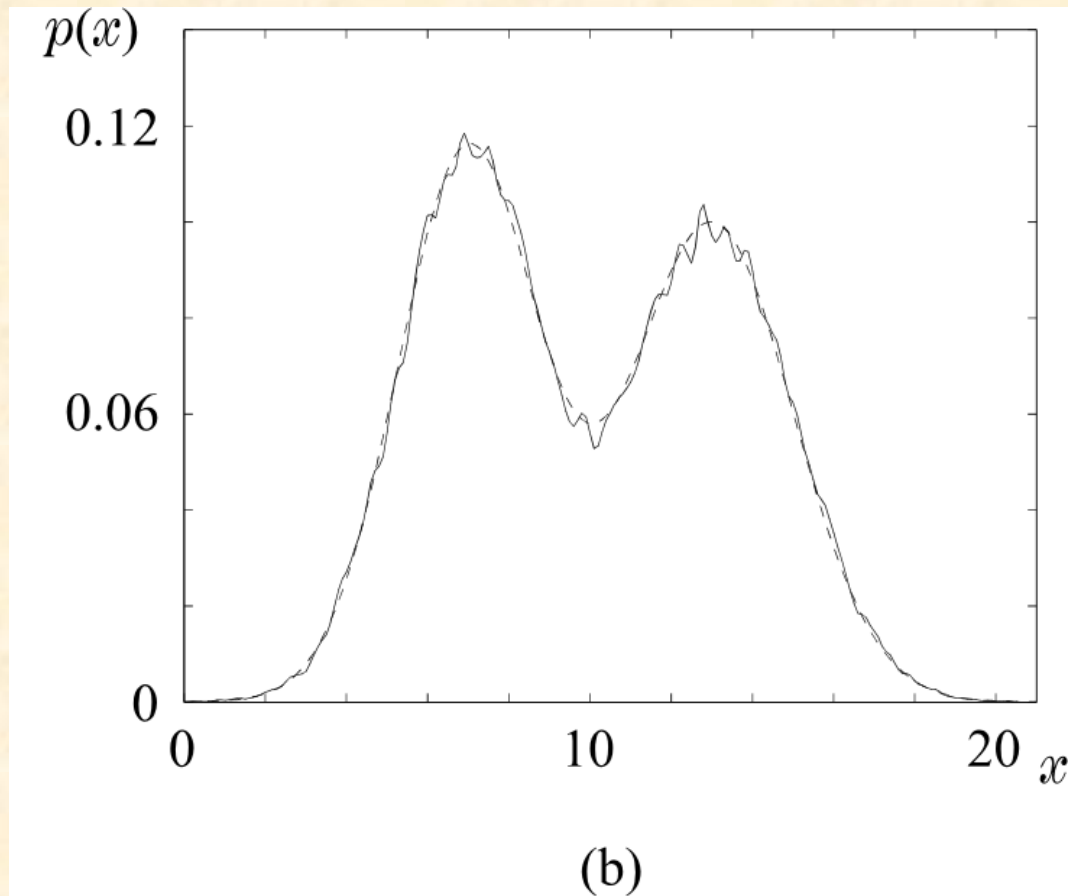
(a)

$h=0.8, N=1000$



(b)

$$h=0.1, N=10000$$



➤ The **higher** the N the **better** the accuracy

➤ If

- $h \rightarrow 0$
- $N \rightarrow \infty$
- $h_N \rightarrow \infty$

asymptotically unbiased

➤ The method

- Remember:

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

$$\cdot \frac{\frac{1}{N_1 h^l} \sum_{i=1}^{N_1} \varphi\left(\frac{\underline{x}_i - \underline{x}}{h}\right)}{\frac{1}{N_2 h^l} \sum_{i=1}^{N_2} \varphi\left(\frac{\underline{x}_i - \underline{x}}{h}\right)} (><) \theta$$

❖ CURSE OF DIMENSIONALITY

- In all the methods, so far, we saw that the **highest** the number of points, N , the **better** the resulting estimate.
- If in the one-dimensional space an interval, filled with N points, is **adequately** (for good estimation), in the two-dimensional space the corresponding square will require N^2 and in the ℓ -dimensional space the ℓ -dimensional cube will require N^ℓ points.
- The exponential increase in the number of necessary points is known as **the curse of dimensionality**. This is a major problem one is confronted with in high dimensional spaces.

❖ NAIVE – BAYES CLASSIFIER

➤ Let $\underline{x} \in \mathcal{R}^\ell$ and the goal is to estimate $p(\underline{x} | \omega_i)$ $i = 1, 2, \dots, M$. For a “good” estimate of the pdf one would need, say, N^ℓ points.

➤ Assume $x_1, x_2, \dots, x_\ell$ **mutually independent**. Then:

$$p(\underline{x} | \omega_i) = \prod_{j=1}^{\ell} p(x_j | \omega_i)$$

➤ In this case, one would require, roughly, N points for each pdf. Thus, a number of points of the order $N \cdot \ell$ would suffice.

➤ It turns out that the Naïve – Bayes classifier works reasonably well even in cases that violate the independence assumption.

❖ K Nearest Neighbor Density Estimation

➤ In Parzen:

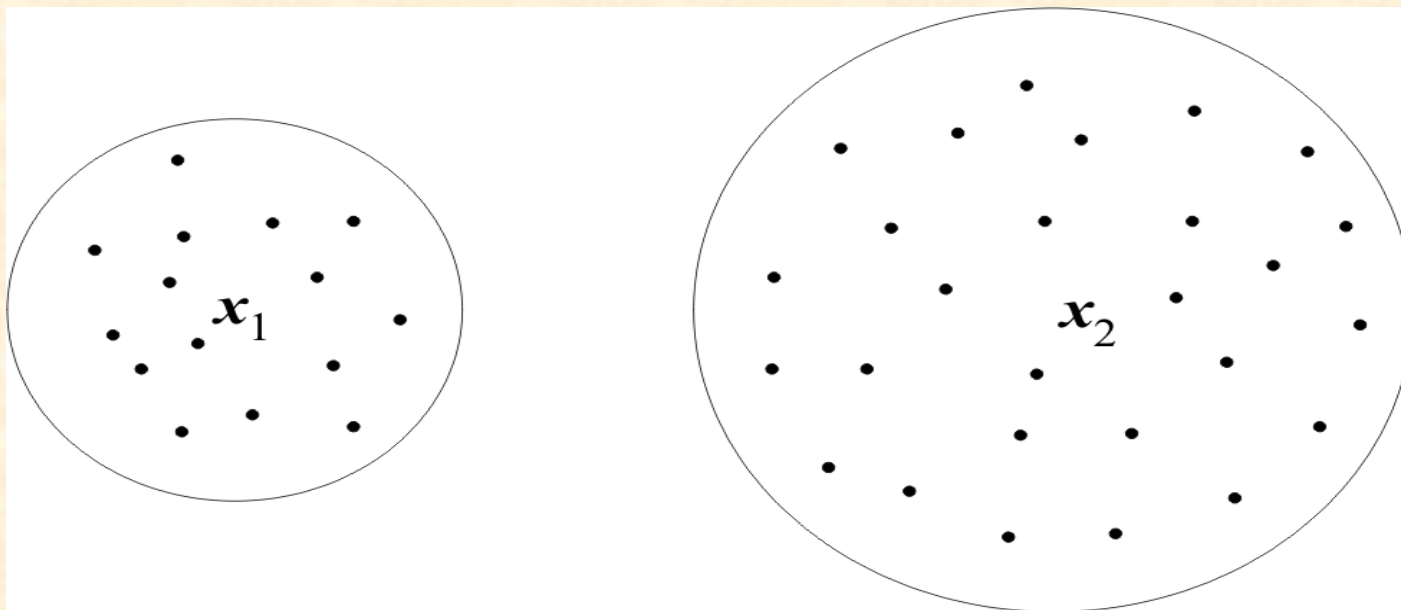
- The volume is constant
- The number of points in the volume is varying

➤ Now:

- Keep the number of points $k_N = k$
constant

- Leave the volume to be varying

- $$\hat{p}(\underline{x}) = \frac{k}{NV(\underline{x})}$$



$$\frac{\frac{k}{N_1 V_1}}{\frac{k}{N_2 V_2}} = \frac{N_2 V_2}{N_1 V_1} \langle \theta \rangle$$

❖ The Nearest Neighbor Rule

- Choose k out of the N training vectors, identify the k nearest ones to $\underline{x}$
- Out of these k identify k_i that belong to class ω_i
- Assign $\underline{x} \rightarrow \omega_i : k_i > k_j \quad \forall i \neq j$
- The simplest version

$k=1$!!!
- For large N this is not bad. It can be shown that:
if P_B is the optimal Bayesian error probability, then:

$$P_B \leq P_{NN} \leq P_B \left(2 - \frac{M}{M-1} P_B \right) \leq 2P_B$$

$$\triangleright P_B \leq P_{kNN} \leq P_B + \sqrt{\frac{2P_{NN}}{k}}$$

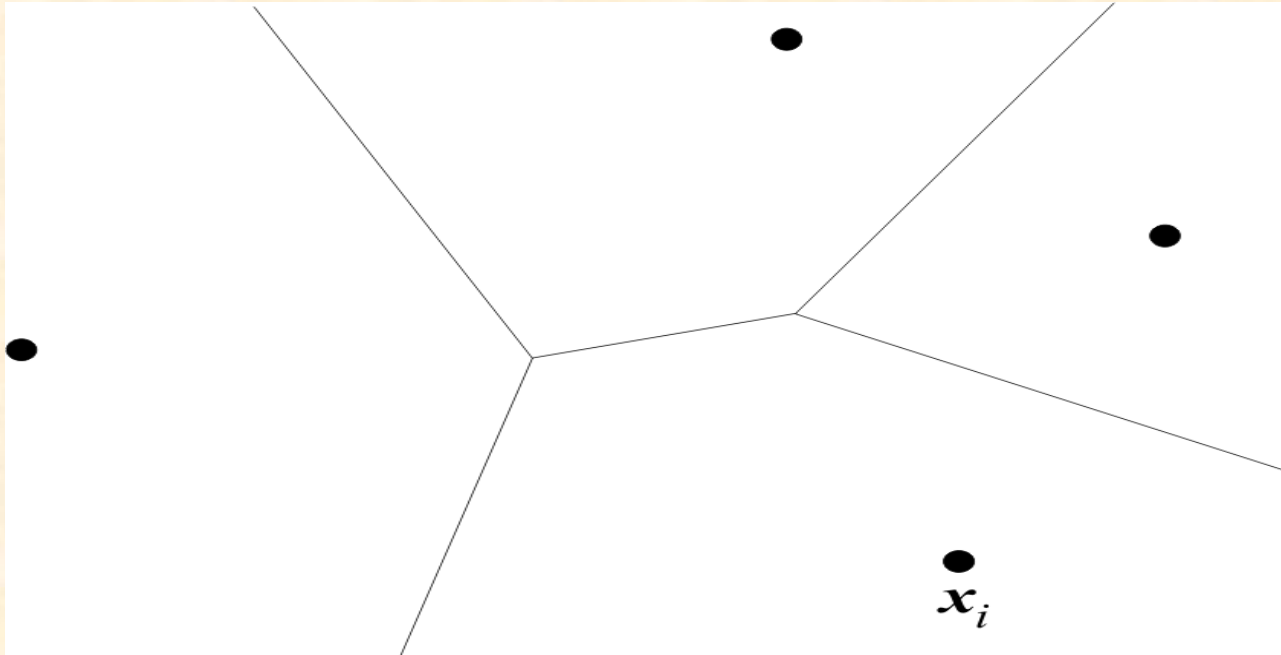
$$\triangleright k \rightarrow \infty, P_{kNN} \rightarrow P_B$$

$\triangleright$ For small P_B :

$$P_{NN} \cong 2P_B$$

$$P_{3NN} \cong P_B + 3(P_B)^2$$

❖ Voronoi tessellation



$$R_i = \{ \underline{x} : d(\underline{x}, \underline{x}_i) < d(\underline{x}, \underline{x}_j) \text{ } i \neq j \}$$

BAYESIAN NETWORKS

❖ Bayes Probability Chain Rule

$$p(x_1, x_2, \dots, x_\ell) = p(x_\ell \mid x_{\ell-1}, \dots, x_1) \cdot p(x_{\ell-1} \mid x_{\ell-2}, \dots, x_1) \cdot \dots \\ \dots \cdot p(x_2 \mid x_1) \cdot p(x_1)$$

- Assume now that the **conditional** dependence for each x_i is limited to a subset of the features appearing in each of the product terms. That is:

$$p(x_1, x_2, \dots, x_\ell) = p(x_1) \cdot \prod_{i=2}^{\ell} p(x_i \mid A_i)$$

where

$$A_i \subseteq \{x_{i-1}, x_{i-2}, \dots, x_1\}$$

- For example, if $\ell=6$, then we could assume:

$$p(x_6 \mid x_5, \dots, x_1) = p(x_6 \mid x_5, x_4)$$

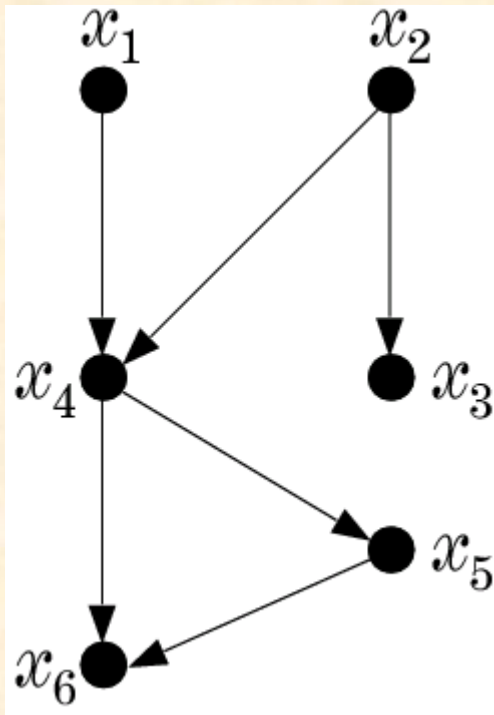
Then:

$$A_6 = \{x_5, x_4\} \subseteq \{x_5, \dots, x_1\}$$

- The above is a generalization of the Naïve – Bayes. For the Naïve – Bayes the assumption is:

$$A_i = \emptyset, \text{ for } i=1, 2, \dots, \ell$$

- A graphical way to portray **conditional dependencies** is given below



- According to this figure we have that:

- x_6 is conditionally dependent on x_4, x_5 .
- x_5 on x_4
- x_4 on x_1, x_2
- x_3 on x_2
- x_1, x_2 are conditionally **independent** on other variables.

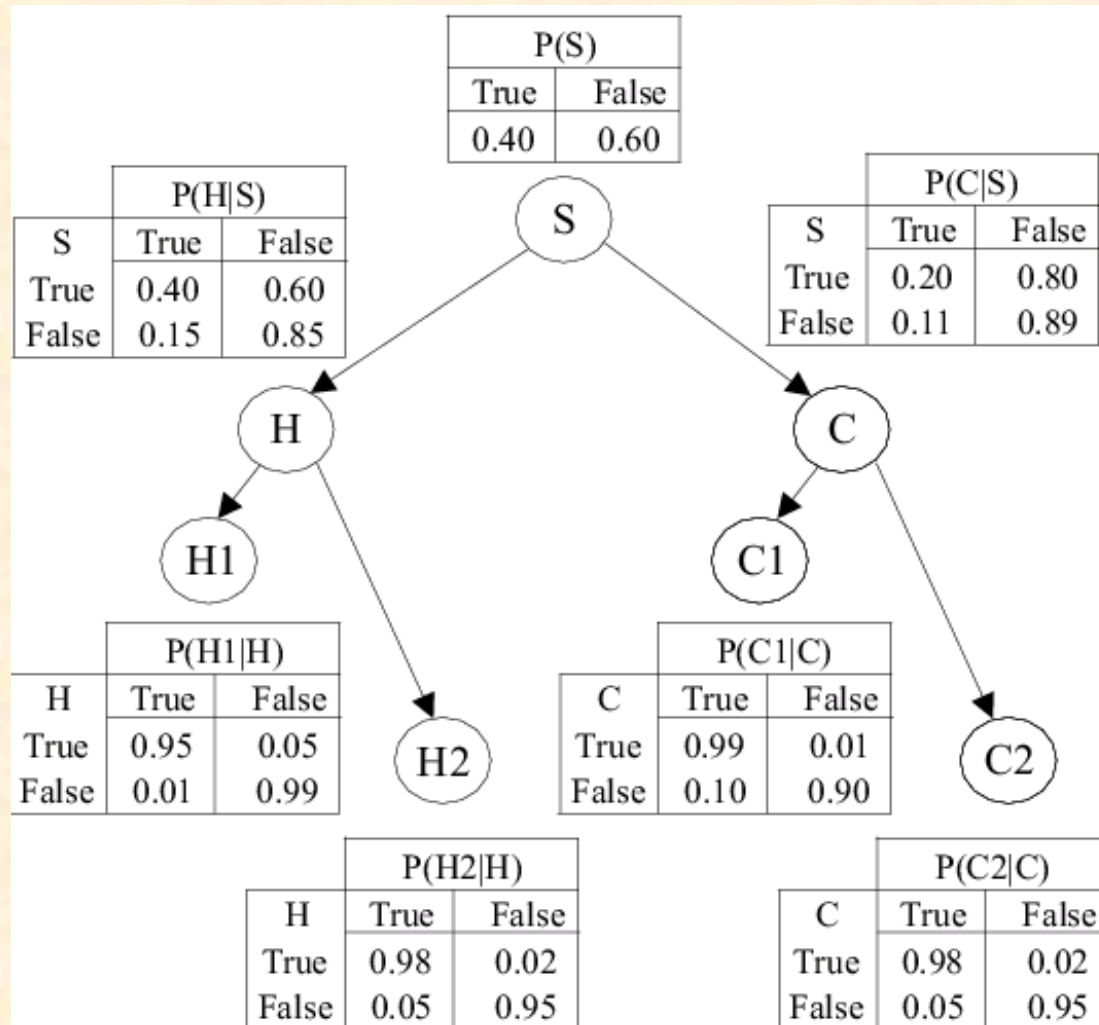
- For this case:

$$p(x_1, x_2, \dots, x_6) = p(x_6 | x_5, x_4) \cdot p(x_5 | x_4) \cdot p(x_3 | x_2) \cdot p(x_2) \cdot p(x_1)$$

❖ Bayesian Networks

- **Definition:** A Bayesian Network is a **directed acyclic graph** (DAG) where the nodes correspond to random variables. Each node is associated with a set of **conditional probabilities (densities)**, $p(x_i | A_i)$, where x_i is the variable associated with the node and A_i is the set of its **parents** in the graph.
- A Bayesian Network is specified by:
 - The marginal probabilities of its root nodes.
 - The conditional probabilities of the non-root nodes, **given their parents**, for **ALL** possible combinations.

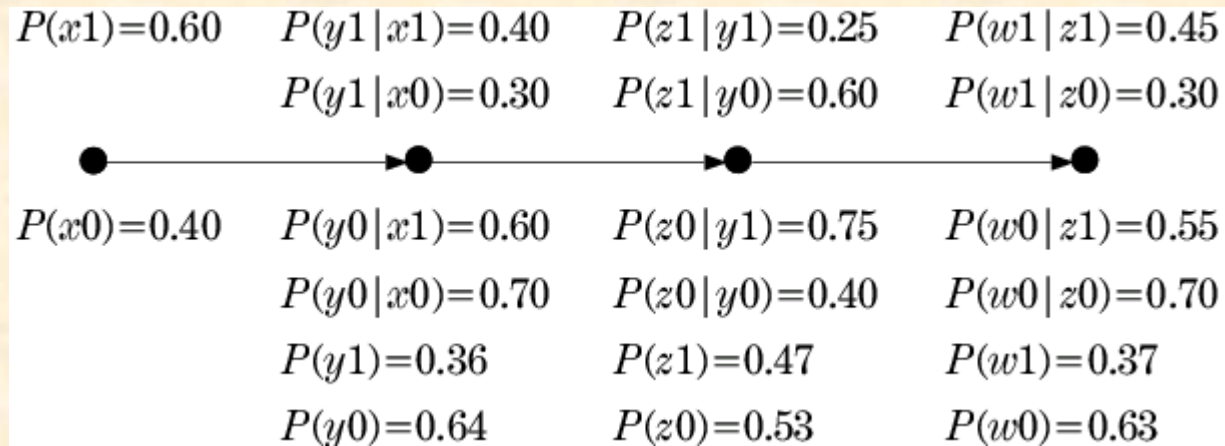
- The figure below is an example of a Bayesian Network corresponding to a paradigm from the medical applications field.



- This Bayesian network models conditional dependencies for an example concerning smokers (S), tendencies to develop cancer (C) and heart disease (H), together with variables corresponding to heart (H1, H2) and cancer (C1, C2) medical tests.

- Once a DAG has been constructed, the joint probability can be obtained by **multiplying the marginal** (root nodes) and the **conditional** (non-root nodes) probabilities.
- **Training**: Once a topology is given, probabilities are estimated via the training data set. There are also methods that learn the topology.
- **Probability Inference**: This is the most common task that Bayesian networks help us to solve **efficiently**. Given the values of some of the variables in the graph, known as **evidence**, the goal is to compute the conditional probabilities for some of the other variables, **given the evidence**.

❖ **Example:** Consider the Bayesian network of the figure:



a) If x is measured to be $x=1$ ($x1$), compute $P(w=0|x=1)$ [$P(w0|x1)$].

b) If w is measured to be $w=1$ ($w1$) compute $P(x=0|w=1)$ [$P(x0|w1)$].

- For a), a set of calculations are required that **propagate** from node x to node w . It turns out that $P(w0|x1) = 0.63$.
- For b), the **propagation** is reversed in direction. It turns out that $P(x0|w1) = 0.4$.
- In general, the required inference information is computed via a combined process of “**message passing**” among the nodes of the DAG.

❖ Complexity:

- For singly connected graphs, message passing algorithms amount to a complexity **linear** in the **number of nodes**.