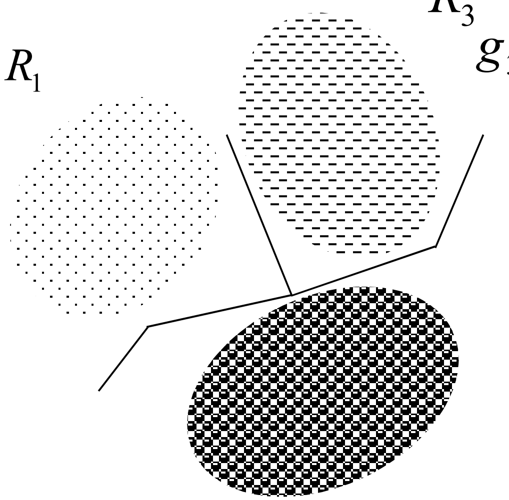
PART 2: Statistical Pattern Classification: Optimal Classification with Bayes Rule

Statistical Approach to P.R

$X = [X_1, X_2, ..., X_d]$



Dimension of the feature space: d

Set of different states of nature: $\{\omega_1, \omega_2, ..., \omega_c\}$

Categories: c

find R_i $\quad R_i \cap R_j = \varphi \quad uR_i = R^d$

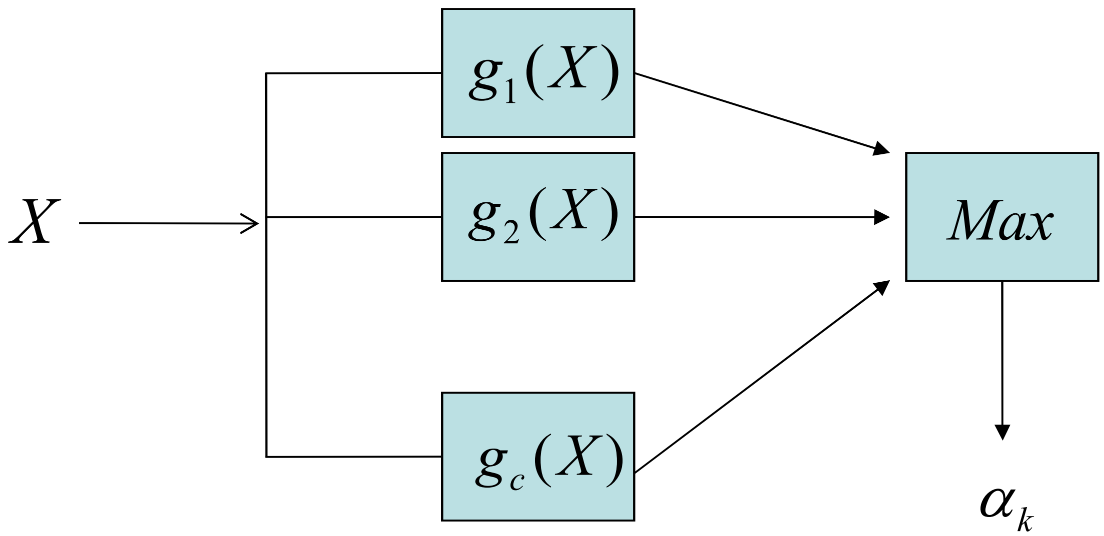
set of possible actions (decisions): $\{\alpha_1, \alpha_2, ..., \alpha_a\}$

Here, a decision might include a 'reject option'

A Discriminant Function $g_i(X) \geq g_j(X) \quad g_i(X) \quad 1 \leq i \leq c$

in region R_i; decision rule : α_k if $\quad g_k(X) > g_j(X)$

A Pattern Classifier



So our aim now will be to define these functions $g_1, g_2, \ldots, g_c$ to *minimize* or *optimize* a criterion.

Parametric Approach to Classification

- 'Bayes DecisionTheory' is used for minimum-error/minimum risk pattern classifier design.

- Here, it is assumed that if a sample X is drawn from a class ω_i it is a random variable represented with a multivariate probability density function.

'Class- conditional density function'

$$P(X|\omega_i)$$

- We also know a-priori probability $P(\omega_i)$

$$1 \leq i \leq c \quad \text{(c is no. of classes)}$$

- Then, we can talk about a decision rule that minimizes the probability of error.

- Suppose we have the observation X
- This observation is going to change a-priori assumption to a-posteriori probability:

$$P(\omega_i | X)$$

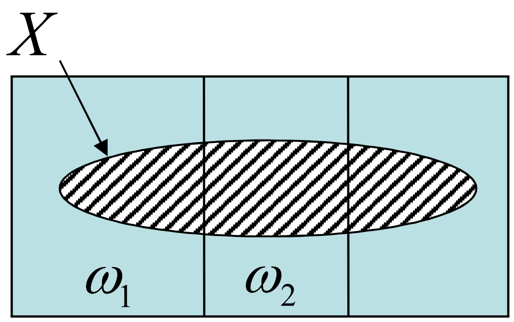
- which can be found by the Bayes Rule.

$$P(\omega_i|X) = P(\omega_i, X) / P(X)$$

$$= \frac{P(X|\omega_i).P(\omega_i)}{P(X)}$$

- $P(X)$ can be found by Total Probability Rule:

When ω_i's are disjoint,



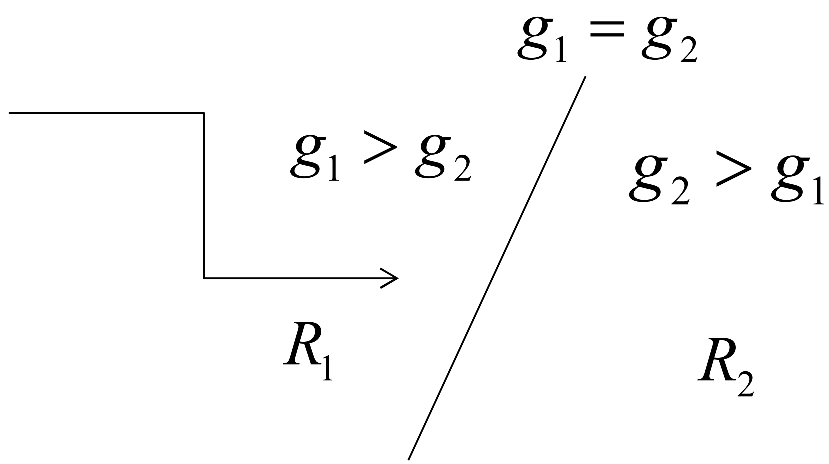
$$P(X) = \sum_{i=1}^{c} P(\omega_i, X)$$

$$P(X) = \sum_{i=1}^{c} P(X|\omega_i).P(\omega_i)$$

- *Decision Rule: Choose the category with highest a-posteriori probability, calculated as above, using Bayes Rule.*

then, $g_i(X) = P(\omega_i | X)$ 1

Decision boundary:



or in general, decision boundaries are where:

$$g_i(X) = g_j(X)$$

between regions R_i and R_j

- Single feature – decision boundary – point
  2 features – curve
  3 features – surface
  More than 3 – hypersurface

$$g_i(X) = P(X|\omega_i).P(\omega_i)$$

$$gi(X) = \frac{P(X|\omega_i).P(\omega_i)}{P(X)}$$

- Sometimes, it is easier to work with logarithms

$$g_i(X) = \log[P(X|\omega_i).P(\omega_i)]$$

$$g_i(X) = \log P(X|\omega_i) + \log P(\omega_i)$$

- Since logarithmic function is a monotonically increasing function, log fn will give the same result.

2 Category Case: c_1, c_2

Assign to c_1 if (α_1) $P(\omega_1|X) > P(\omega_2|X)$

c_2 if (α_2) $P(\omega_1|X) < P(\omega_2|X)$

But this is the same as:

c_1 if

$$\frac{P(X|\omega_1).P(\omega_1)}{P(X)} > \frac{P(X|\omega_2).P(\omega_2)}{P(X)}$$

By throwing away $P(X)$'s, we end up with:

c_1 if

$$P(X|\omega_i).P(\omega_1) > P(X|\omega_2).P(\omega_2)$$

Which the same as:

$$\frac{P(X|\omega_1)}{P(X|\omega_2)} > \frac{P(X|\omega_2)}{P(X|\omega_1)} = k$$

Likelihood ratio

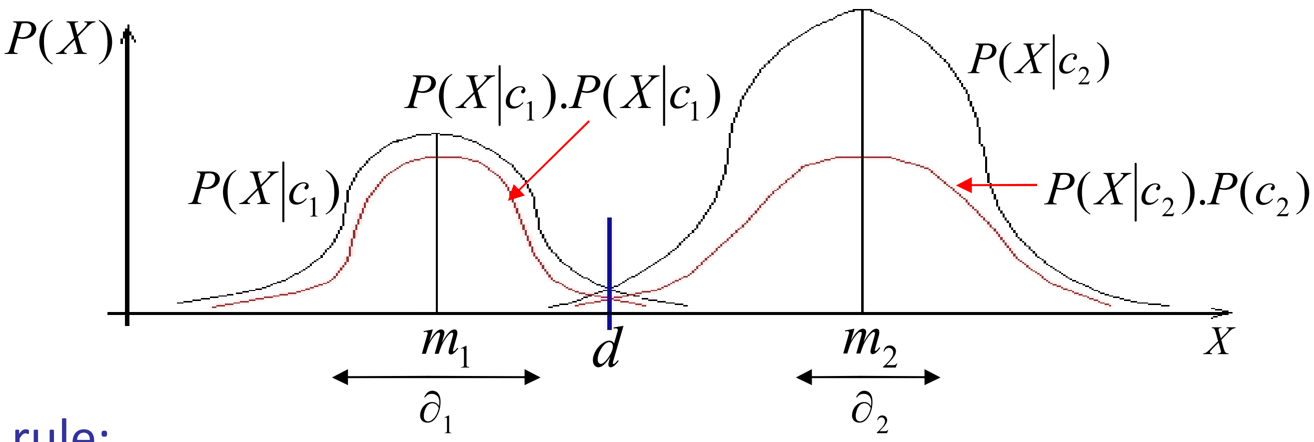
Example: a single feature, 2 category problem with gaussian density

: Diagnosis of diabetes using sugar count X

c_1 state of being healthy $P(c_1) = 0.7$

c_2 state of being sick (diabetes) $P(c_2) = 0.3$

$P(X\|c_1) = \frac{1}{\sqrt{2\pi\partial_1^2}}.e^{-(X-m_1)^2/2\partial_1^2}$	$P(X\|c_2) = \frac{1}{\sqrt{2\pi\partial_2^2}}.e^{-(X-m_2)^2/2\partial_2^2}$



The decision rule:

c_1 if $P(X|c_1).P(c_1) > P(X|c_2).P(c_2)$ or

$$0.7P(X|c_1) > 0.3P(X|c_2)$$

Assume now: $m_1 = 10$ $m_2 = 20$ $\partial_1 = \partial_2 = 2$

And we measured: $X = 17$

Assign the unknown sample: X to the correct category.

Find likelihood ratio: $$= \frac{e^{-(X-10)^2/8}}{e^{-(X-20)^2/8}}$$ for $X = 17$

$$= e^{-4.9} = 0.006$$

Compare with: $$\frac{P(c_2)}{P(c_1)} = \frac{0.3}{0.7} = 0.43 > 0.006$$

So assign: X to c_2.

Example: A discrete problem

Consider a 2-feature, 3 category case

where:

$$P(X_1, X_2 | c_i) = \begin{cases} = \dfrac{1}{(a_i - b_i)^2} & for \quad \begin{matrix} a_i < X_1 < b_i \\ a_i < X_2 < b_i \end{matrix} \\ = 0 & other\ wise \end{cases}$$

And $P(c_1) = 0.4$, $P(c_2) = 0.4$, $P(c_3) = 0.2$

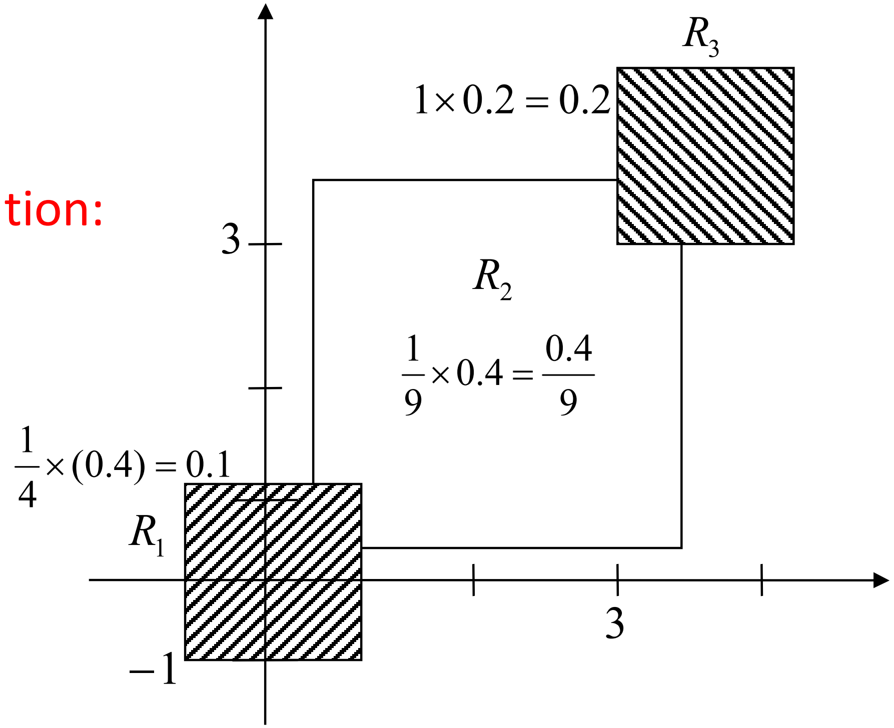
Find the decision boundaries and regions:

$a_1 = -1 \qquad b_1 = 1$

$a_2 = 0.5 \qquad b_2 = 3.5$

$a_3 = 3 \qquad b_3 = 4$

Solution:



Remember now that for the 2-class case:

if c_1 $\quad P(X|c_1).P(c_1) > P(X|c_2).P(c_2)$

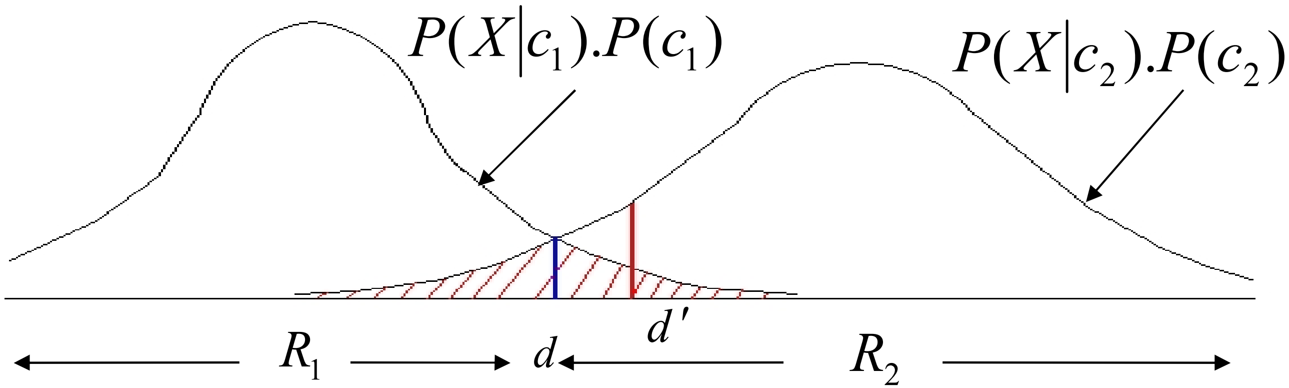
or

$$\frac{P(X|c_1)}{P(X|c_2)} > \frac{P(X|c_2)}{P(X|c_1)} = k$$

Likelihood ratio

Error probabilities and a simple proof of minimum error

Consider again a 2-class 1-d problem:



Let's show that: if the decision boundary is d (intersection point) rather than any arbitrary point d'.

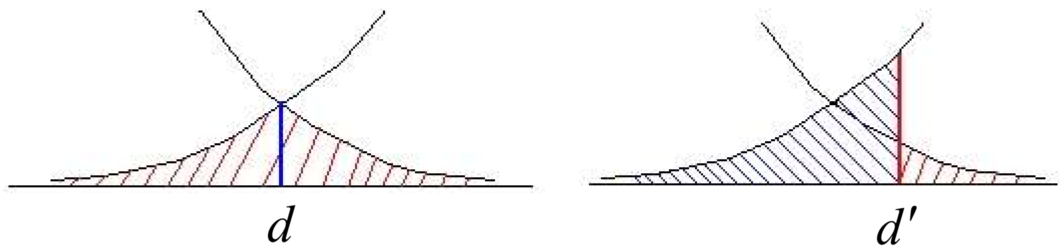
Then $P(E)$ (probability of error) is minimum.

$$P(E) = P(X \in R_2, c_1) + P(X \in R_1, c_2)$$

$$= P(X \in R_2 | c_1).P(c_1) + P(X \in R_1 | c_2).P(c_2)$$

$$= [\int_{R_2} P(X|c_1)dX].P(c_1) + [\int_{R_1} P(X|c_2)dX].P(c_2)$$

$$= \int_{R_2} P(X|c_1).P(c_1)dX + \int_{R_1} P(X|c_2).P(c_2)dX$$



It can very easily be seen that the $P(E)$ is minimum if $d' = d$.

Minimum Risk Classification

Risk associated with incorrect decision might be more important than the probability of error.

So our decision criterion might be modified to minimize the *average risk* in making an incorrect decision.

We define a conditional risk (expected loss) for decision α_i when X occurs as:

$$R^i(X) = \sum_{j=1}^{c} \lambda(\alpha_i | \omega_j).P(\omega_j | X)$$

Where $\lambda(\alpha_i | \omega_j)$ is defined as the conditional loss associated with decision α_i when the true class is ω_j . It is assumed that λ is known.

The decision rule: decide on c_i if $R^i(X) < R^j(X)$

for all $1 \le j \le c \quad i \neq j$

The discriminant function here can be defined as: $g_i(X) = -R^i(X)$

- We can show that minimum – error decision is a special case of above rule where:

$$\lambda(\alpha_i | \omega_i) = 0$$

$$\lambda(\alpha_i | \omega_j) = 1$$

then, $$R^i(X) = \sum_{j \neq i} P(\omega_j | X)$$

$$= 1 - P(\omega_i | X)$$

so the rule is α_i if $1 - P(\omega_i | X) < 1 - P(\omega_j | X)$

$$\equiv P(\omega_i | X) > R(\omega_j | X)$$

For the 2 - category case, minimum - risk classifier becomes:

$$R^{\alpha_1}(X) = \lambda_{11} P(\omega_1|X) + \lambda_{12} P(\omega_2|X)$$

$$R^{\alpha_2}(X) = \lambda_{22} P(\omega_2|X) + \lambda_{21} P(\omega_1|X)$$

α_1 if $\quad \lambda_{11} P(\omega_1|X) + \lambda_{12} P(\omega_2|X) > \lambda_{22} P(\omega_2|X) + \lambda_{21} P(\omega_1|X)$

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

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

α_1 if $\quad \dfrac{P(X|\omega_1)}{P(X|\omega_2)} > \dfrac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \cdot \dfrac{P(\omega_2)}{P(\omega_1)}$

Otherwise, α_2.

This is the same as likelihood rule if $\lambda_{22} = \lambda_{11} = 0$ and $\lambda_{12} = \lambda_{21} = 1$

Discriminant Functions so far

For Minimum Error: $+P(\omega_i|X)$

$$+P(X|\omega_i).P(\omega_i)$$

$$+\log P(X|\omega_i)+\log P(\omega_i)$$

For Minimum Risk: $-R^i(X)$

Where

$$R^i(X)=\sum_{j=1}^{c}\lambda(\alpha_i|\omega_j).P(\omega_j|X)$$

Bayes (Maximum Likelihood)Decision:

- Most general optimal solution
- Provides an upper limit(you cannot do better with other rule)
- Useful in comparing with other classifiers

Special Cases of Discriminant Functions

Multivariate Gaussian (Normal) Density $N(M, \Sigma)$:

The general density form: $P(X) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-1/2 (X-M)^T \Sigma^{-1} (X-M)}$

Here X in the feature vector of size $.d$

M : d element mean vector $E(X) = M = [\mu_1, \mu_2, \ldots, \mu_d]^T$

Σ_{dxd} : covariance matrix

$$\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$$

$$\Sigma_{ii} = E[(X_i - \mu_i)^2]$$

$(= \sigma_i^2$ variance of feature X_i)

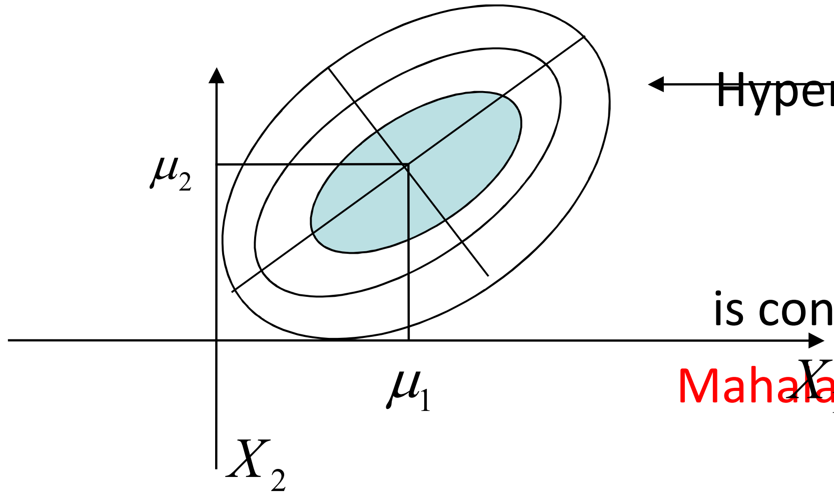
Σ - symmetric

$\Sigma_{ij} = 0$ when X_i and X_j are statistically independent.

$|\Sigma|$ - determinant of Σ

General shape: Hyper ellipsoids

where



$$(X-M)^T \Sigma^{-1} (X-M)$$

is constant:

Mahalanobis

Distance

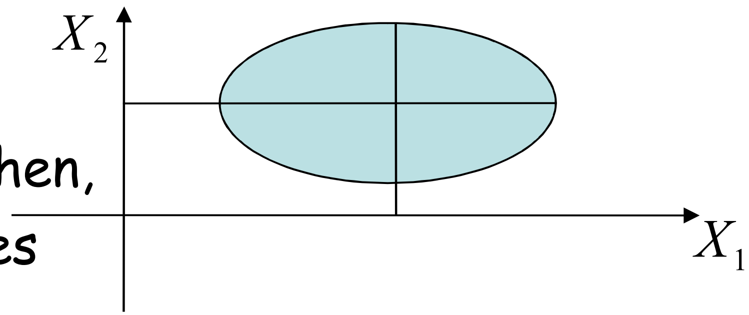
$$M = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^{\ 2} & \Sigma_{12} \\ \Sigma_{21} & \sigma_2^{\ 2} \end{bmatrix}$$

2 - d problem: $X_1, \quad X_2$,

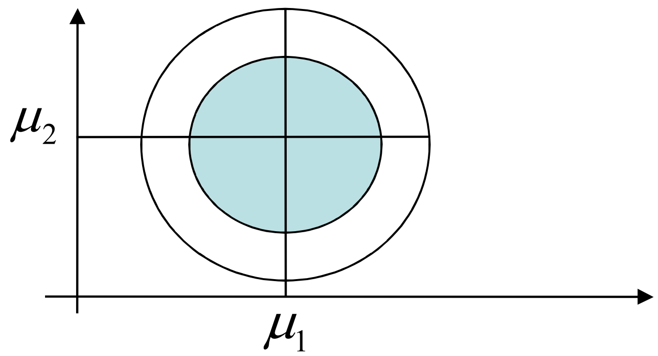
If $\Sigma_{12} = 0$, $\Sigma_{21} = 0$

(statistically independent features) then,

major axes are parallel to major ellipsoid axes



if in addition $\sigma_1{}^2 = \sigma_2{}^2$



in general, the equal density curves are hyper ellipsoids. Now

$$g_i(X) = \log_e P(X|\omega_i) + \log_e P(\omega_i)$$

is used for $N(M_i, \Sigma_i)$ since its ease in manipulation

---

$$g_i(X) = -(1/2).(X - M_i)^T \sum_i^{-1} (X - M_i)$$
$$-(1/2)\log|\Sigma_i| + \log P(\omega_i)$$

$g_i(X)$ is a quadratic function of X as will be shown.

$$g_i(X) = -1/2.X^T \sum_i^{-1} X - 1/2.M_i^T \sum_i^{-1} M_i$$
$$+1/2.X^T \sum_i^{-1} M_i + 1/2 M_i^T \sum_i^{-1} X$$
$$-1/2.\log\left|\Sigma_i^{-1}\right| + \log P(\omega_i)$$

$$W_i = -1/2.\sum_i^{-1}$$

$$V_i = M_i^T \sum_i^{-1}$$

a scalar $W_{io} = -1/2.M_i^T \Sigma_i^{-1} M_i - 1/2.\log\left|\Sigma_i\right| + \log P(\omega_i)$

Then,

$$g_i(X) = X^T W_i X + V_i X + W_{io}$$

On the decision boundary,

$$g_i(X) = g_j(X)$$

$$X^T W_i X - X^T W_j X + V_i X - V_j X + W_{io} - W_{jo} = 0$$

$$X^T(W_i - W_j)X + (V_i - V_j)X + (W_{io} - W_{jo}) = 0$$

$$X^T WX + VX + W_0 = 0$$

Decision boundary function is hyperquadratic in general.

Example in 2d.

$$W = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}$$

$$V = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$$

$$X = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$

Then, above boundary becomes

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + W_0 = 0$$

$$\omega_{11} x_1^{\ 2} + 2\omega_{12} x_1 x_2 + \omega_{22} x_2^{\ 2} + v_1 x_1 + v_2 x_2 + W_0 = 0$$

General form of hyper quadratic boundary IN 2-d.

The special cases of Gaussian:

Assume $\Sigma_i = \sigma^2 I$

Where I is the unit matrix

$$|\Sigma_i| = \sigma^{2d}$$

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

$$\Sigma_i^{-1} = \frac{1}{\sigma^2} I$$

$$g_i(X) = -\tfrac{1}{2\sigma^2}(X - M_i)^T.(X - M_i) - \tfrac{1}{2}\log \sigma^{2d} + \log P(\omega_i)$$

$$g_i(X) = -\tfrac{1}{2\sigma^2}\left\| X, M_i \right\|^2 + \log P(\omega_i)$$

(not a function of M_i X so can be removed)

Now assume

$$P(\omega_i) = P(\omega_j)$$

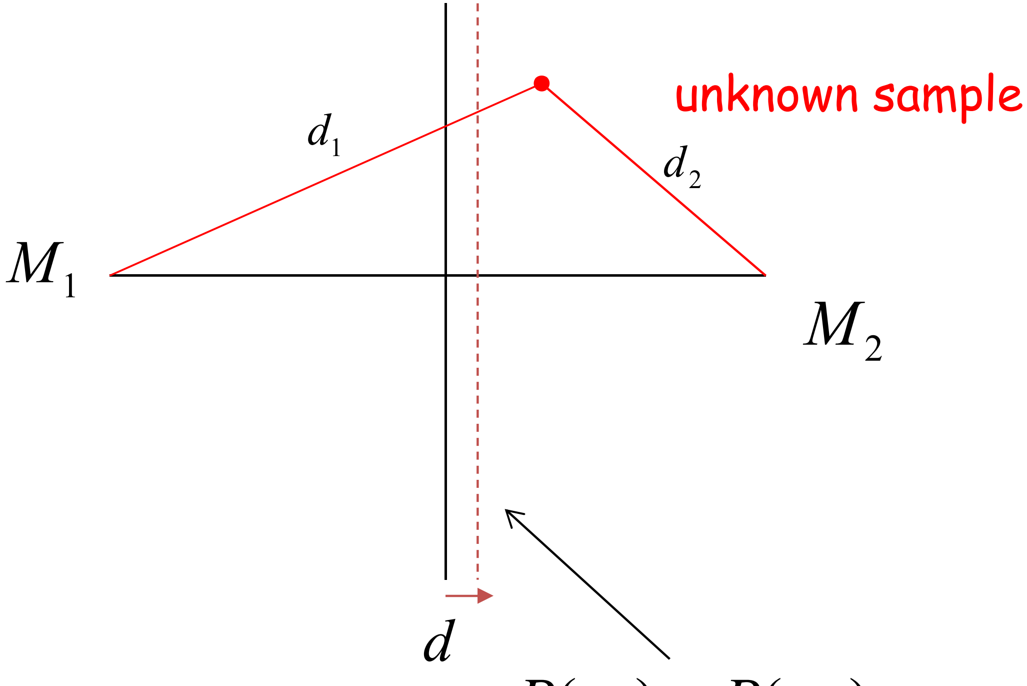
$$g_i(X) = -\tfrac{1}{2\sigma^2}\left\| X, M_i \right\|^2 = -d^2\left(X, M_i\right)$$

euclidian distance between X and Mi

Then,the decision boundary is linear !

Decision

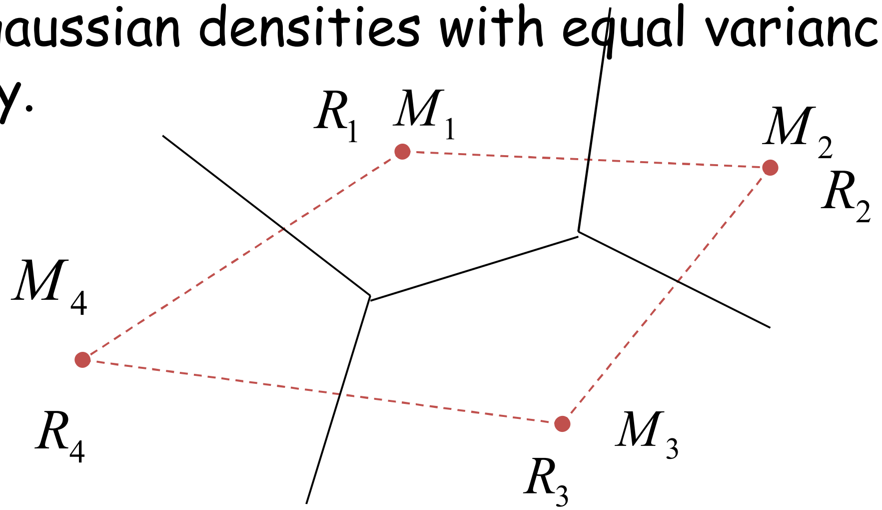
Rule: Assign the unknown sample to the closest mean's category



d= Perpendicular bisector that will move towards the less probable category

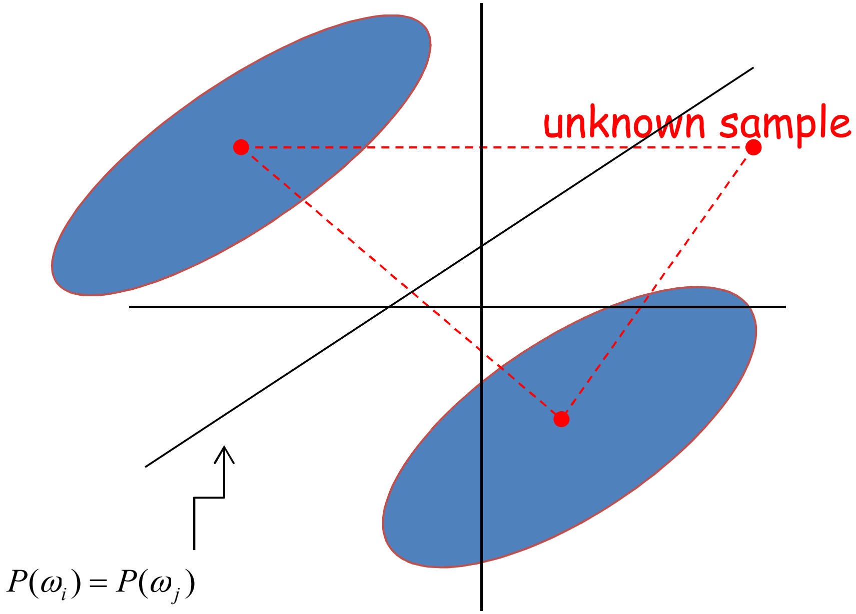
Minimum Distance Classifier

- Classify an unknown sample X to the category with closest mean !
- Optimum when gaussian densities with equal variance and equal a-priori probability.



Piecewise linear boundary in case of more than 2 categories.

- Another special case: It can be shown that when (Covariance matrices are the same) $\Sigma_i = \Sigma$
- Samples fall in clusters of equal size and shape



is called Mahalonobis Distance

$$g_i(X) = -\tfrac{1}{2}(X - M_i)^T \Sigma^{-1}(X - M_i) + \log P(\omega_i)$$

is called Mahalonobis Distance

$$-\tfrac{1}{2}(X - M_i)^T \Sigma^{-1}(X - M_i)$$

Then, if $P(\omega_i) = P(\omega_j)$

The decision rule:

α_i if (Mahalanobis Distance of unknown sample to M_i) > (Mahalanobis Distance of unknown sample to M_j)

If $P(\omega_i) \neq P(\omega_j)$

The boundary moves toward the less probable one.

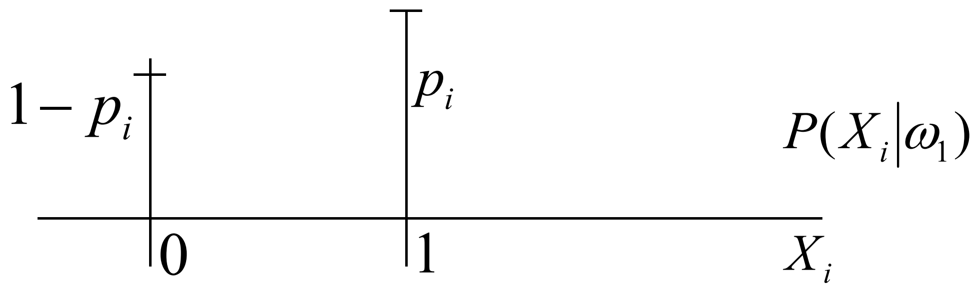
Binary Random Variables

- <u>Discrete features:</u> Features can take only discrete values. Integrals are replaced by summations.

- <u>Binary Features:</u> 0 or 1

$$p_i = (X_i = 1|\omega_1)$$

$$q_i = (X_i = 1|\omega_2)$$

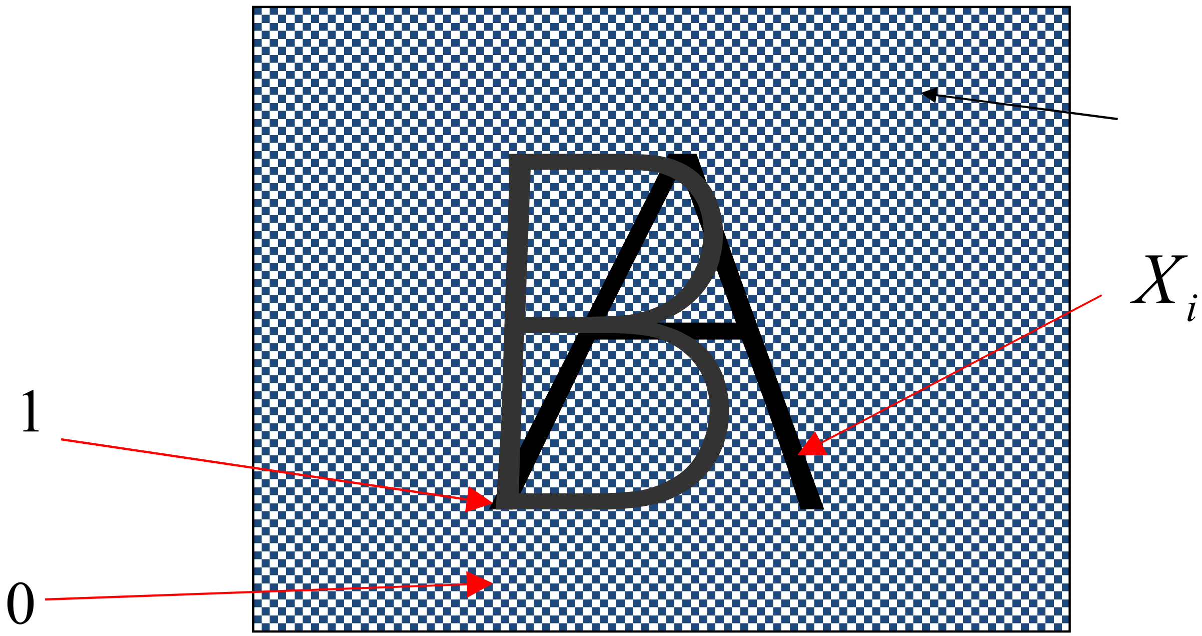


- Assume binary features are statistically independent.
- Where X_i is binary

$$X = [X_1, X_2, \ldots, X_d]^T$$

Binary Random Variables

<u>Example:</u> Bit – matrix for machine – printed characters



Here, each pixel may be taken as a feature X_i

For above problem, we have $d = 10 \times 10 = 100$

p_i is the probability that $X_i = 1$ for letter A,B,...

$$P(x_i) = (p_i)^{x_i}(1-p_i)^{1-x_i}$$

defined for $x_i = 0,1$ undefined elsewhere:

$$P(X) = \prod_{i=1}^{d} P(x_i) = \prod_{i=1}^{d} (p_i)^{x_i}(1-p_i)^{1-x_i}$$

$$g_k(X) = \log(P(X|w_k) + \log P(w_k)) = \sum_{i=1}^{d} x_i \log p_i + \sum (1-x_i)\log(1-p_i) + \log P(w_k)$$

- If statistical independence of features is assumed.

- Consider the 2 category problem; assume:

$$p_i = (x = 1|\omega_1)$$

$$q_i = (x = 1|\omega_2)$$

then, the decision boundary is:

$$\sum x_i \log p_i + \sum (1-x_i)\log(1-p_i) - \sum x_i \log q_i - \sum (1-x_i)\log(1-q_i) +$$
$$\log P(\omega_1) - \log P(\omega_2) = 0$$

So if

$$\sum x_i \log\tfrac{p_i}{q_i} + \sum (1-x_i)\log\tfrac{1-p_i}{1-q_i} + \log\tfrac{P(\omega_1)}{P(\omega_2)}$$

$$>0 \quad category\ 1$$
$$else \quad 2$$

$$g(X) = \sum W_i X_i + W_0$$

The decision boundary is linear in X.

a weighted sum of the inputs

where:

and

$$W_i = \ln \frac{P_i(1-q_i)}{q_i(1-p_i)} \qquad W_0 = \sum \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$