CLASSIFIERS BASED ON BAYES DECISION THEORY

CLASSIFIERS BASED ON BAYES DECISION THEORY

- Statistical nature of feature vectors

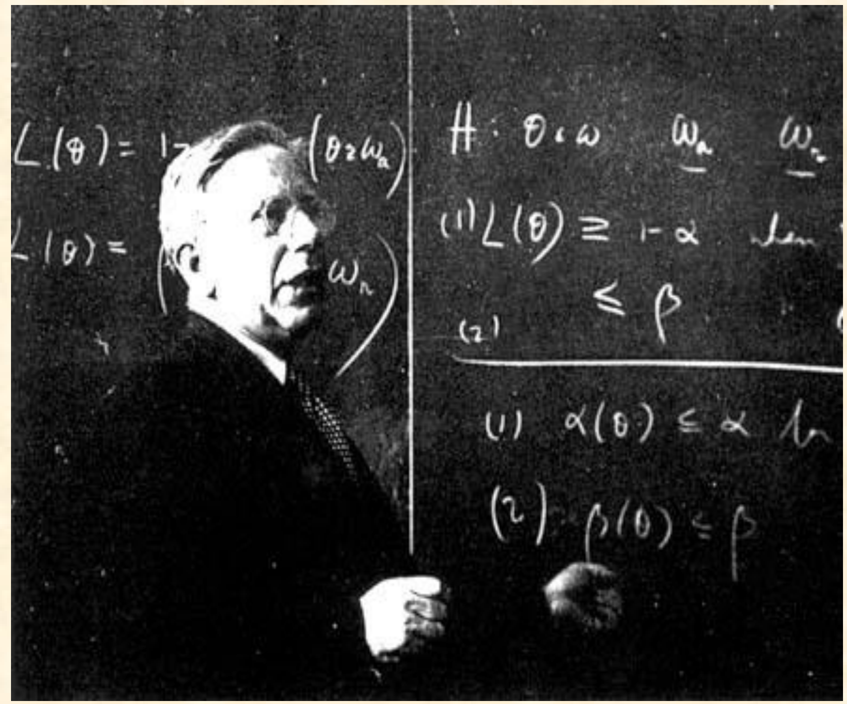
$$\underline{x} = \left[x_1, x_2, \ldots, x_l\right]^T$$

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

CLASSIFIERS BASED ON BAYES DECISION THEORY



Abraham Wald (1902-1950)

Thomas Bayes (1707-1761)

- 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

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

- 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

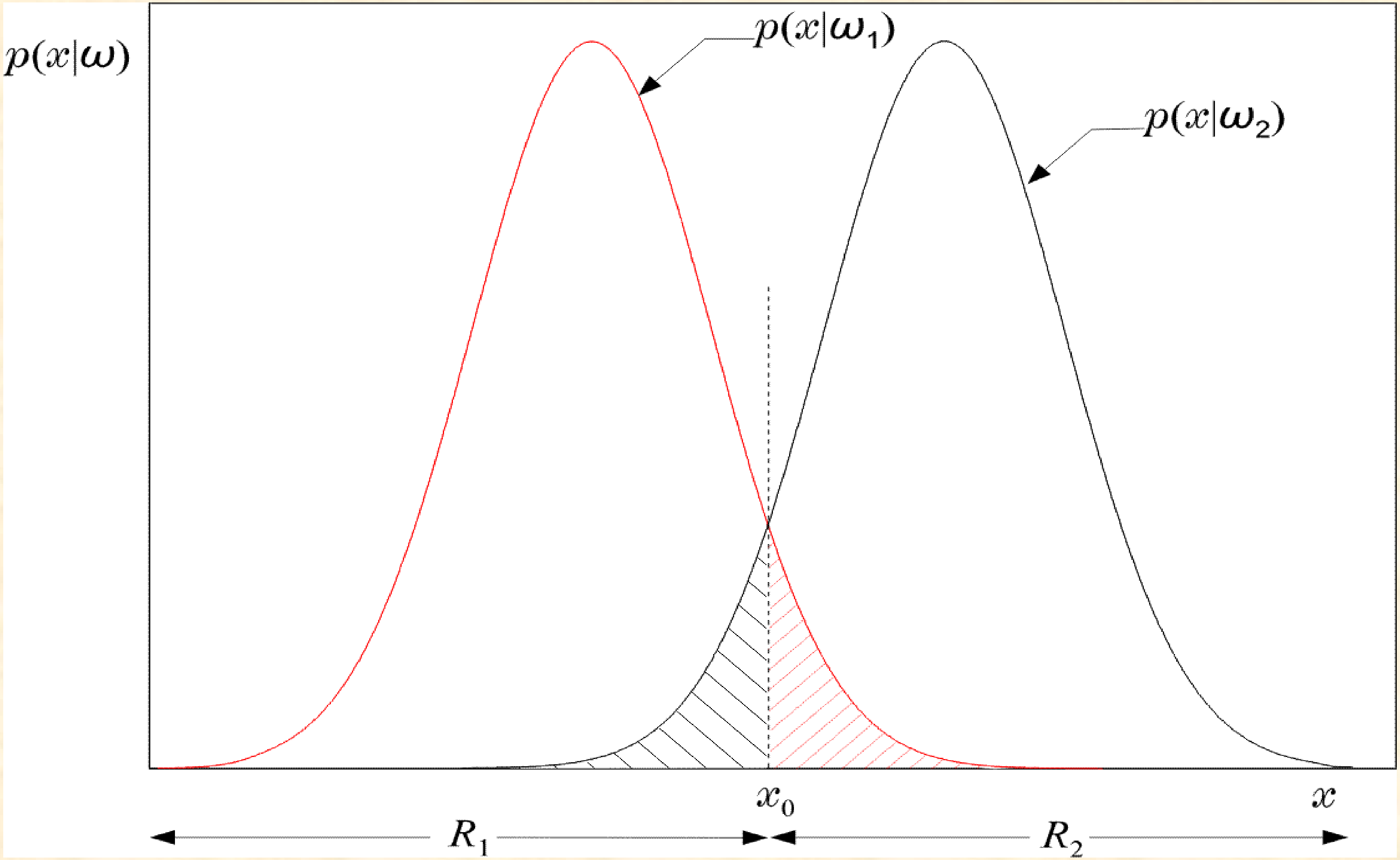
$$
\begin{aligned}
&If\ P(\omega_1|\underline{x}) > P(\omega_2|\underline{x})\ \ \underline{x} \rightarrow \omega_1 \\
&If\ P(\omega_2|\underline{x}) > P(\omega_1|\underline{x})\ \ \underline{x} \rightarrow \omega_2
\end{aligned}
$$

- 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

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



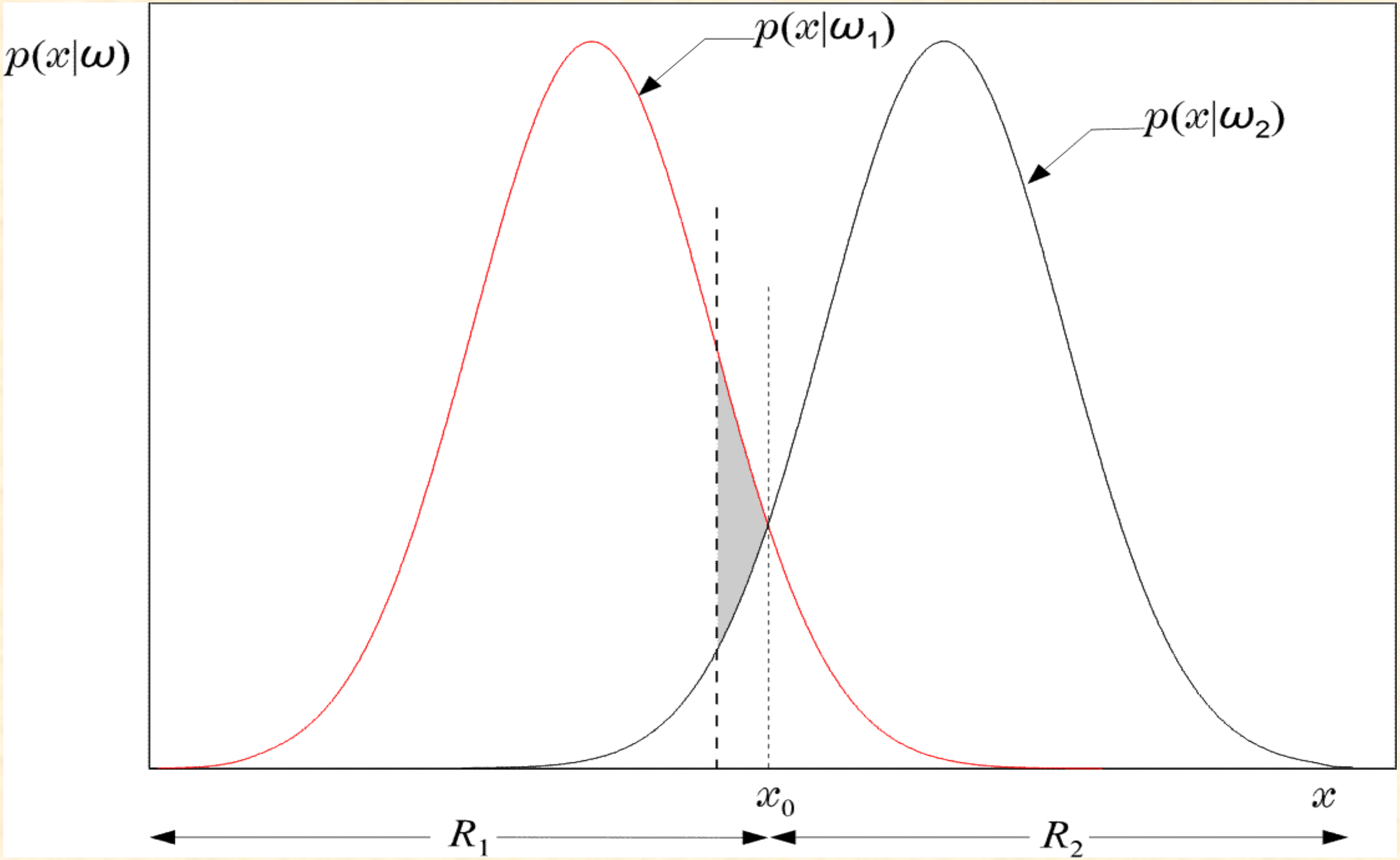
$$R_1(\rightarrow \omega_1) \ and \ R_2(\rightarrow \omega_2)$$

- Equivalently in words: Divide space in two regions

$$\text{If } \underline{x} \in R_1 \Rightarrow \underline{x} \text{ in } \omega_1$$
$$\text{If } \underline{x} \in R_2 \Rightarrow \underline{x} \text{ in } \omega_2$$

- Probability of error
  - Total shaded area
  - $P_e = \int_{-\infty}^{x_0} p(x|\omega_2)dx + \int_{x_0}^{+\infty} p(x|\omega_1)dx$

- 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

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

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

- Equivalently:

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

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

ℓ_{12} : likelihood ratio

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

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

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

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

❖ 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:

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

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

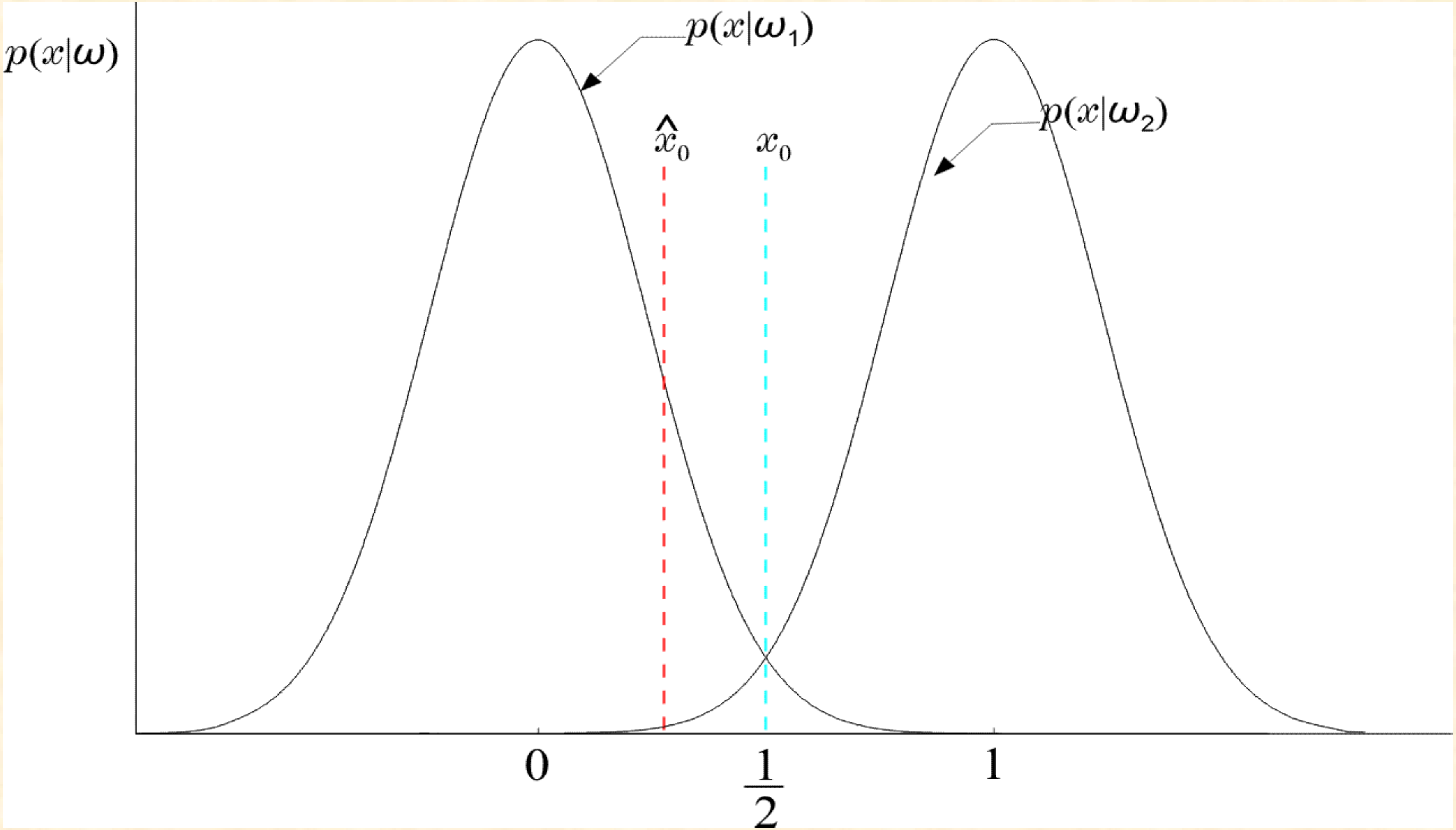
$$x_0 = \frac{1}{2}$$

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

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

$$\hat{x}_0 = \frac{(1-\ell n2)}{2} < \frac{1}{2}$$

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: $g(\underline{x}) \equiv P(\omega_i | \underline{x}) - P(\omega_j | \underline{x}) = 0$

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

+

-

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

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

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

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

$$\underline{\mu}_i = E[\underline{x}]$$

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

called covariance matrix

- $\ln(\cdot)$ 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: $$\Sigma_i = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

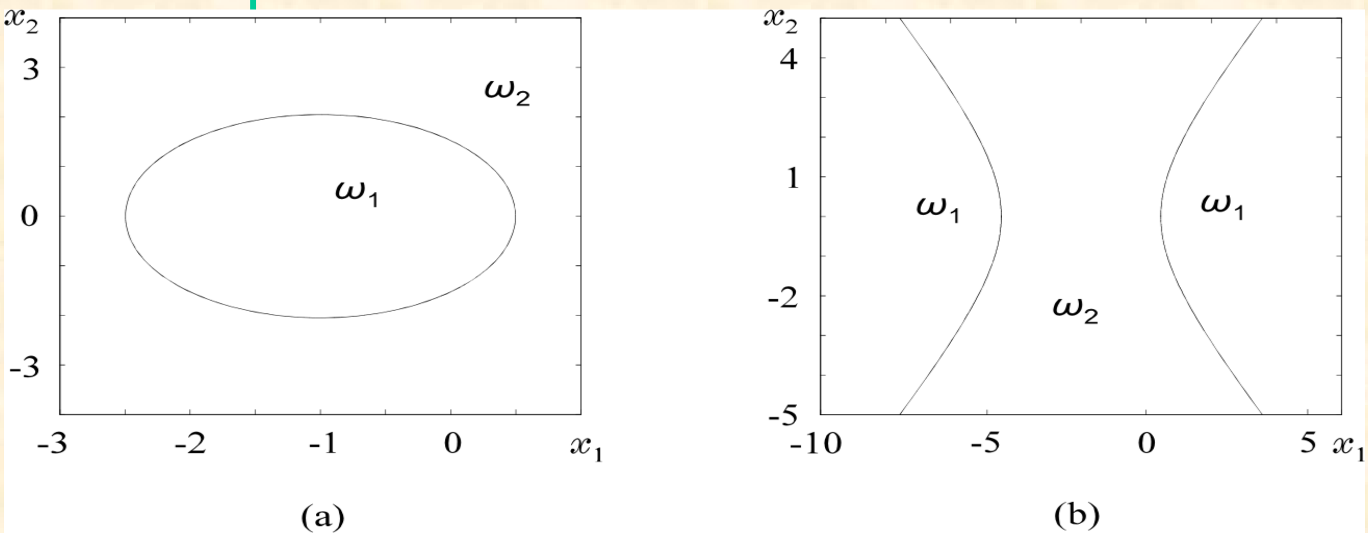
$$g_i(\underline{x}) = -\frac{1}{2\sigma^2}(x_1^2 + x_2^2) + \frac{1}{\sigma^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln(P\omega_i) + C_i$$

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.

For example:



❖ 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:

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

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

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

Discriminant functions are LINEAR

➢ Let in addition:

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

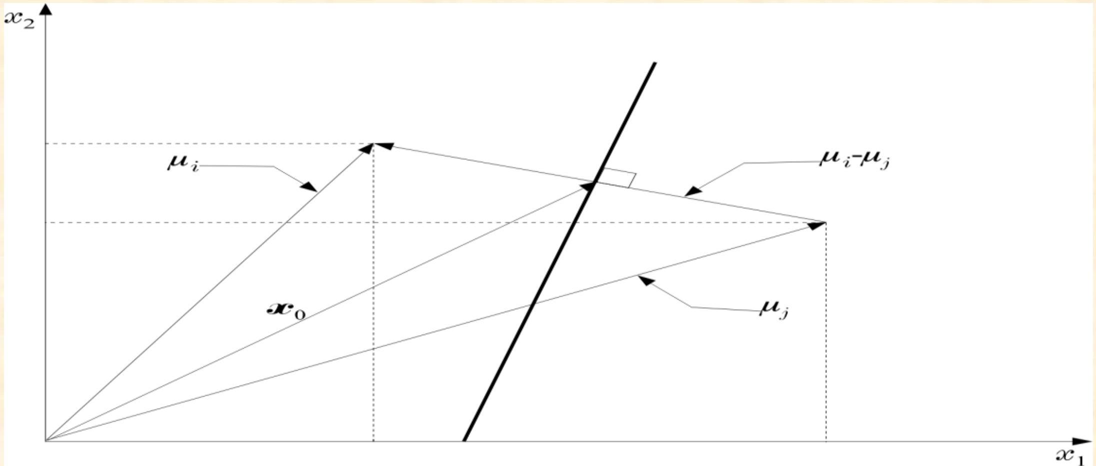
$$g_i(\underline{x}) = \frac{1}{\sigma^2}\underline{\mu}_i^T \underline{x} + w_{i0}$$

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

$$= \underline{w}^T(\underline{x} - \underline{x}_o)$$

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

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



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

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

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

➢ Decision hyperplane
- not normal to $\underline{\mu}_i - \underline{\mu}_j$
- normal to $\Sigma^{-1}(\underline{\mu}_i - \underline{\mu}_j)$

❖ Minimum Distance Classifiers

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

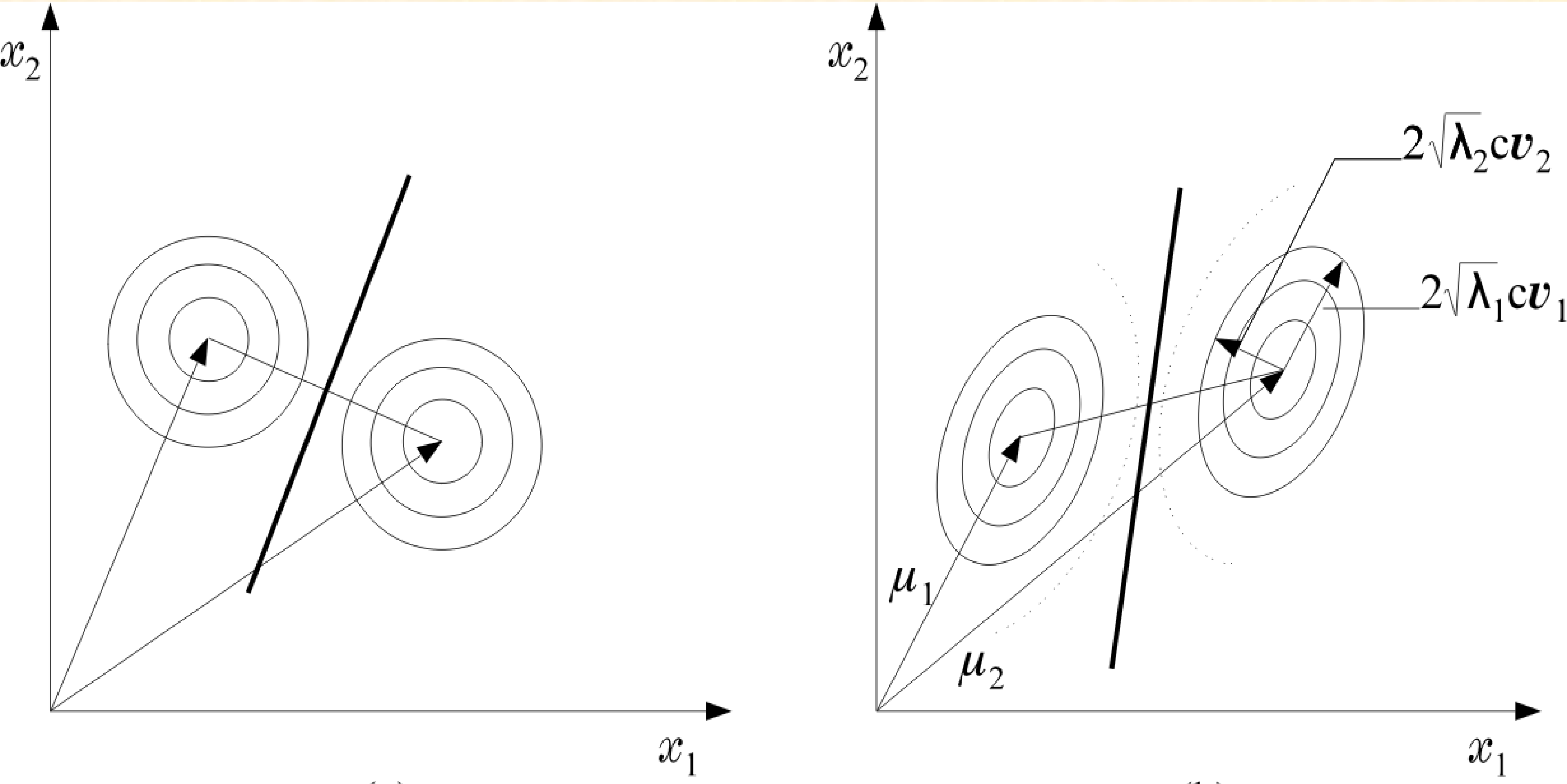
Euclidean Distance: $d_E \equiv \left\| \underline{x} - \underline{\mu}_i \right\|$

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



x_2
x_1
x_2
$2\sqrt{\lambda_2}c\boldsymbol{v}_2$
$2\sqrt{\lambda_1}c\boldsymbol{v}_1$
μ_1
μ_2
x_1

(a) (b)

❖ 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 $d_{E,2} < d_{E,1}$

❖ 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 in known as the curse of dimensionality. This is a major problem one is confronted with in high dimensional spaces.

❖ NAIVE – BAYES CLASSIFIER

- Let $\underline{x} \in \Re^{\ell}$ and the goal is to estimate $p(\underline{x} \mid \omega_i)$ $i = 1, 2, \ldots, M$. For a "good" estimate of the pdf one would need, say, N^{ℓ} points.

- Assume $x_1, x_2, \ldots, x_{\ell}$ mutually independent. Then:

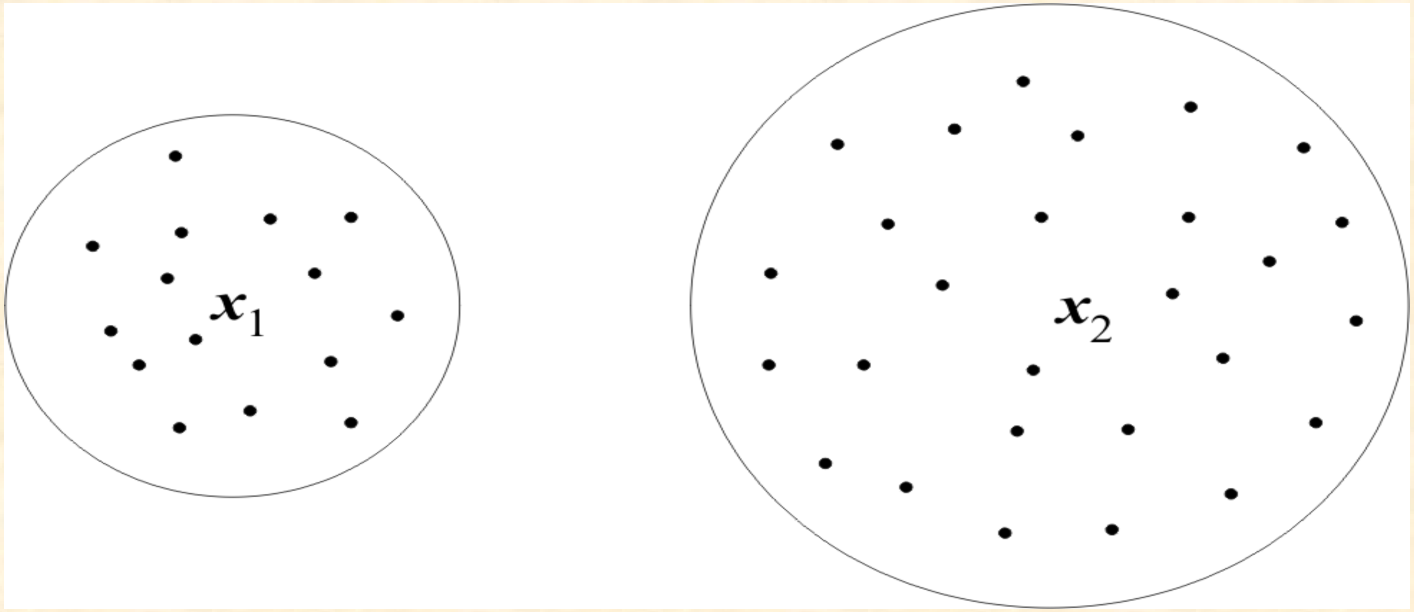
$$p(\underline{x} \mid \omega_i) = \prod_{j=1}^{\ell} p(x_j \mid \omega_i)$$

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

- 

$$\frac{\dfrac{k}{N_1V_1}}{\dfrac{k}{N_2V_2}} = \frac{N_2V_2}{N_1V_1}(><)\theta$$

❖ 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:

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

- $P_B \leq P_{kNN} \leq P_B + \sqrt{\dfrac{2P_{NN}}{k}}$

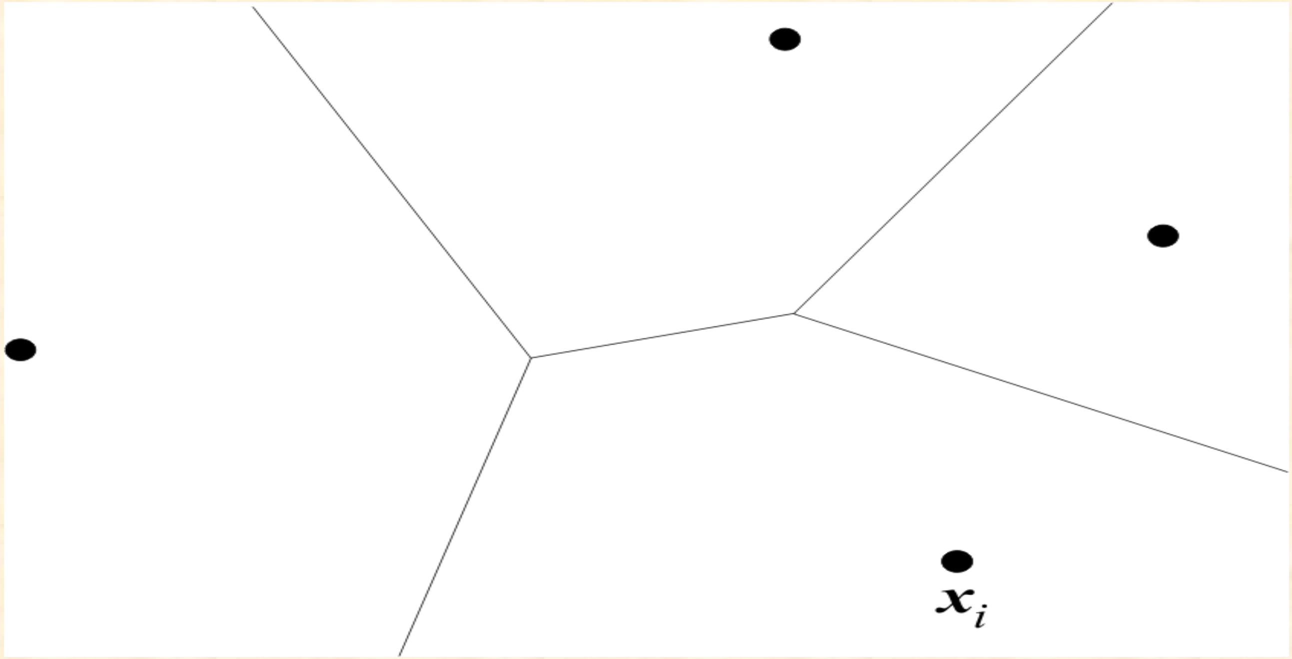
- $k \to \infty, \; P_{kNN} \to P_B$

- For small P_B:

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

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

❖ Voronoi tesselation



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

BAYESIAN NETWORKS

❖ Bayes Probability Chain Rule

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

$$\ldots \cdot p(x_2 \mid x_1) \cdot p(x_1)$$

- 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:

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

where

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

- For example, if ℓ=6, then we could assume:

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

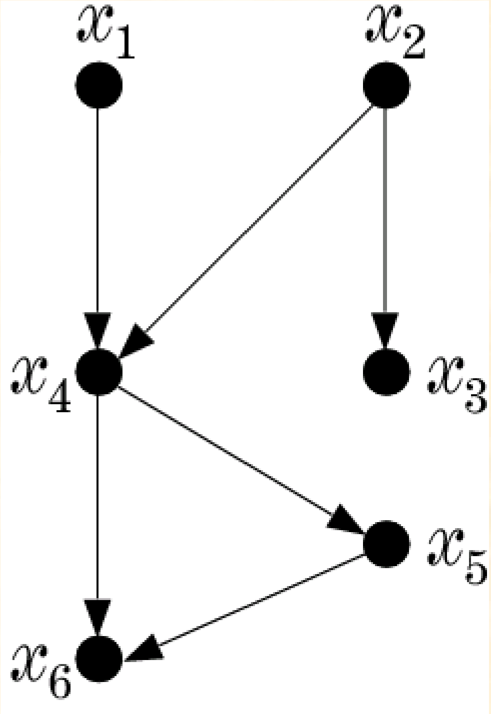
Then:

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

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

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

- 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:

$$p(x_1, x_2, \ldots, x_6) =$$

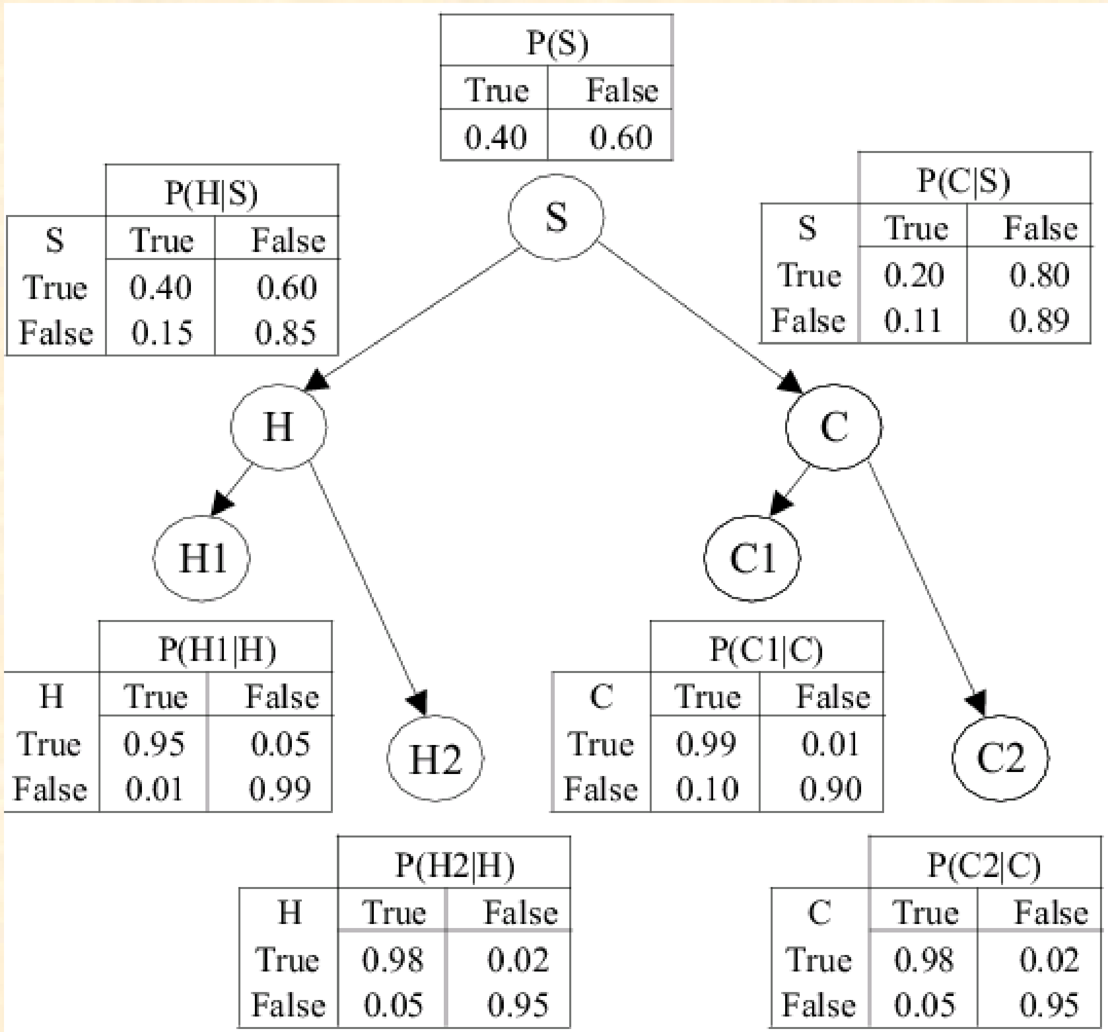
$$p(x_6 \mid x_5, x_4) \cdot p(x_5 \mid x_4) \cdot p(x_4 \mid x_2, x_1) \cdot p(x_3 \mid x_2) \cdot p(x_2) \cdot p(x_1)$$

❖ 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.



P(S)	
True	False
0.40	0.60

	P(H|S)	
S	True	False
True	0.40	0.60
False	0.15	0.85

	P(C|S)	
S	True	False
True	0.20	0.80
False	0.11	0.89

	P(H1|H)	
H	True	False
True	0.95	0.05
False	0.01	0.99

	P(C1|C)	
C	True	False
True	0.99	0.01
False	0.10	0.90

	P(H2|H)	
H	True	False
True	0.98	0.02
False	0.05	0.95

	P(C2|C)	
C	True	False
True	0.98	0.02
False	0.05	0.95

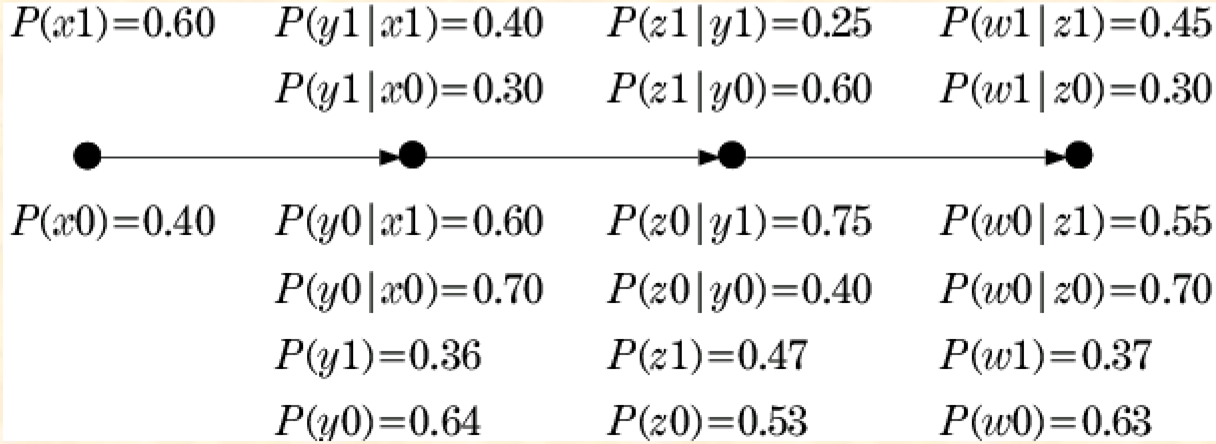
- This Bayesian network models conditional dependencies for an example concerning smokers (S), tendencies to develop cancer (C) and heart disease (H), together with variables corresponding to heart (H1, H2) and cancer (C1, C2) medical tests.

- 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 $P(x0|w1) = 0.4$.

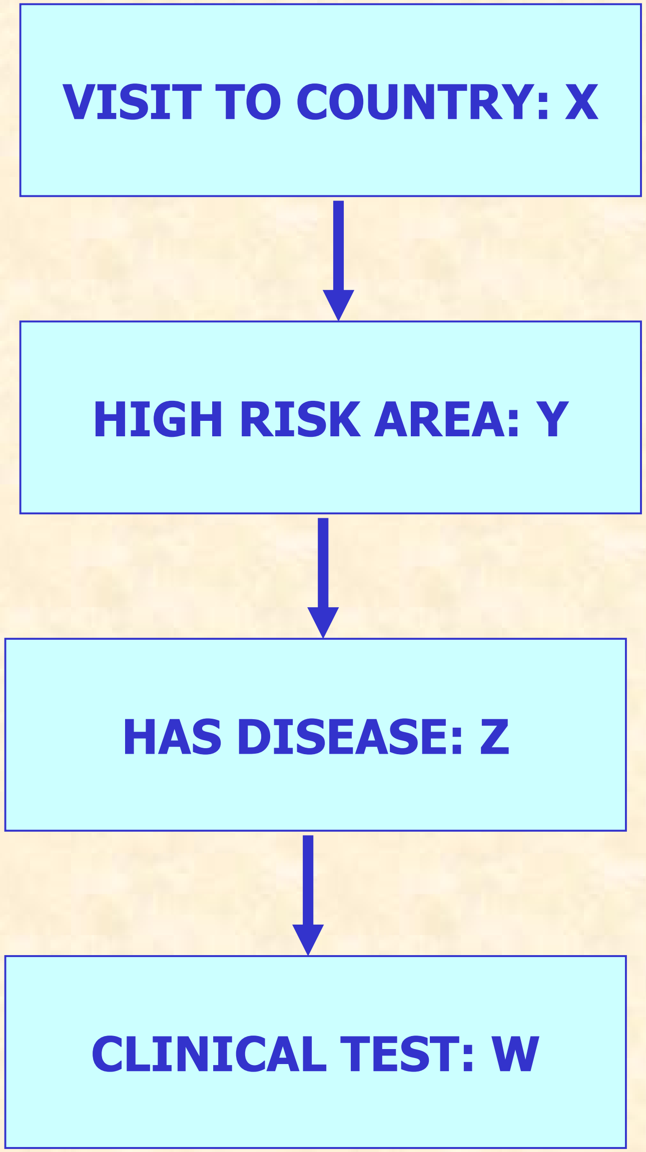
- In general, the required inference information is computed via a combined process of "message passing" among the nodes of the DAG.

Complexity:

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

Example

- 0 = NO
- 1 = YES



$$P(x_1) = 0{,}1$$
$$P(x_0) = 0{,}9$$

$$P(y_1 \mid x_1) = 0{,}3$$
$$P(y_1 \mid x_0) = 0{,}05$$

$$P(z_1 \mid y_1) = 0{,}5$$
$$P(z_1 \mid y_0) = 0{,}02$$

$$P(w_1 \mid z_1) = 0{,}95$$
$$P(w_1 \mid z_0) = 0{,}03$$

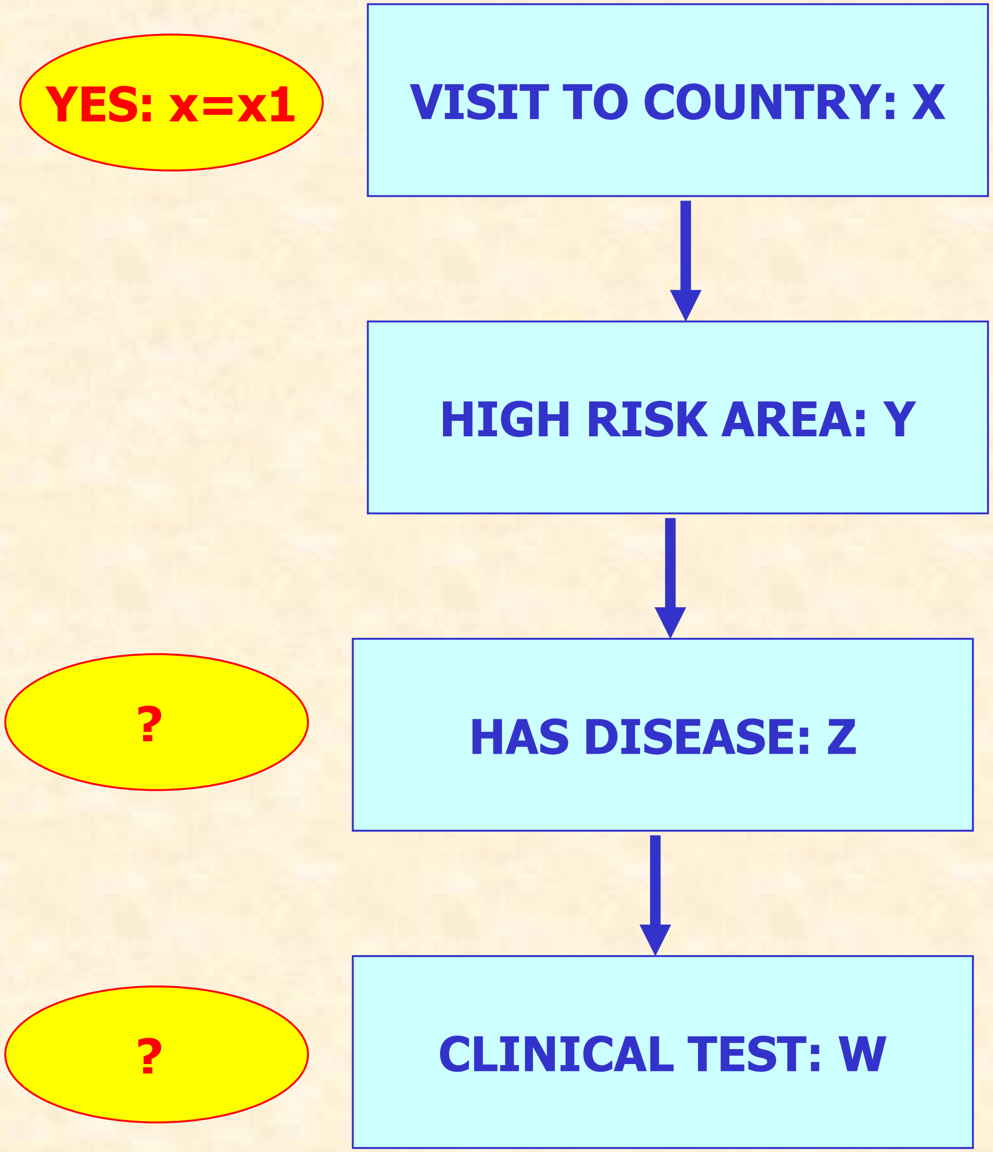
Inference

Answer questions like:

- What is the probability of a person having caught the disease given that he/she has visited the high risk country?
- What is the probability of the clinical test of someone coming out positive, given that he/she has visited the high risk country?
- Given that the clinical test of a person has come out positive, what is the probability that he/she has visited the high risk country?
- Given that the clinical test of a person has come out positive, what is the probability that he/she has the disease?

Inference

- 0 = NO
- 1 = YES



$$P(x_1) = 0,1$$
$$P(x_0) = 0,9$$

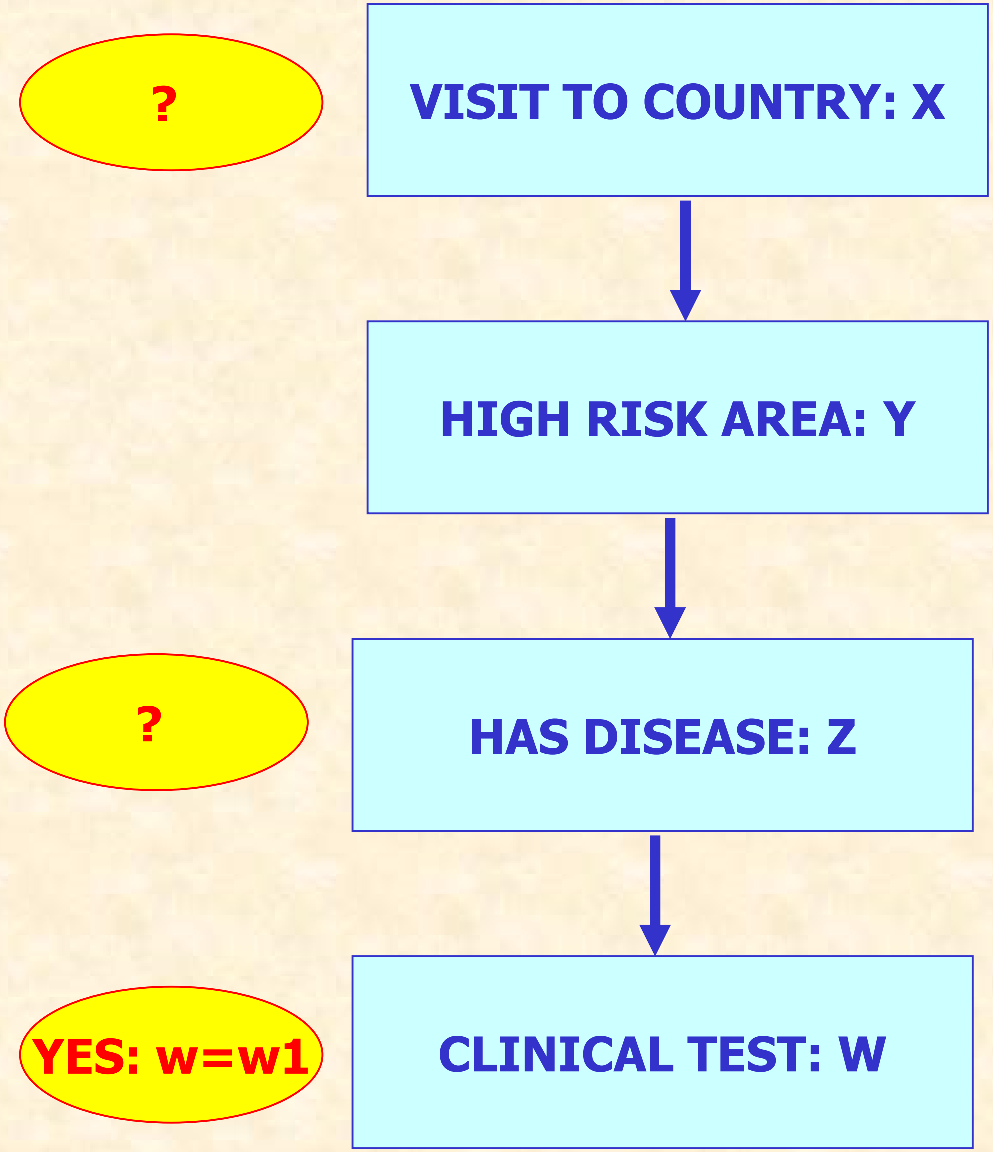
$$P(y_1 | x_1) = 0,3$$
$$P(y_1 | x_0) = 0,05$$

$$P(z_1 | y_1) = 0,5$$
$$P(z_1 | y_0) = 0,02$$

$$P(w_1 | z_1) = 0,95$$
$$P(w_1 | z_0) = 0,03$$

Inference

- 0 = NO
- 1 = YES



$$P(x_1) = 0{,}1$$
$$P(x_0) = 0{,}9$$

$$P(y_1 \mid x_1) = 0{,}3$$
$$P(y_1 \mid x_0) = 0{,}05$$

$$P(z_1 \mid y_1) = 0{,}5$$
$$P(z_1 \mid y_0) = 0{,}02$$

$$P(w_1 \mid z_1) = 0{,}95$$
$$P(w_1 \mid z_0) = 0{,}03$$