Bayesian Decision Theory

Computer Science- Pattern Recognition

Prof. Dr. Dhahir A. Abdullah

Case 2: $\Sigma_i = \Sigma$

- Covariance matrices for all of the classes are identical,
- But covariance matrices are arbitrary.

Case 2: $\Sigma_i = \Sigma$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1} (x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$\frac{d}{2}\ln 2\pi$ and $\frac{1}{2}\ln|\Sigma_i|$ are independent of i.

So,

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1} (x-\mu_i) + \ln P(\omega_i)$$

Case 2: $\Sigma_i = \Sigma$

Discriminant functions are

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad \text{(linear discriminant)}$$

where

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

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \ln P(w_i).$$

Case 2: $\Sigma_i = \Sigma$

Decision boundaries can be written as

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x_0}) = 0$$

where

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

$$\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln(P(w_i)/P(w_j))}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

Hyperplane passes through $\mathbf{x}_0$ but is not necessarily orthogonal to the line between the means.

Case 2: $\Sigma_i = \Sigma$

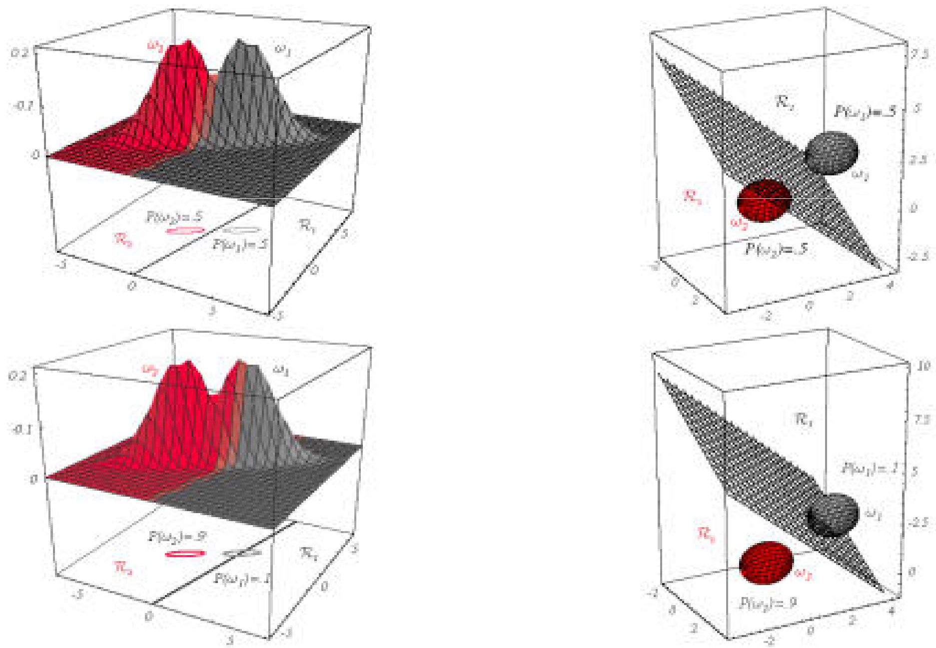


Figure 6: Probability densities with equal but asymmetric Gaussian distributions. The decision hyperplanes are not necessarily perpendicular to the line connecting the means.

Case 3: Σ_i = arbitrary

- Covariance matrices are different for each category.

Case 3: Σ_i = arbitrary

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1} (x-\mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln|\Sigma_i| + \ln P(\omega_i)$$

only $\frac{d}{2} \ln 2\pi$ is independent of i.

So,

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

Case 3: Σ_i = arbitrary

Discriminant functions are

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad \text{(quadratic discriminant)}$$

where

$$\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1}$$

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

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(w_i).$$

Decision boundaries are hyperquadrics.

Case 3: Σ_i = arbitrary

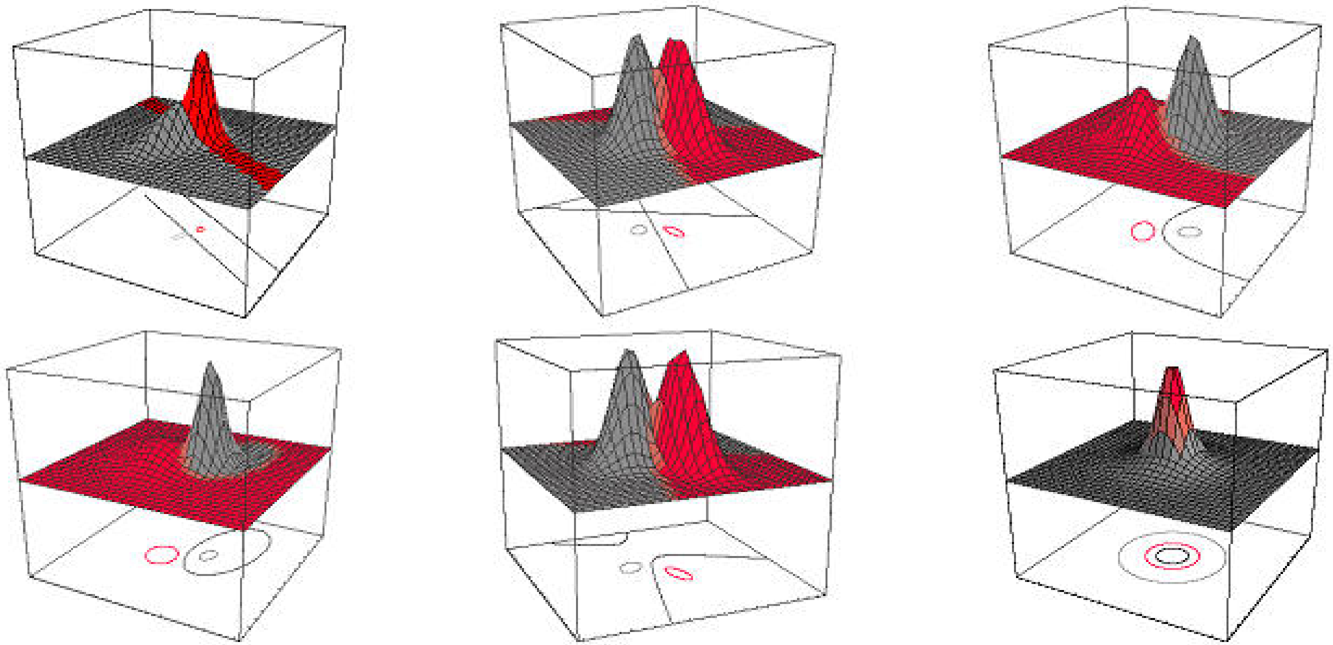


Figure 7: Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics.

Case 3: Σ_i = arbitrary

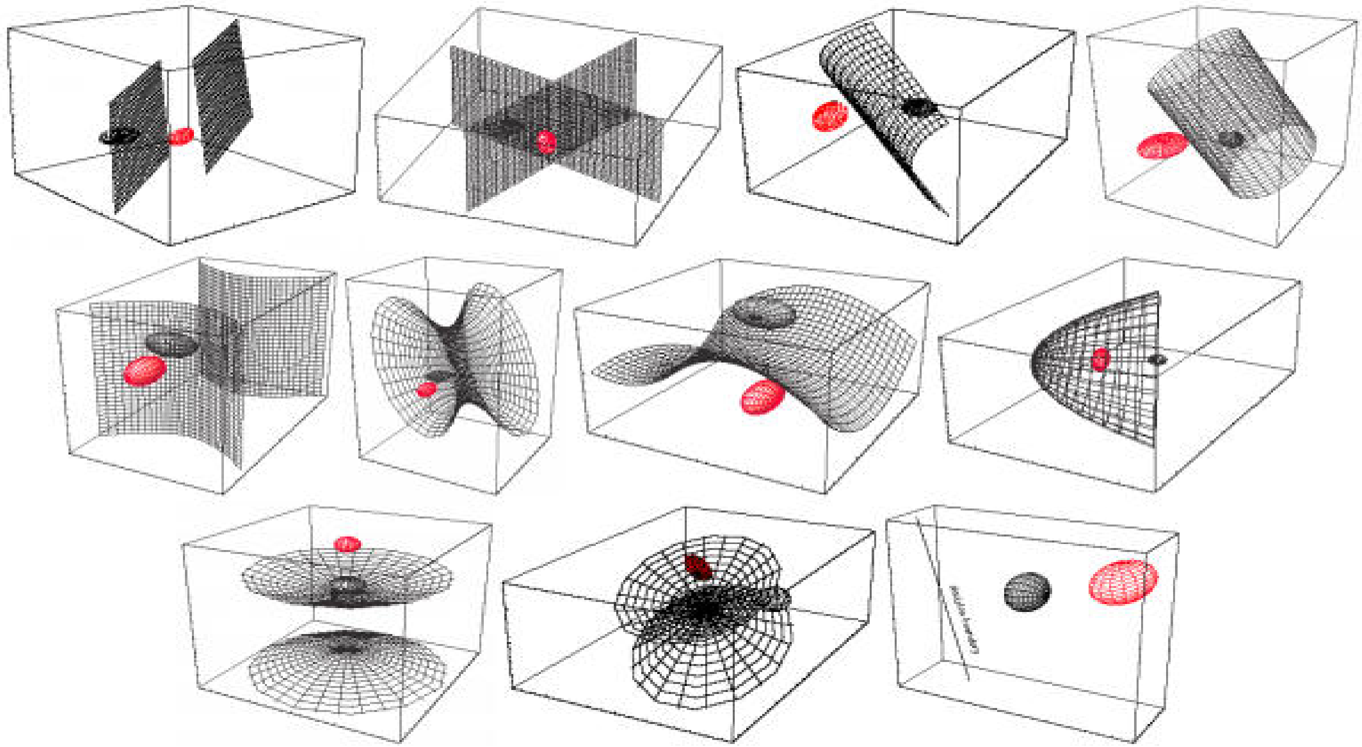


Figure 8: Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics.

EXAMPLE:

Consider a two-category classification problem

with two-dimensional feature vector $X = (x_1, x_2)^t$.

The two categories are ω_1 and ω_2.

$$p(X|\omega_1) \sim N\left(\begin{bmatrix}-1\\1\end{bmatrix}, \Sigma_1\right),$$

$$p(X|\omega_2) \sim N\left(\begin{bmatrix}1\\0\end{bmatrix}, \Sigma_2\right), \qquad \Sigma_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \Sigma_2 = \begin{bmatrix}2 & 2\\2 & 3\end{bmatrix}.$$

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

EXAMPLE (cont.):

Calculate the Bayes decision boundary.

EXAMPLE (cont.):

Find the **discriminant function for the first class.**

$$g_1(x) = -\ln(2\pi) - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\left(\bar{x} - \overline{\mu_1}\right)^t \Sigma_1^{-1}\left(\bar{x} - \overline{\mu_1}\right)$$

$$= -\ln(2\pi) - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\begin{bmatrix} x_1 - \mu_{11} & x_2 - \mu_{12} \end{bmatrix}\Sigma_1^{-1}\begin{bmatrix} x_1 - \mu_{11} \\ x_2 - \mu_{12} \end{bmatrix}$$

EXAMPLE (cont.):

Find the **discriminant function for the first class**.

$$-\ln(2\pi)-\frac{1}{2}\ln|\Sigma_1|-\frac{1}{2}[x_1-\mu_{11} \quad x_2-\mu_{12}]\Sigma_1^{-1}\begin{bmatrix}x_1-\mu_{11}\\x_2-\mu_{12}\end{bmatrix}=$$

$$=-\ln(2\pi)-\frac{1}{2}\ln 1-\frac{1}{2}[x_1+1 \quad x_2-1]\begin{bmatrix}1&0\\0&1\end{bmatrix}\begin{bmatrix}x_1+1\\x_2-1\end{bmatrix}$$

$$=-\ln(2\pi)-\frac{1}{2}\left((x_1+1)^2+(x_2-1)^2\right)$$

$$=-\ln(2\pi)-\frac{1}{2}\left(x_1^2+2x_1+1+x_2^2-2x_2+1\right)$$

EXAMPLE (cont.):

Similarly, find the **discriminant function for the second class**.

$$g_2(x) = -\ln(2\pi) - \frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}\left(\bar{x} - \overline{\mu_2}\right)^t \Sigma_2^{-1}\left(\bar{x} - \overline{\mu_2}\right)$$

$$= -\ln(2\pi) - \frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}\begin{bmatrix} x_1 - \mu_{21} & x_2 - \mu_{22} \end{bmatrix}\Sigma_2^{-1}\begin{bmatrix} x_1 - \mu_{21} \\ x_2 - \mu_{22} \end{bmatrix}$$

$$= -\ln(2\pi) - \frac{1}{2}\ln 2 - \frac{1}{2}\begin{bmatrix} x_1 - 1 & x_2 \end{bmatrix}\frac{1}{2}\begin{bmatrix} 3 & -2 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} x_1 - 1 \\ x_2 \end{bmatrix}$$

$$= -\ln(2\pi) - \frac{1}{2}\ln 2 - \frac{1}{4}\begin{bmatrix} 3(x_1 - 1) - 2x_2 & -2(x_1 - 1) + 2x_2 \end{bmatrix}\begin{bmatrix} x_1 - 1 \\ x_2 \end{bmatrix}$$

$$= -\ln(2\pi) - \frac{1}{2}\ln 2 - \frac{1}{4}\left(3(x_1 - 1)^2 - 2(x_1 - 1)x_2 - 2(x_1 - 1)x_2 + 2x_2^2\right)$$

$$= -\ln(2\pi) - \frac{1}{2}\ln 2 - \frac{1}{4}\left(3x_1^2 - 6x_1 + 3 - 4x_1x_2 + 4x_2 + 2x_2^2\right)$$

EXAMPLE (cont.):

The **decision boundary:**

$$g(x) = g_1(x) - g_2(x)$$

$$= -\ln(2\pi) - \frac{1}{2}\left(x_1^2 + 2x_1 + 1 + x_2^2 - 2x_2 + 1\right) + \ln(2\pi) + \frac{1}{2}\ln 2 - \frac{1}{4}\left(3x_1^2 - 6x_1 + 3 - 4x_1x_2 + 4x_2 + 2x_2^2\right)$$

$$= -\frac{1}{2}\left(x_1^2 + 2x_1 + 1 + x_2^2 - 2x_2 + 1\right) + \frac{1}{2}\ln 2 + \frac{1}{4}\left(3x_1^2 - 6x_1 + 3 - 4x_1x_2 + 4x_2 + 2x_2^2\right)$$

$$= -\frac{1}{2}x_1^2 - x_1 - \frac{1}{2} - \frac{1}{2}x_2^2 + x_2 - \frac{1}{2} + \frac{1}{2}\ln 2 + \frac{3}{4}x_1^2 - \frac{6}{4}x_1 + \frac{3}{4} - x_1x_2 + x_2 + \frac{1}{2}x_2^2$$

$$= -\frac{1}{2}x_1^2 + \frac{3}{4}x_1^2 - x_1 - \frac{6}{4}x_1 - x_1x_2 + x_2 + x_2 + \frac{1}{2}x_2^2 - \frac{1}{2}x_2^2 - \frac{1}{2} - \frac{1}{2} + \frac{3}{4} + \frac{1}{2}\ln 2$$

$$= \frac{1}{4}x_1^2 - \frac{10}{4}x_1 - x_1x_2 + 2x_2 - \frac{1}{4} + \frac{1}{2}\ln 2$$

EXAMPLE (cont.):

The **decision boundary:**

$$g(x) = g_1(x) - g_2(x) \quad = \frac{1}{4}x_1^2 - \frac{10}{4}x_1 - x_1x_2 + 2x_2 - \frac{1}{4} + \frac{1}{2}\ln 2$$

$$= x_1^2 - 10x_1 - 4x_1x_2 + 8x_2 - 1 + 2\ln 2$$

EXAMPLE (cont.):

Using MATLAB we can draw the decision boundary:

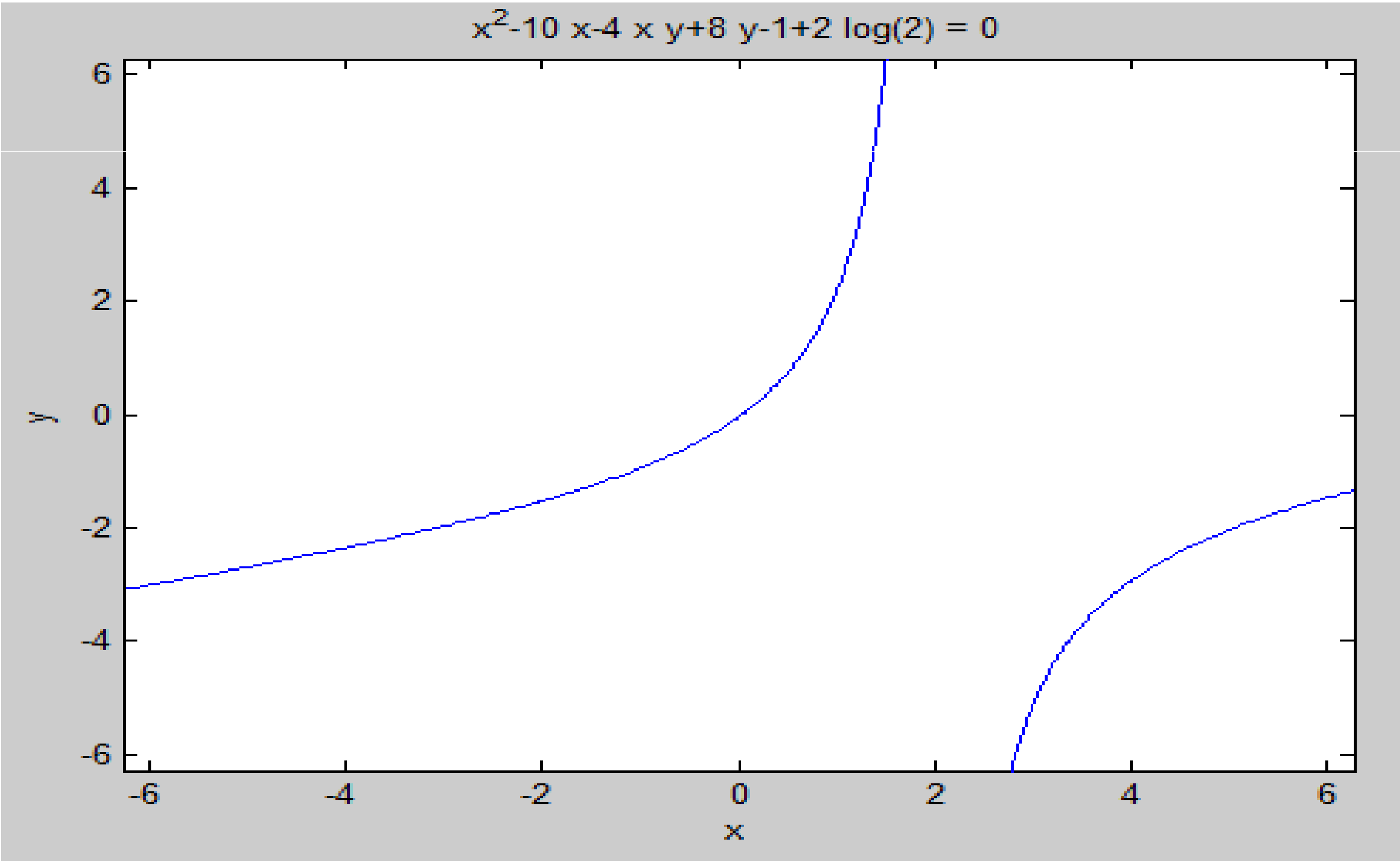
$$g(x) = x_1^2 - 10x_1 - 4x_1x_2 + 8x_2 - 1 + 2\ln 2$$

(to draw the decision boundary in MATLAB)

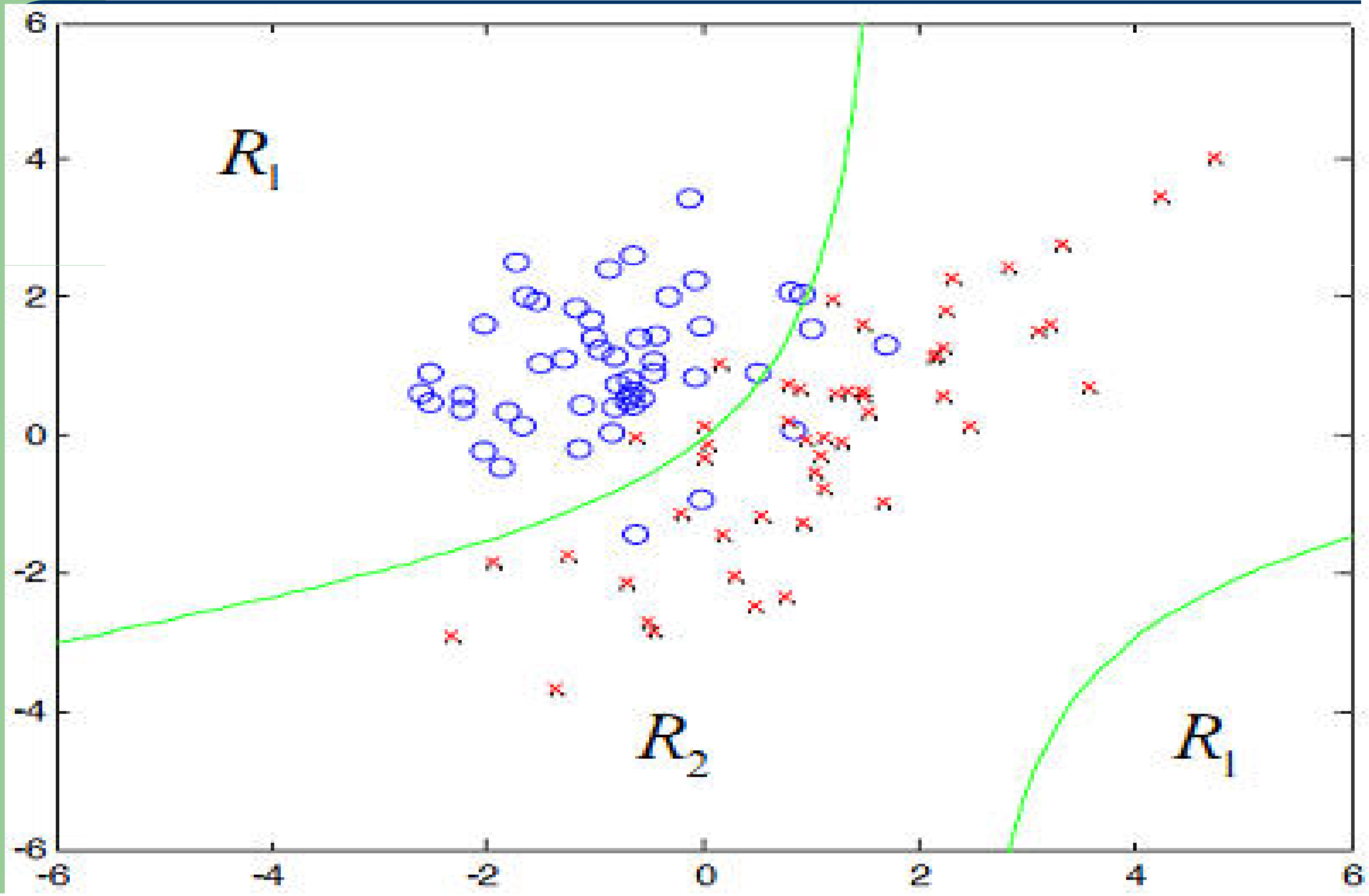
```
>> s = 'x^2-10*x-4*x*y+8*y-1+2*log(2)';
>> ezplot(s)
```

EXAMPLE (cont.):

Using MATLAB we can draw the decision boundary:



EXAMPLE (cont.):



Error Probabilities and Integrals

For the two-category case

$$\begin{aligned} P(error) &= P(\mathbf{x} \in \mathcal{R}_2, w_1) + P(\mathbf{x} \in \mathcal{R}_1, w_2) \\ &= P(\mathbf{x} \in \mathcal{R}_2 | w_1) P(w_1) + P(\mathbf{x} \in \mathcal{R}_1 | w_2) P(w_2) \\ &= \int_{\mathcal{R}_2} p(\mathbf{x}|w_1)\, P(w_1)\, d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|w_2)\, P(w_2)\, d\mathbf{x}. \end{aligned}$$

Error Probabilities and Integrals

For the multicategory case

$$
\begin{aligned}
P(error) &= 1 - P(correct) \\
&= 1 - \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i, w_i) \\
&= 1 - \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i | w_i) P(w_i) \\
&= 1 - \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathbf{x}|w_i)\, P(w_i)\, d\mathbf{x}.
\end{aligned}
$$

Error Probabilities and Integrals

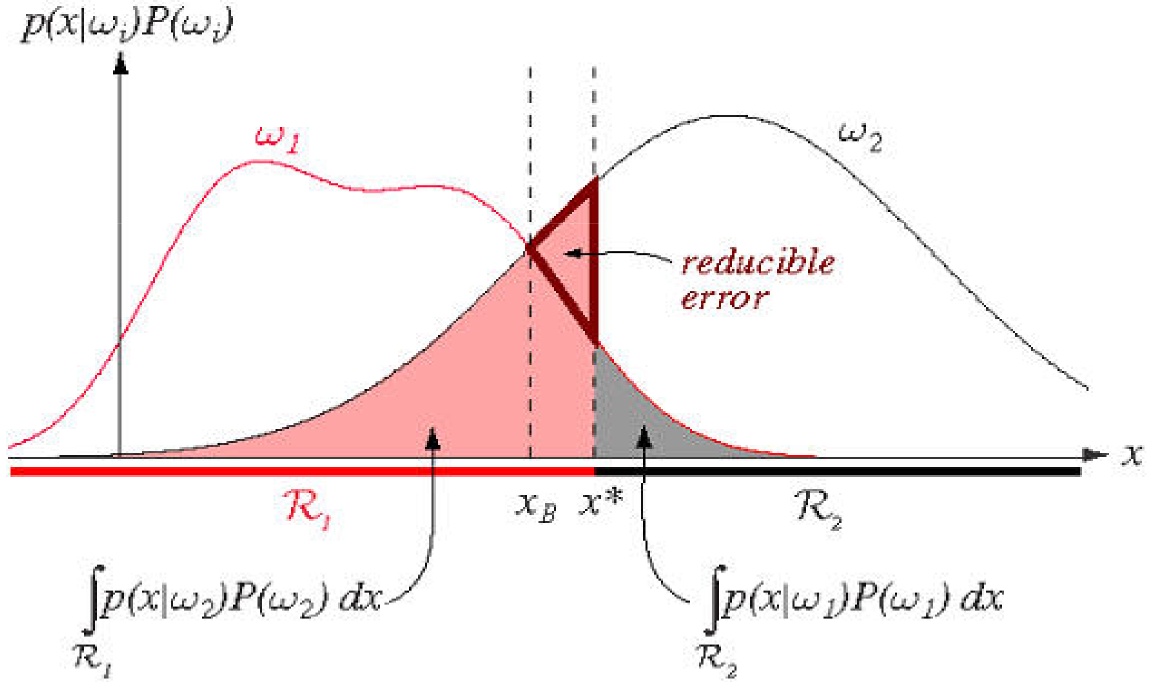


Figure 9: Components of the probability of error for equal priors and the non-optimal decision point x^*. The optimal point x_B minimizes the total shaded area and gives the Bayes error rate.

Receiver Operating Characteristics

- Another measure of distance between two Gaussian distributions.
- found a great use in medicine, radar detection and other fields.

Receiver Operating Characteristics

- Consider the two-category case and define
  - w_1: target is present,
  - w_2: target is not present.

Table 1: *Confusion matrix*.

		Assigned	
		w_1	w_2
True	w_1	correct detection	mis-detection
	w_2	false alarm	correct rejection

- Mis-detection is also called false negative or Type I error.
- False alarm is also called false positive or Type II error.

Receiver Operating Characteristics

- If we use a parameter (e.g., a threshold) in our decision, the plot of these rates for different values of the parameter is called the *receiver operating characteristic* (ROC) curve.

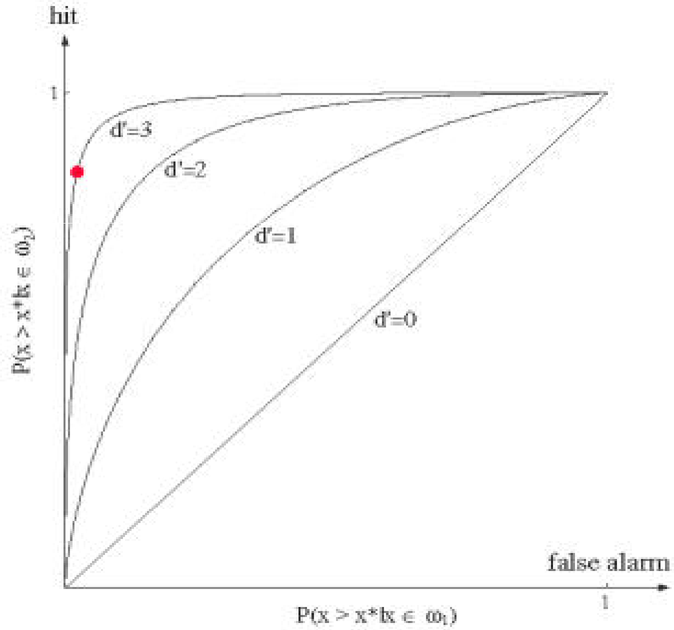


Figure 10: Example receiver operating characteristic (ROC) curves for different settings of the system.

Receiver Operating Characteristics

Let p_i be the probability that patient i will get a positive diagnosis (i.e., the patient is ill) and q_i be patient i's probability of a positive test. The *prevalence*, P, of the positive diagnosis in the population[2] is theoretically $P = \text{mean}(p_i)$. The *level* of the test, Q, is $Q = \text{mean}(q_i)$. We also define $P' = 1 - P$ and $Q' = 1 - Q$.

	Test result		
	Positive	Negative	
Diagnosis Positive	TP	FN	P
Negative	FP	TN	P'
	Q	Q'	1

Receiver Operating Characteristics

- If both diagnosis and test are positive, it is called a true positive. The probability of a TP to occur is estimated by counting the true positives in the sample and divide by the sample size.

- If the diagnosis is positive and the test is negative it is called a false negative (FN).

- False positive (FP) and true negative (TN) are defined similarly.

Receiver Operating Characteristics

- The values described are used to calculate different measurements of the quality of the test.
- The first one is sensitivity, SE, which is **the probability of having a positive test among the patients who have a positive diagnosis.**

$$\mathrm{SE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) = \mathrm{TP}/P.$$

Receiver Operating Characteristics

- Specificity, SP, is **the probability of having a negative test among the patients who have a negative diagnosis.**

$$\mathrm{SP} = \mathrm{TN}/(\mathrm{FP} + \mathrm{TN}) = \mathrm{TN}/P'.$$

Receiver Operating Characteristics

- Example:

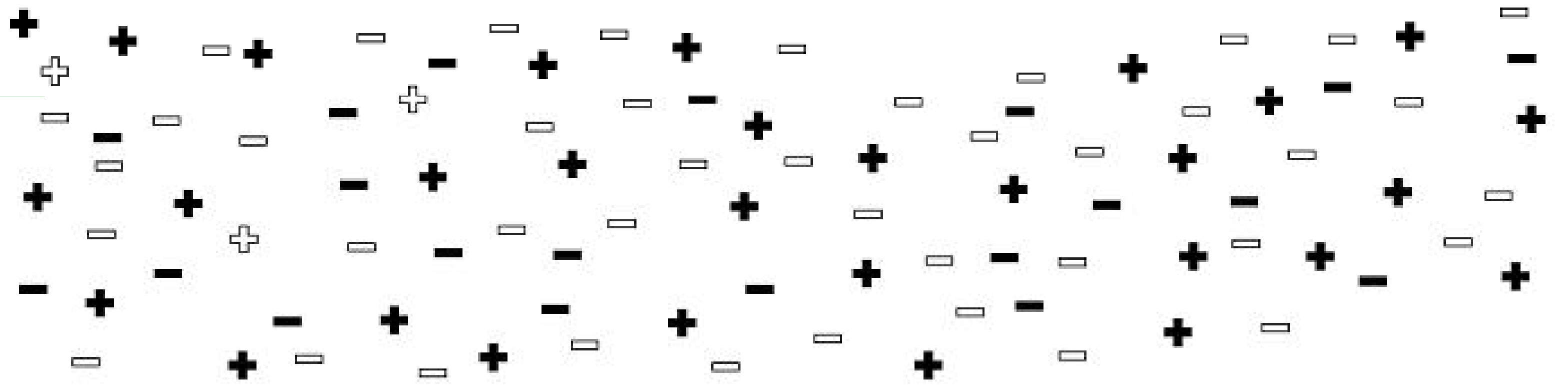


Figure 1.1: A sample population of $N = 95$ patients. The minus signs denote patients with a negative diagnosis and the plus signs denotes patients with a positive diagnosis. The colour shows the result of a test. White is a negative test and black is a positive test.

Receiver Operating Characteristics

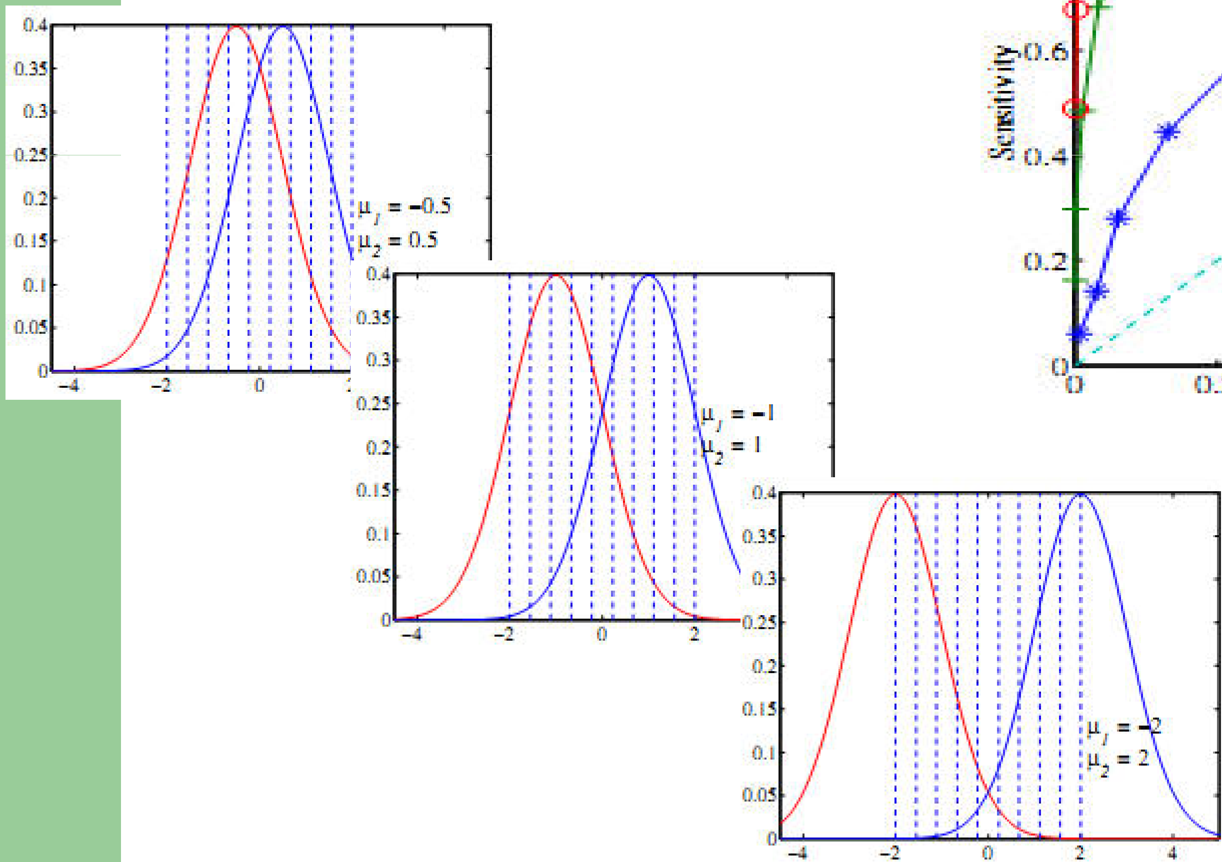
- Example (cont.):

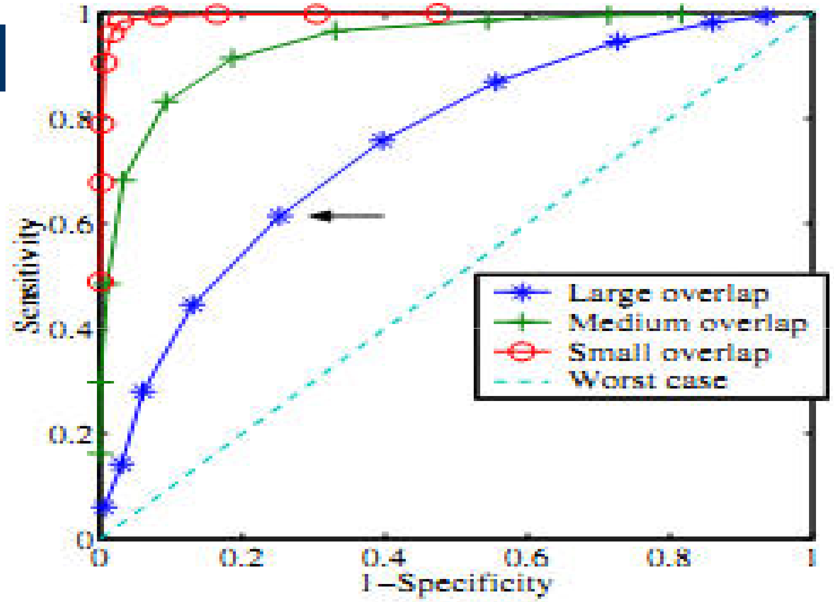
#Diagnosis	#Test result Positive	Negative	
Positive	30	3	33
Negative	20	42	62
	50	45	95

$$
\begin{array}{lllllllll}
\mathrm{TP} & = & 30/95 & = & 0.316(\pm 0.048) & P & = & 33/95 & = & 0.347(\pm 0.049) \\
\mathrm{FN} & = & 3/95 & = & 0.032(\pm 0.018) & Q & = & 50/95 & = & 0.526(\pm 0.051) \\
\mathrm{FP} & = & 20/95 & = & 0.211(\pm 0.042) & \mathrm{SE} & = & 30/33 & = & 0.909(\pm 0.050) \\
\mathrm{TN} & = & 42/95 & = & 0.442(\pm 0.051) & \mathrm{SP} & = & 42/62 & = & 0.677(\pm 0.059)
\end{array}
$$

Receiver Operating Characteristics

- Overlap in distributions:





Bayes Decision Theory – Discrete Features

- Assume **x = (x1.......xd)** takes only **m** discrete values
- Components of x are binary or integer valued, x can take only one of m **discrete values**

$$v_1, v_2, \ldots, v_m$$

Bayes Decision Theory – Discrete Features (Decision Strategy)

- *Maximize class posterior using bayes decision rule*

$$P(\omega_1 / \underline{x}) = \frac{P(\mathbf{x} / \omega_i) \cdot P(\omega_i)}{\Sigma P(\underline{\mathbf{x}} / \omega_j) \cdot P(\omega_j)}$$

Decide ω_i if $P(\omega_i / \underline{x}) > P(\omega_i / \underline{x})$ *for all* $i \neq j$

or

- Minimize the overall risk by selecting the action $\alpha_i = \alpha_* :$ *deciding* ω_2

$$\alpha_* = \arg\min R(\alpha_i / x)$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem)

Here, we consider a 2-category problem in which the components of the feature vector are binary-valued and *conditionally independent* (which yields a simplified decision rule):

$$x = (x_1, \ldots, x_d)^t, x_i \epsilon \{0, 1\}$$

We also assign the following probabilities (*p* and *q*) to each x_i in **X**:

$$p_i = \Pr[x_i = 1 | \omega_1] \text{ and } q_i = \Pr[x_i = 1 | \omega_2]$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem)

If $p_i > q_i$, we expect to x_i to be 1 more frequently when the state of nature is w_1 than when it is w_2. If we assume *conditional independence*, we can write P(**X**|w_i) as the product of probabilities for the components of **X**. The class conditional probabilities are then:

$$P(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$$

$$P(\mathbf{x}|\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem)

Since this is a two class problem the discriminant function $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$ where:

$$g_1(x) = \log\left[p(\mathbf{x}|\omega_1)p(\omega_1)\right] \text{ and } g_2(x) = \log\left[p(\mathbf{x}|\omega_2)p(\omega_2)\right]$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem)

The likelihood ratio is therefore given by:

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

which yields the discriminant function as follows:

$$g(x) = \ln \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1-x_i) \ln \frac{1-p_i}{1-q_i}\right] + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem)

If we notice that this function is linear in x_i, we can rewrite it as a linear function of x_i

$$\boxed{g(x) = \sum_{i=1}^{d} w_i x_i + w_0}$$

where

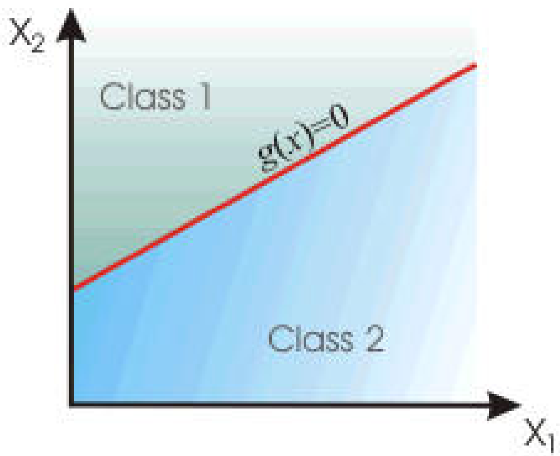
$$\boxed{w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)}, i = 1, \ldots, d}$$

and

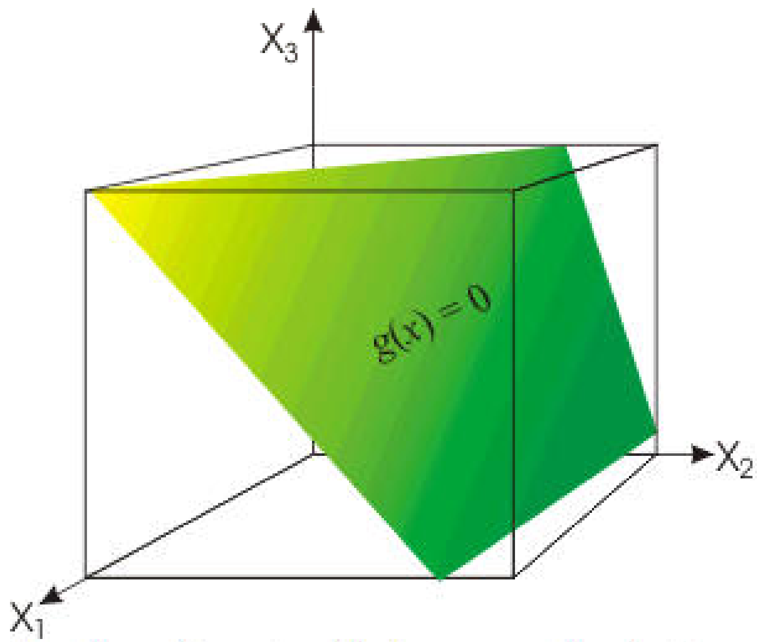
$$\boxed{w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}}$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem)

The discriminant function g(x) will therefore indicate whether the current feature vector belongs to class 1 and class 2. It is important to note that w_0 and w_i are weights calculated for the linear discriminant. A decision boundary lies wherever g(x) = 0. This decision boundary can be a line, or hyper-plane depending upon the dimension of the feature space.



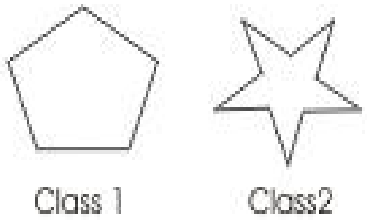
The decision boundry $g(x) = 0$ is a line on a cartesian plan for a two dimensional ($d = 2$) feature space.



In a three dimensional feature space, the decision boundary $g(x) = 0$ is a plane,

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem) EXAMPLE

Let's consider a three dimensional binary feature vector $X=(x_1,x_2,x_3) = (0,1,1)$ that we will attempt to classify with one of the following classes:



and lets say that the prior probability for class 1 is $P(\omega_1)= 0.6$ while for class 2 is $P(\omega_2)= 0.4$. Hence, it is already evident that there is a bias towards class 1.

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem) EXAMPLE

Additionally, we know that likelihoods of each independent feature is given by p and q where:

$$p_i = P(x_i=1|\omega_1) \text{ and } q_i = P(x_i=1|\omega_2)$$

meaning that we know the probability (or likelihood) of each independent feature given each class - these values are known and given:

$$p = \{0.8, 0.2, 0.5\} \text{ and } q = \{0.2, 0.5 , 0.9\}$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem) EXAMPLE

therefore, the discriminant function is $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$ or by taking the log of both sides:

$$g(x) = \log \frac{p(\mathbf{x}|\omega_2)}{p(\mathbf{x}|\omega_2)} + \log \frac{p(\omega_1)}{p(\omega_2)}$$

however, since the problem definition assumes that $\boldsymbol{X}$ is independent, the discriminant function can be calculated by:

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_o$$

with

$$w_i(x) = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1, \ldots, d$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem) EXAMPLE

$$w_i(x) = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \quad i = 1, \ldots, d$$

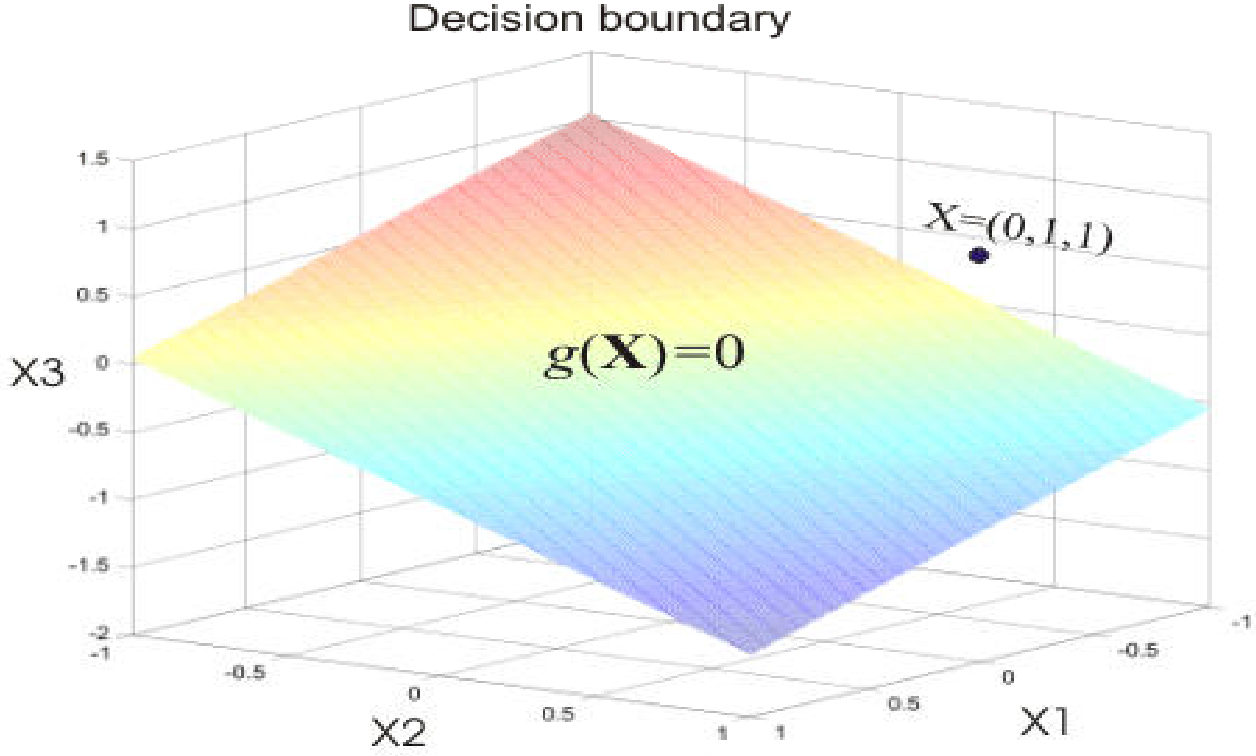
$$w_1 = \ln \frac{0.8(1-0.2)}{0.2(1-0.8)} = 2.77; w_2 = \ln \frac{0.2(1-0.5)}{0.5(1-0.2)} = -1.39; w_3 = \ln \frac{0.5(1-0.9)}{0.9(1-0.5)} = -2.19$$

$$w_0 = \ln\left(\frac{0.6}{0.4}\right) + \ln\left(\frac{1-0.8}{1-0.2}\right) + \ln\left(\frac{1-0.2}{1-0.5}\right) + \ln\left(\frac{1-0.5}{1-0.9}\right) = 1.0986$$

$$g(x) = 2.77x_1 - 1.39x_2 - 2.19x_3 + 1.0986$$

Bayes Decision Theory – Discrete Features (Independent binary features in 2-category problem) EXAMPLE

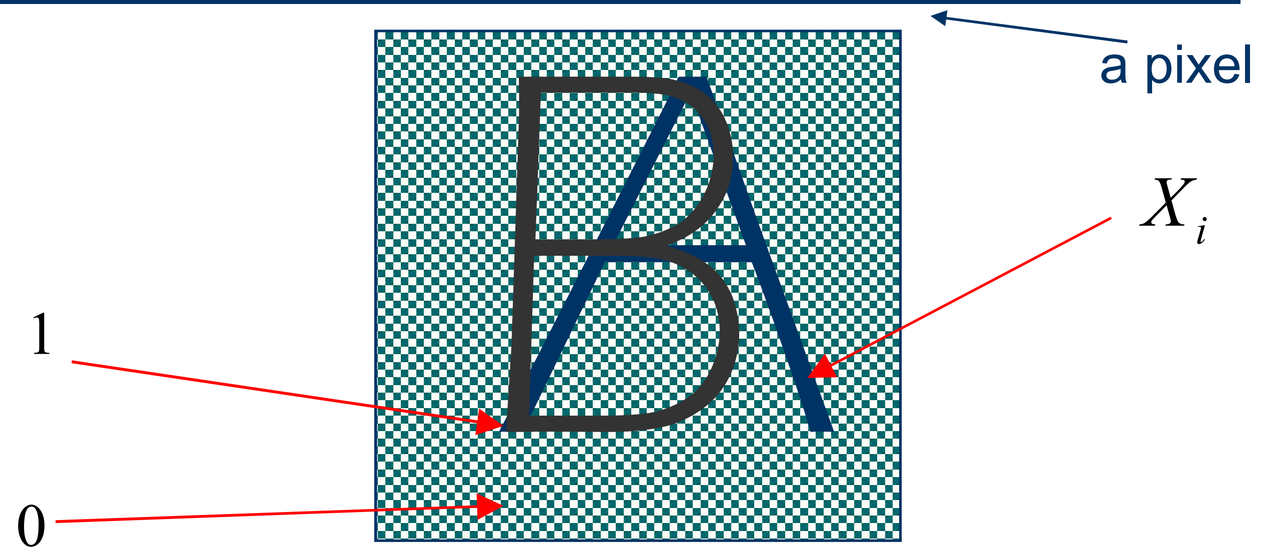
After inputting the x_i values into the discriminant function, the answer $g(\mathbf{x})$ = -2.4849. Therefore this belongs to class 2. Below is a plot of the decision boundary surface.



All points above the plane belong to class ω_2 since if X = (0,1,1), $g(\mathbf{x})$ = -2.4849 < 0.

APPLICATION EXAMPLE

Bit – matrix for machine – printed characters



Here, each pixel may be taken as a feature X_i $d = 10 \times 10 = 100$

For above problem, we have

p_i is the probabilty that $X_i = 1$ for letter A,B,…

Summary

•To minimize the overall risk, choose the action that minimizes the conditional risk **R (α |x).**
•To minimize the probability of error, choose the class that maximizes the posterior probability **P (wj |x).**
•If there are different penalties for misclassifying patterns from different classes, the posteriors must be weighted according to such penalties before taking action.
•Do not forget that these decisions are the optimal ones under the assumption that the “true” values of the probabilities are known.