Discriminant functions

Classification based on discriminant functions

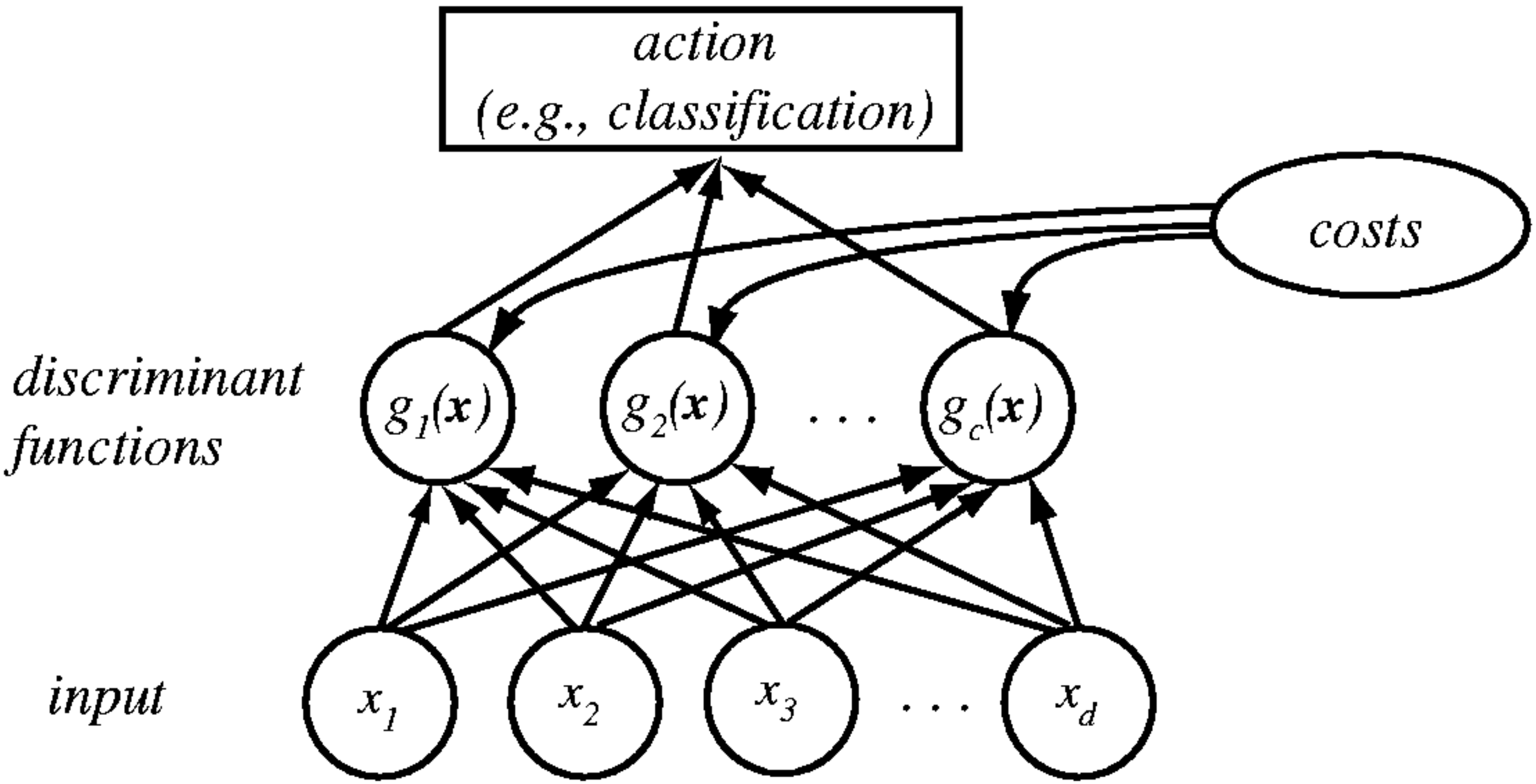
Compute values of discriminant functions $g_i(x)$ and assign to x the class corresponding to the discriminant function with the largest value.

The classifier assigns a feature vector x to class ω_i if

$$g_i(\mathrm{x}) > g_j(\mathrm{x}), \quad \text{for all } j \neq i$$

where $g_i(\mathrm{x})$ is the set of discriminant functions.

General structure of a classifier



(from Duda, Hart, Stork (2001) Pattern classification)

Examples of discriminant functions

Risks:

$$g_i(\mathrm{x}) = -R(\alpha_i \mid \mathrm{x})$$

Posterior probabilities:

$$g_i(\mathrm{x}) = P(\omega_i \mid \mathrm{x})$$

Freedom to change discriminant functions

The choice of discriminant functions is not unique.

In general: If every discriminant function $g_i(\mathrm{x})$ is replaced by a monotonically increasing function

$$f(g_i(\mathrm{x}))$$

the classification result does not change.

Examples of changing of discriminant functions

Motivation: Other quantities, simpler to understand or to compute, lead to identical classification results.

$$g_i(\mathrm{x}) = P(\omega_\mathrm{i} \mid \mathrm{x}) = \frac{p(\mathrm{x} \mid \omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathrm{x} \mid \omega_j)P(\omega_j)}$$

$$g'_i(\mathrm{x}) = p(\mathrm{x} \mid \omega_i)P(\omega_i)$$

$$f(g'_i(\mathrm{x})) = \ln p(\mathrm{x} \mid \omega_i) + \ln P(\omega_i)$$

(***ln*** is useful in the important case of normal distribution)

Dichotomizer

In a two-class problem, a ‘dichotomizer’ is defined

$$g(\mathrm{x}) = g_1(\mathrm{x}) - g_2(\mathrm{x})$$

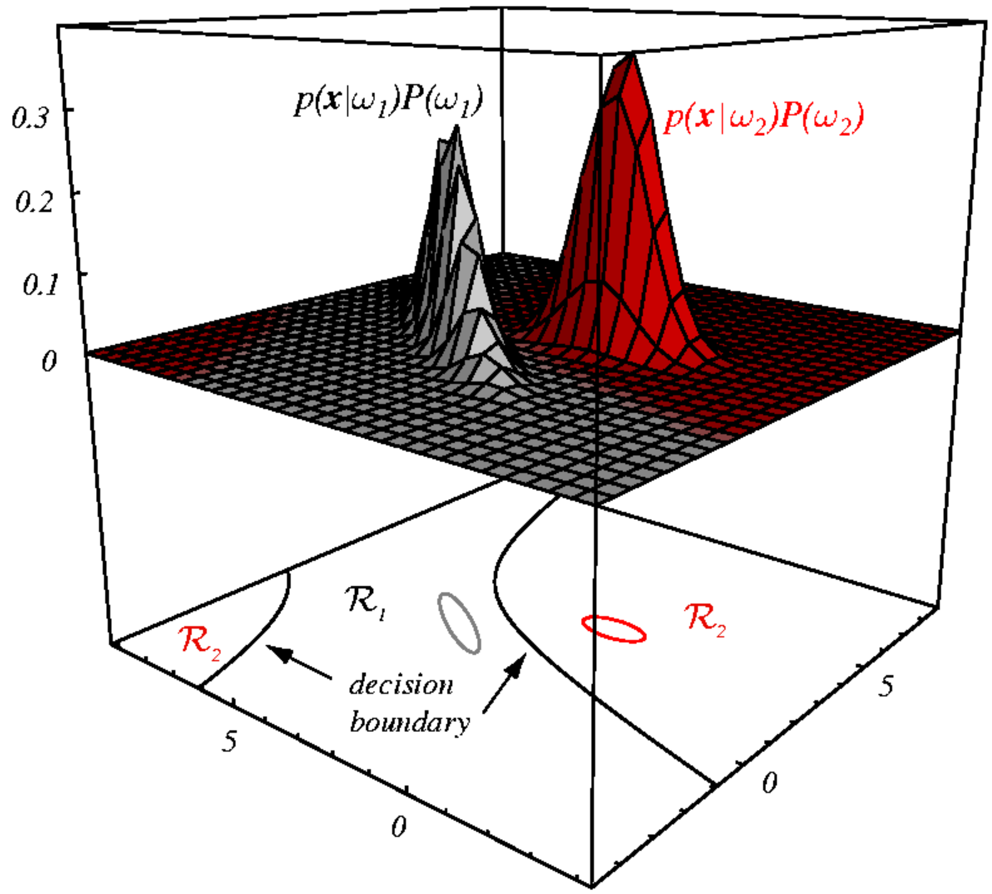
Decide ω_1 if $g(\mathrm{x}) > 0$, and ω_2 otherwise.

Alternative (sometimes more convenient) forms:

$$g(\mathrm{x}) = P(\omega_1 \mid \mathrm{x}) - P(\omega_2 \mid \mathrm{x})$$

$$g(\mathrm{x}) = \ln \frac{p(\mathrm{x} \mid \omega_1)}{p(\mathrm{x} \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Decision boundary for a dichotomizer



(from Duda, Hart, Stork (2001) Pattern classification)

Discriminant functions and decision boundaries for normal distributions

Univariate normal density

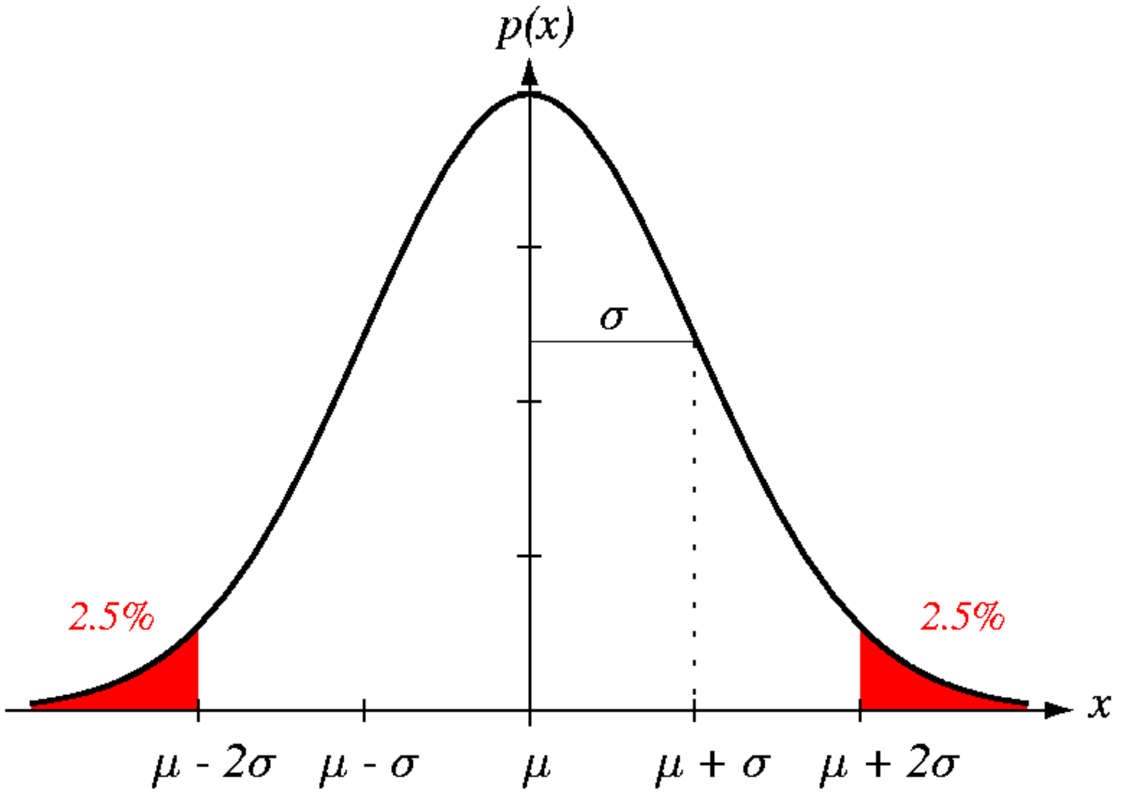
$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

Mean: $$\varepsilon[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx = \mu$$

Variance: $$\varepsilon[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, p(x)\, dx = \sigma^2$$

Univariate normal density

In 95% of the cases x is in the range $|x-\mu| \leq 2\sigma$



(from Duda, Hart, Stork (2001) Pattern classification)

Multivariate normal density

$$p(\mathrm{x}) = \frac{1}{(2\pi)^{d/2} \,|\sum|^{1/2}} \exp\left[-\frac{1}{2}(\mathrm{x}-\mu)^t \sum{}^{-1} (\mathrm{x}-\mu)\right]$$

- $\mathrm{x} \in \mathrm{R}^{\mathrm{d}}$ is a d dimensional vector
- μ is the mean (a d-dimensional vector itself)
- Σ is the covariance matrix ($|\Sigma|$ its determinant and Σ^{-1} its inverse)

Common notation:

$$p(\mathrm{x}) \sim N(\mu, \Sigma)$$

Covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{21} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \dots & \dots & & \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} {\sigma_1}^2 & \sigma_{21} & \dots & \sigma_{1d} \\ \sigma_{21} & {\sigma_2}^2 & \dots & \sigma_{2d} \\ \dots & \dots & & \\ \sigma_{d1} & \sigma_{d2} & \dots & {\sigma_d}^2 \end{bmatrix}$$

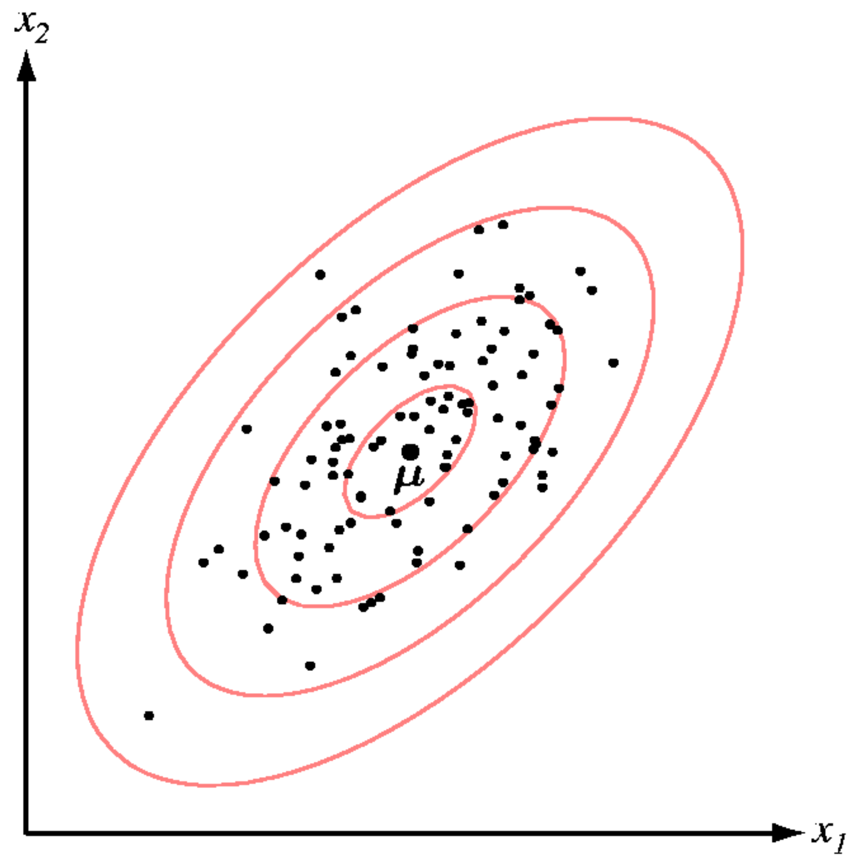
$$\varepsilon[(x_i - \mu_i)(x_j - \mu_j)] = \int (x_i - \mu_i)(x_j - \mu_j)\, p(x)\, dx = \sigma_{i,j}$$

Σ is always symmetric and positive semi-definite ($|\Sigma| \geq 0$)

For statistically independent events x_i and x_j , $\sigma_{ij} = 0$

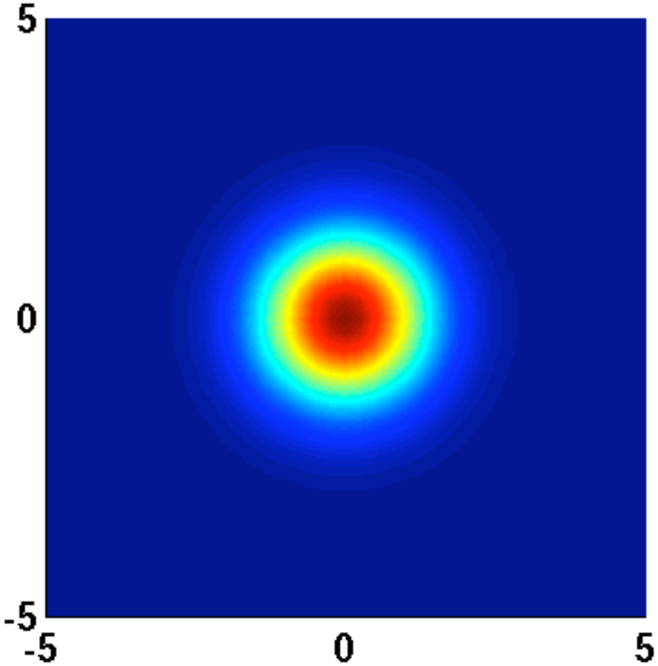
Thus it becomes a diagonal matrix.

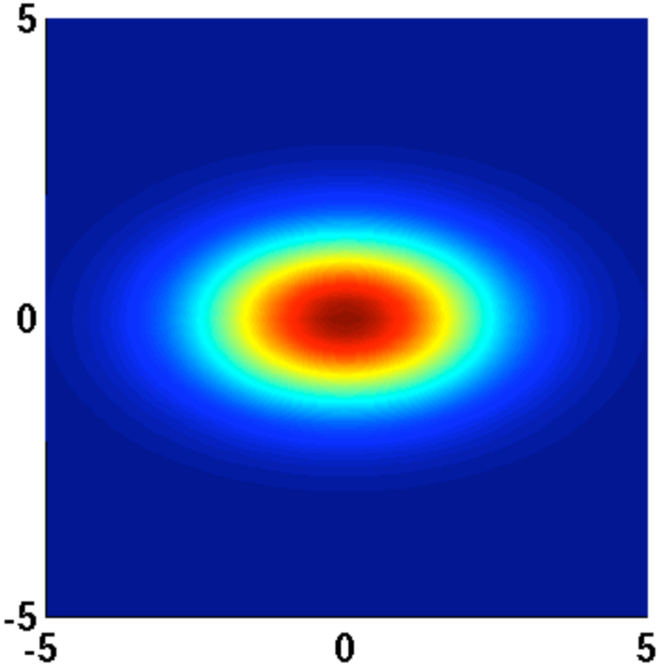
A hyper ellipsoidal cluster formed by points drawn from a population which has normal distribution



(from Duda, Hart, Stork (2001) Pattern classification)

Multivariate Gaussians



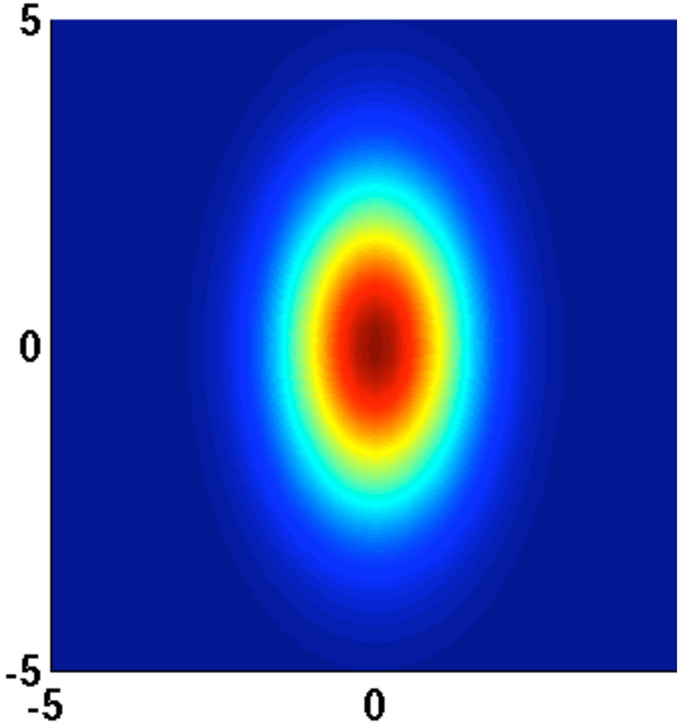


$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

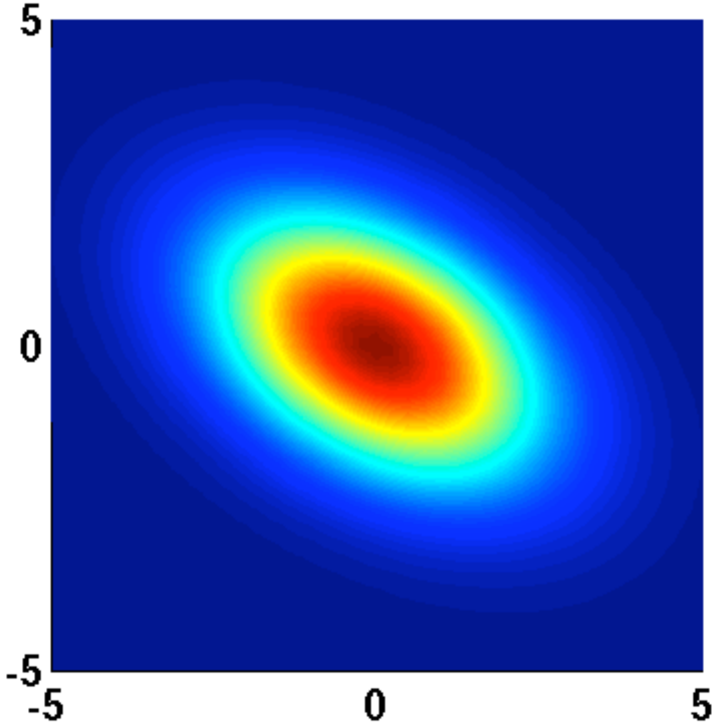
The principal axes of the hyperellipsoids are given by the eigenvectors of Σ .
The eigenvalues determine the length of these axes.

$$\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

Multivariate Gaussians



$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$



$$\Sigma = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}$$

Mahalanobis distance

The squared Mahalanobis distance from a point x to a class $N(\mu, \Sigma)$

$$r^2 = (\mathrm{x} - \mu)^t \Sigma^{-1} (\mathrm{x} - \mu)$$

The contours of constant density are hyperellipsoids of constant Mahalanobis distance.

Linear combinations

Linear combinations of normally distributed random variables (independent or not) are normally distributed, i.e.

$$\begin{matrix} p(\mathrm{x}) \sim N(\mu, \Sigma) \\ \mathrm{y} = \mathrm{A}^t \mathrm{x} \end{matrix} \quad \Rightarrow \quad p(\mathrm{y}) \sim N(\mathrm{A}^t \mu, \mathrm{A}^t \Sigma \mathrm{A})$$

where

- A is a d*k matrix
- y is a k dimensional vector

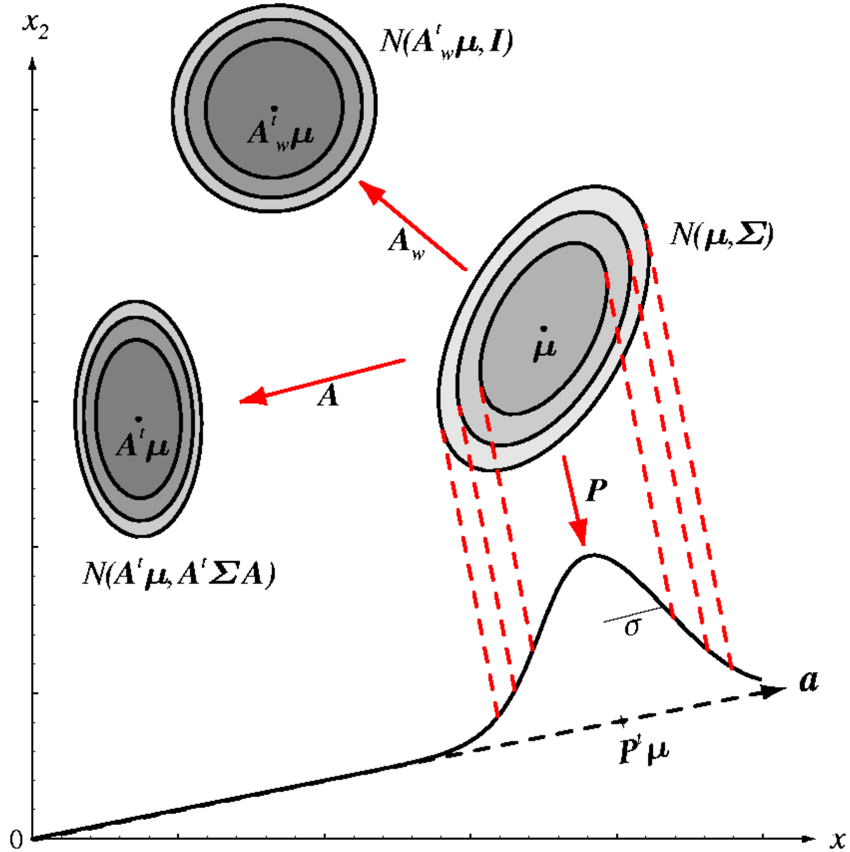
If k = 1, then $\mathrm{A} = [\mathrm{a}]$, and $y = \mathrm{a}^t \mathrm{x}$ is a scalar and represents the projection of **x** onto **a**.

Linear combinations

Whitening transform Aw

General transform A

Projection P



(from Duda, Hart, Stork (2001)Pattern classification)

Discriminant functions for normal densities

Minimum error rate classification can be achieved using the following discriminant functions:

$$g_i(\mathrm{x}) = \ln p(\mathrm{x} \mid \omega_i) + \ln P(\omega_i)$$

Bayesian Decision Theory

If the densities $p(\mathrm{x} \mid \omega_i)$ are multivariate normal densities $p(\mathrm{x} \mid \omega_\mathrm{i}) \sim N(\mu_i, \Sigma_i)$, then from:

$$p(\mathrm{x} \mid \omega_\mathrm{i}) = \frac{1}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \exp\left[-\frac{1}{2} (\mathrm{x} - \mu_i)^t \Sigma_i^{-1} (\mathrm{x} - \mu_i) \right]$$

it follows:

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

Statistically independent features of equal variance

$$\Sigma_i = \sigma^2 I$$

Features are statistically independent and each feature has the same variance σ^2.

The determinant of the covariance matrix and the inverse are $|\Sigma_i| = \sigma^{2d}$ and $\Sigma_i^{-1} = (1/\sigma^2)I$

The form of the discriminant functions simplifies to:

$$g_i(\mathrm{x}) = -\frac{\|\mathrm{x} - \mu_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

where $\|\mathrm{x} - \mu_i\|^2 = (\mathrm{x} - \mu_i)^t(\mathrm{x} - \mu_i)$ is the Euclidian norm.

Equivalently:

$$g_i(\mathrm{x}) = -\frac{\| \mathrm{x} - \mu_i \|^2}{2\sigma^2} + \ln P(\omega_i)$$

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

$$\Leftrightarrow \quad \boxed{g_i(\mathrm{x}) = -\frac{1}{2\sigma^2}[\mathrm{x}^{\mathrm{t}}\mathrm{x} - 2\mu_i^{\mathrm{t}}\mathrm{x} + \mu_i^t \mu_i] + \ln P(\omega_i)}$$

Linear discriminant

The discriminant function:

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

can be rewritten as a linear discriminant:

$$g_i(\mathrm{x}) = \mathrm{w}_{\mathrm{i}}^{\mathrm{t}}\mathrm{x} + w_{i0} \qquad \mathrm{w}_{\mathrm{i}} = \frac{1}{\sigma^2}\mu_i$$

$$w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln P(\omega_i) \qquad \cancel{\left(-\mathrm{x}^{\mathrm{t}}\mathrm{x}/(2\sigma^2)\right)}$$

The term w_{i0} is called the bias for the i-th category.

Linear machine

A classifier that uses linear discriminant functions is called a **linear machine**.

If the priors are equal, the optimum decision rule can be stated as:

- *Classify a feature vector in the class with the closest mean.*

This classifier is known as a minimum-distance classifier.

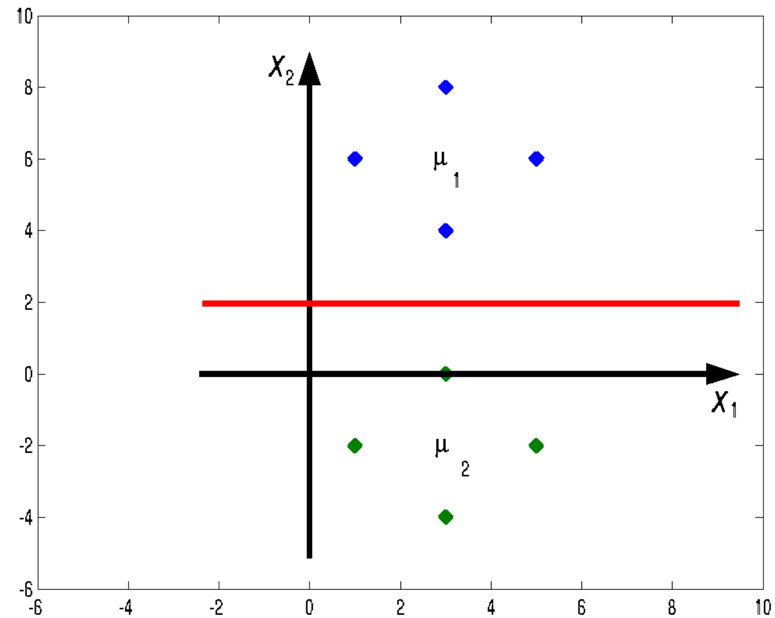
Example

$$g_i(\mathrm{x}) = -\frac{\| \mathrm{x} - \mu_i \|^2}{2\sigma^2} + \ln P(\omega_i)$$

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

$$\sigma_1 = \sigma_2 = \sqrt{2}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

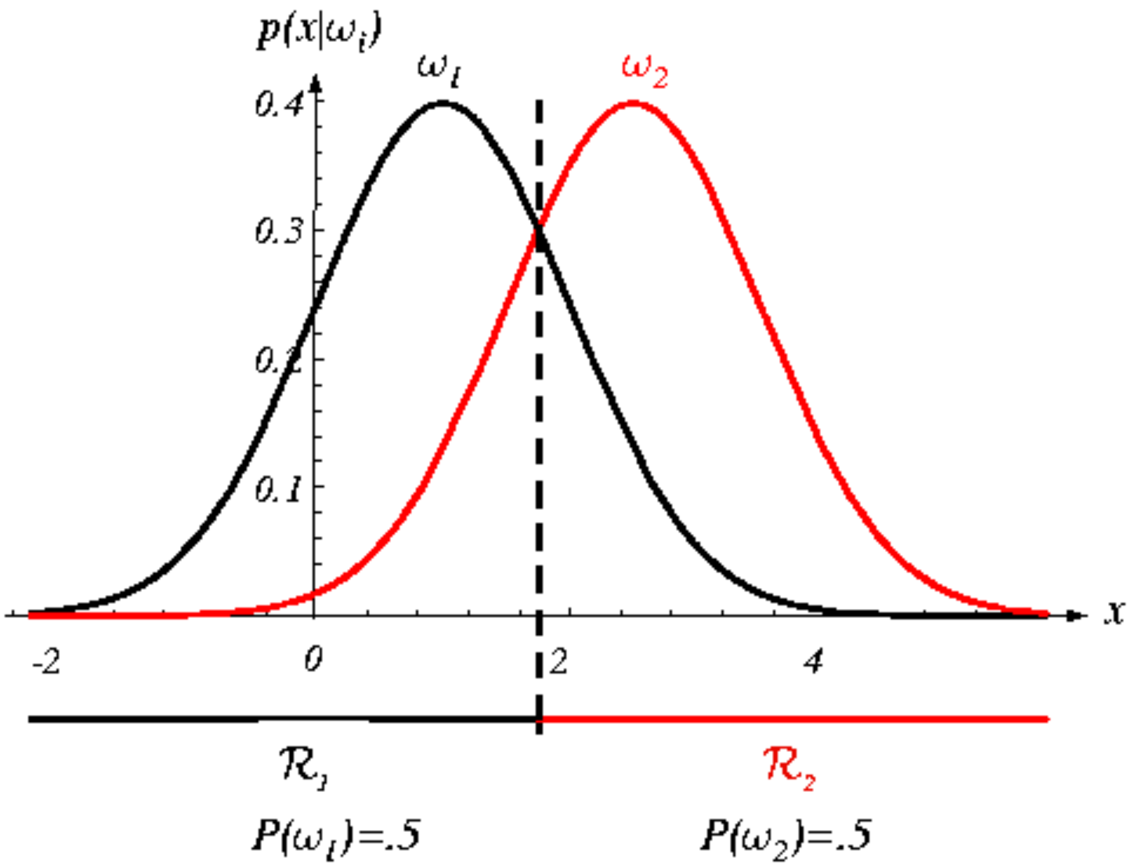


$$g_1(\mathrm{x}) = g_2(\mathrm{x}) \quad \Rightarrow$$

$$(\mathrm{x} - \mu_1)^t (\mathrm{x} - \mu_1) = (\mathrm{x} - \mu_2)^t (\mathrm{x} - \mu_2) \quad \Rightarrow$$

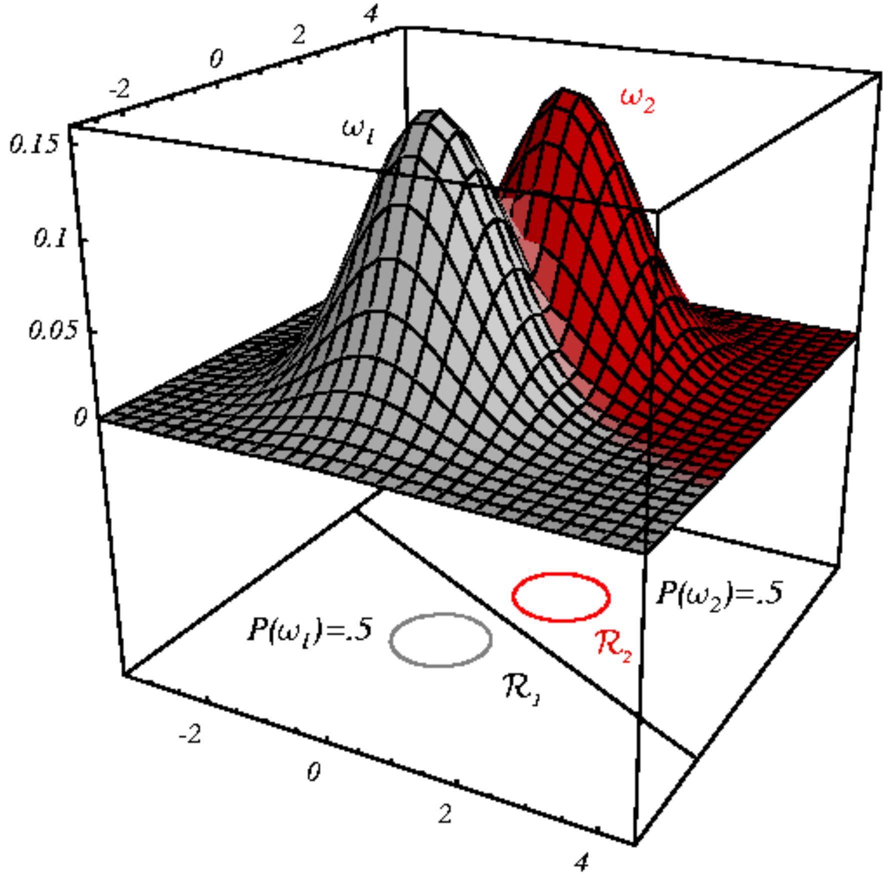
$$\begin{bmatrix} x_1 - 3 & x_2 - 6 \end{bmatrix} \begin{bmatrix} x_1 - 3 \\ x_2 - 6 \end{bmatrix} = \begin{bmatrix} x_1 - 3 & x_2 + 2 \end{bmatrix} \begin{bmatrix} x_1 - 3 \\ x_2 + 2 \end{bmatrix} \quad \Rightarrow \quad x_2 = 2$$

One-dimensional case



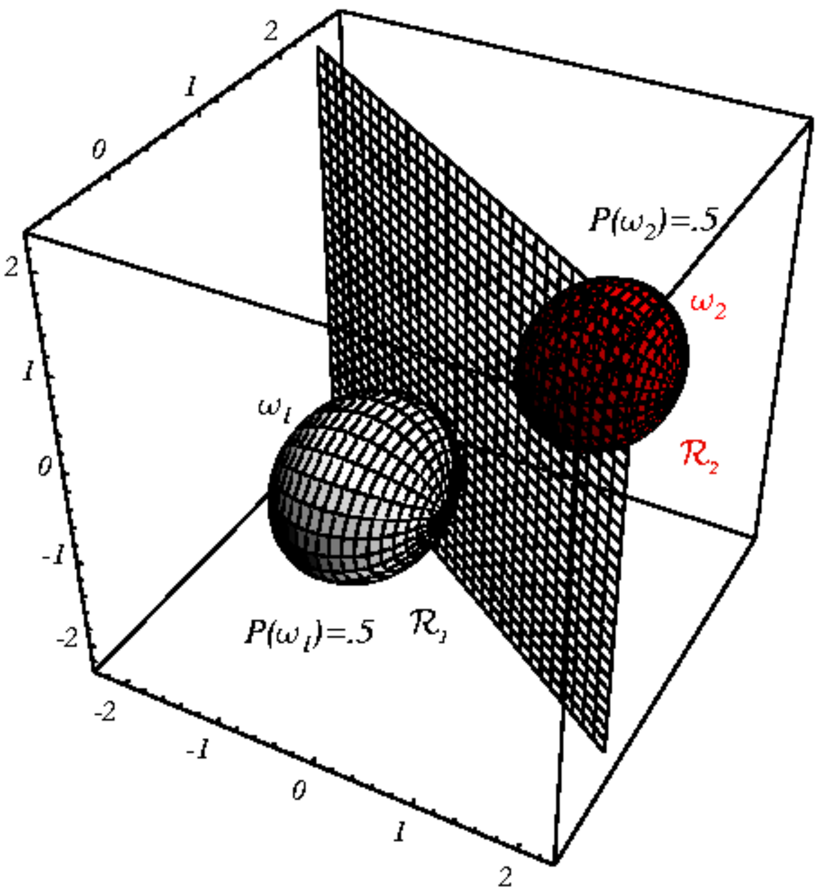
(from Duda, Hart, Stork (2001) Pattern classification)

Two-dimensional case



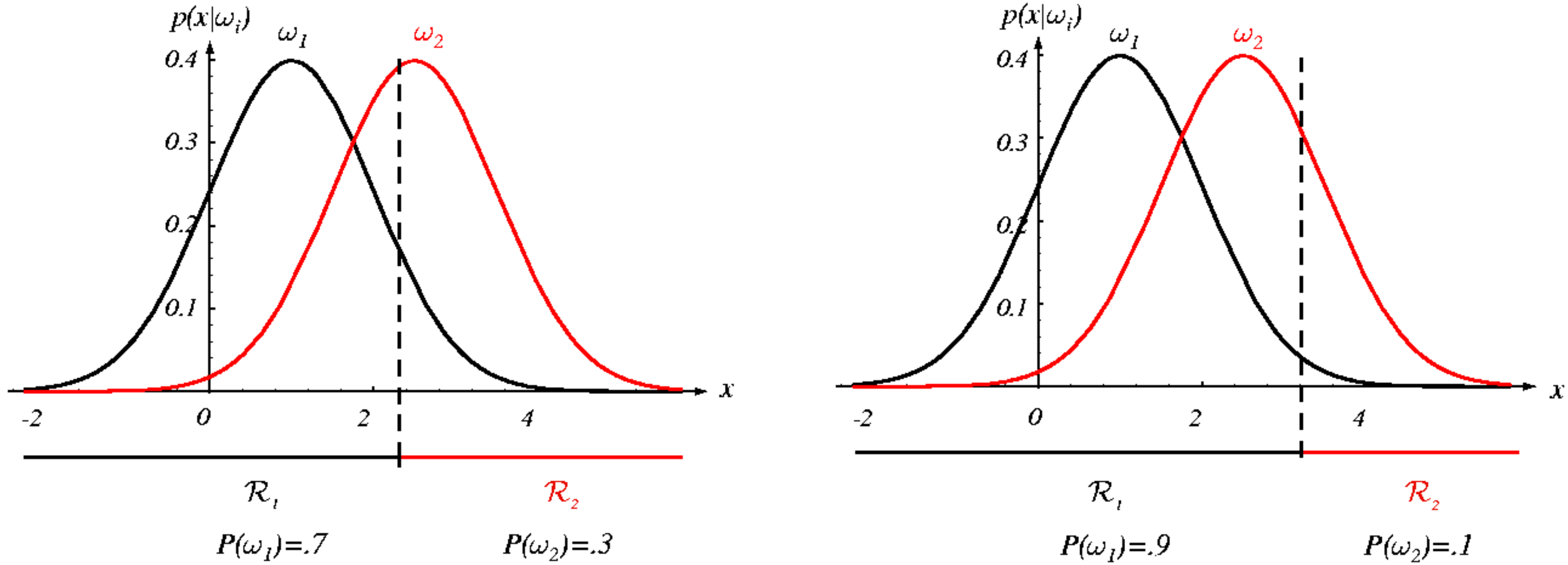
(from Duda, Hart, Stork (2001) Pattern classification)

Three-dimensional case



(from Duda, Hart, Stork (2001) Pattern classification)

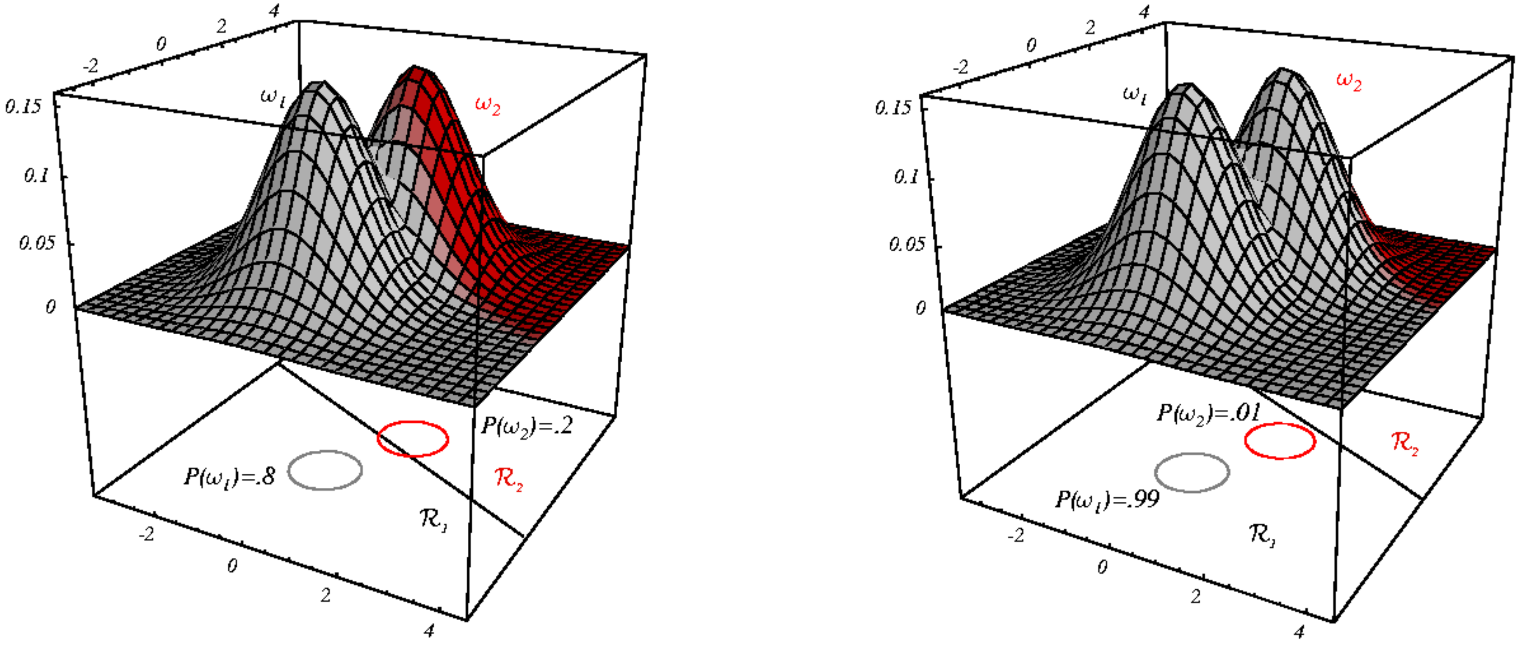
Unequal priors



When $P(\omega_i) \neq P(\omega_j)$, the decision boundary is shifted

(from Duda, Hart, Stork (2001) Pattern classification)

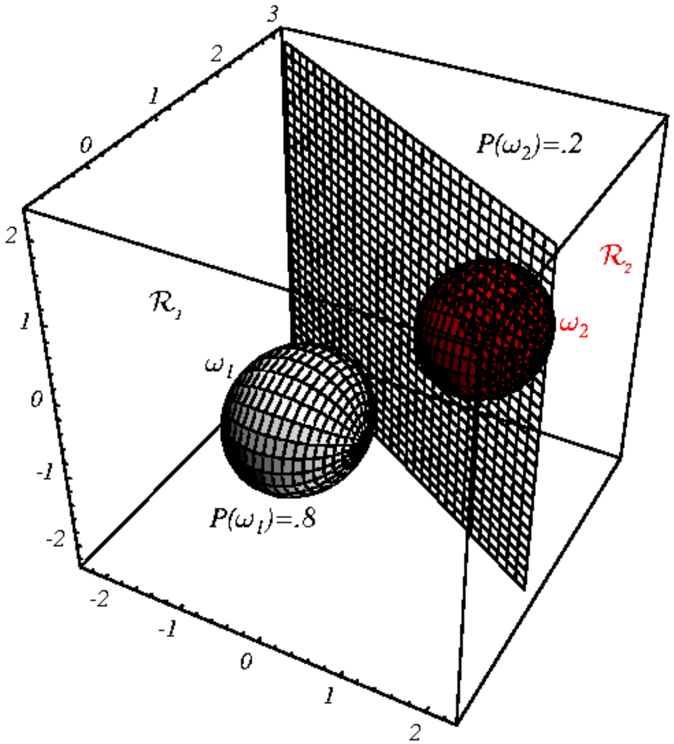
Unequal priors

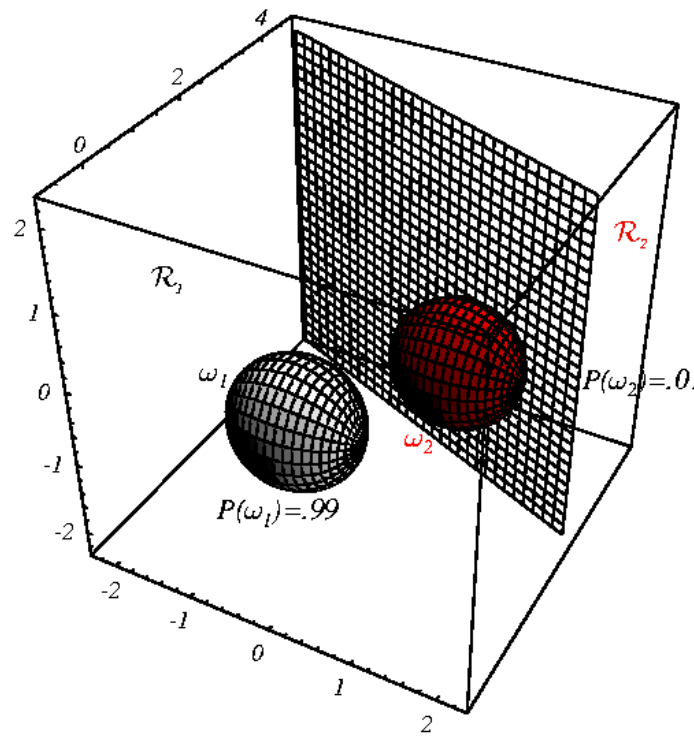


When $P(\omega_i) \neq P(\omega_j)$, the decision boundary is shifted (but still orthogonal to the segment connecting the means).

(from Duda, Hart, Stork (2001) Pattern classification)

Unequal priors





(from Duda, Hart, Stork (2001) Pattern classification)

Arbitrary covariance matrices

$$\Sigma_i = \Sigma$$

Example: $\Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}$ $\Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ $\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$ $\mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$

$$\Rightarrow \quad \Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \Sigma_1^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad |\Sigma_1| = 1 \quad |\Sigma_2| = 4$$

Let $P(\omega_1) = P(\omega_2) = 0.5 \Rightarrow \ln P(\omega_1) = \ln P(\omega_2) = -\ln 2$

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

$$g_1(\mathrm{x}) = -\frac{1}{2}[x_1-3, x_2-6]\begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}\begin{bmatrix} x_1-3 \\ x_2-6 \end{bmatrix} - \ln 2$$

$$g_2(\mathrm{x}) = -\frac{1}{2}[x_1-3, x_2+2]\begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}\begin{bmatrix} x_1-3 \\ x_2+2 \end{bmatrix} - \frac{1}{2}\ln 4 - \ln 2$$

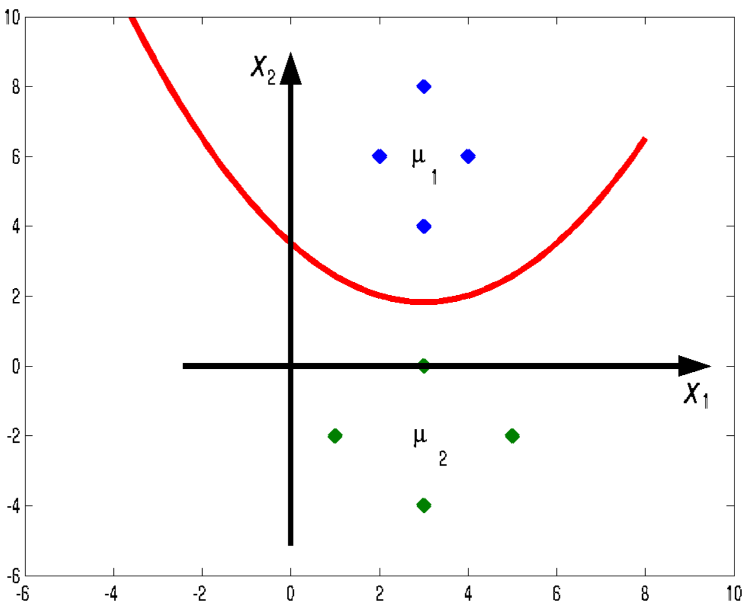
$$g_1(\mathrm{x}) = -(x_1-3)^2 - \frac{1}{4}(x_2-6)^2 - \ln 2$$

$$g_2(\mathrm{x}) = -\frac{1}{4}(x_1-3)^2 - \frac{1}{4}(x_2+2)^2 - \frac{1}{2}\ln 4 - \ln 2$$

Decision boundary is determined by $g_1(\mathrm{x}) = g_2(\mathrm{x}) \quad \Rightarrow$

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

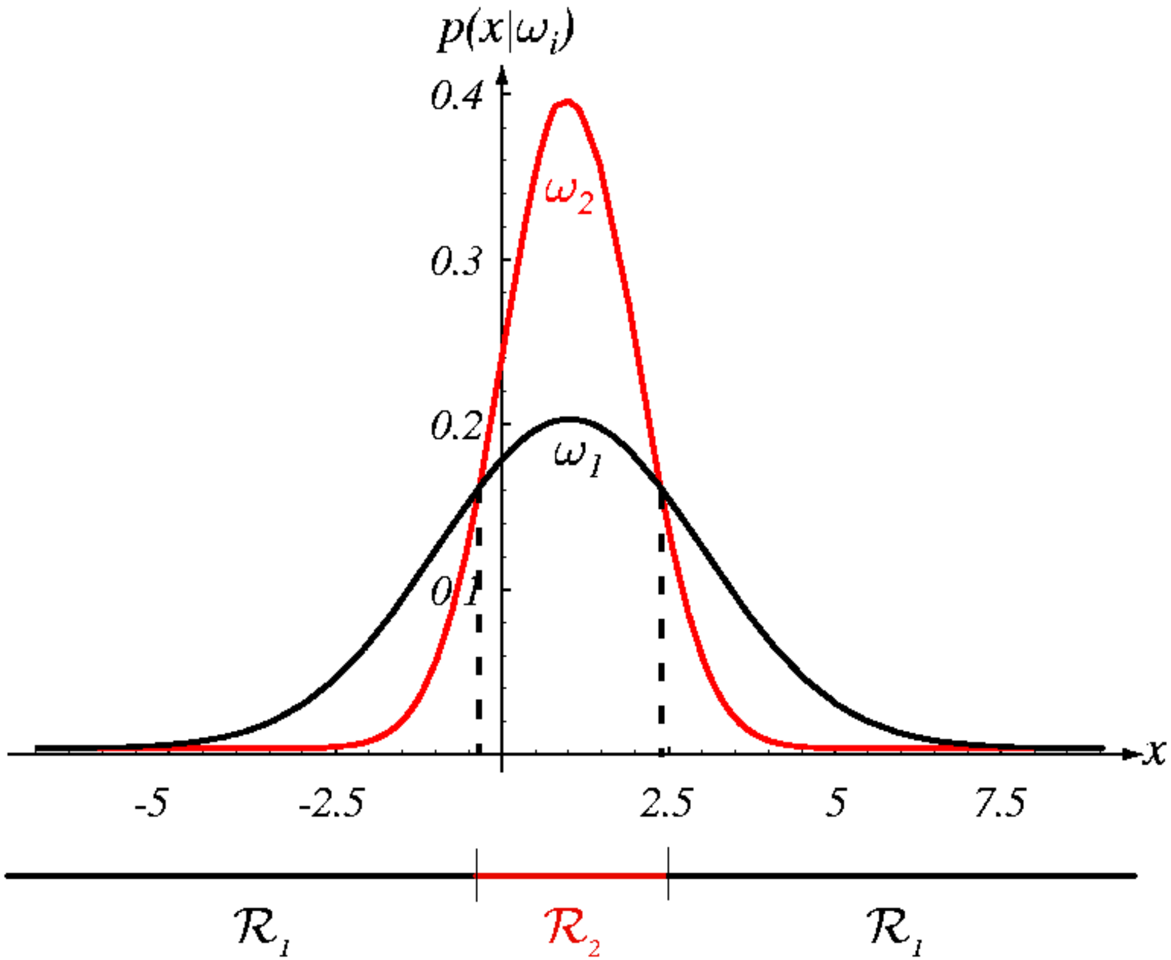
$$-(x_1 - 3)^2 - \frac{1}{4}(x_2 - 6)^2 = -\frac{1}{4}(x_1 - 3)^2 - \frac{1}{4}(x_2 + 2)^2 - \frac{1}{2}\ln 4$$



$$\Rightarrow \boxed{x_2 = 3.514 - 1.125\, x_1 + 0.1875\, x_1^{\,2}}$$

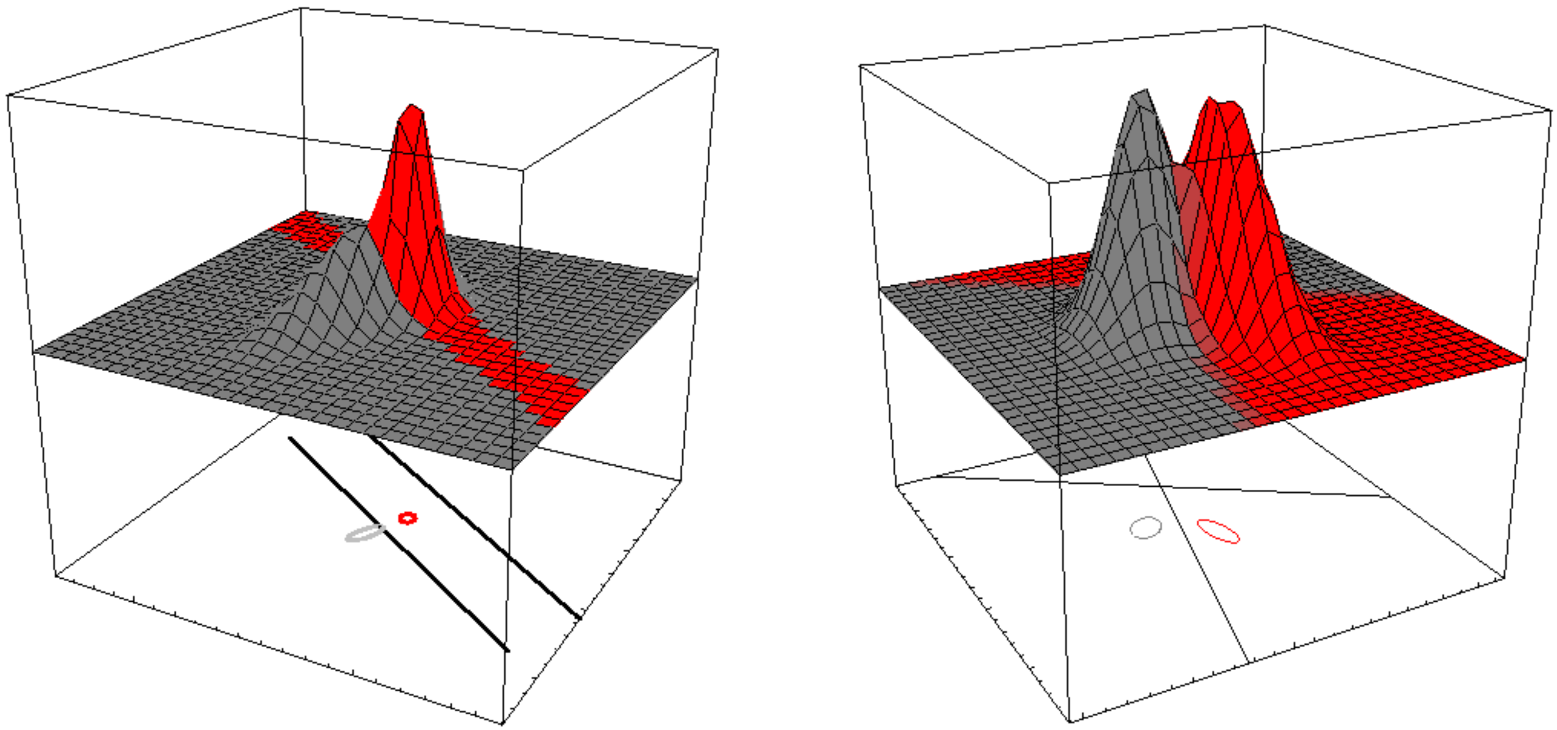
(from Duda, Hart, Stork (2001) Pattern classification)

One dimension



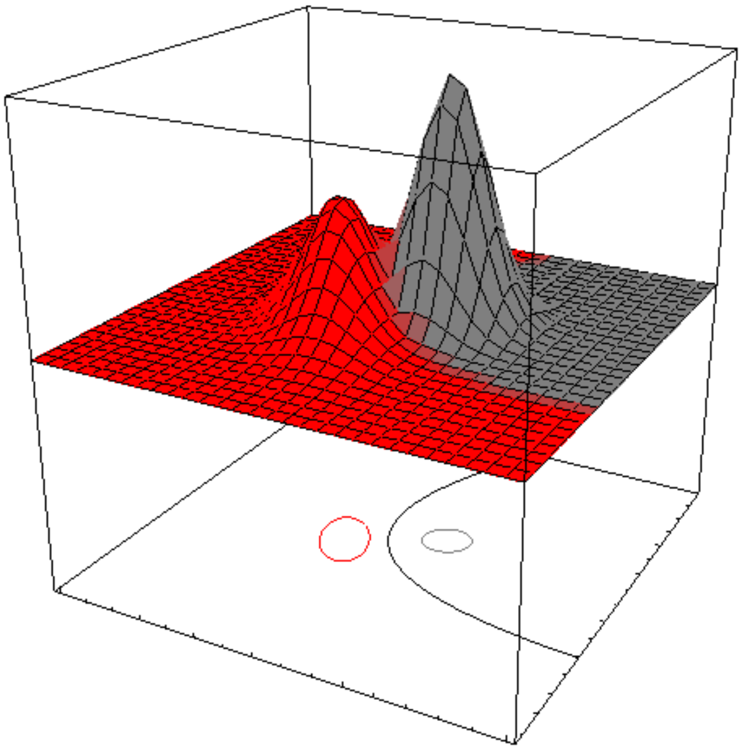
(from Duda, Hart, Stork (2001) Pattern classification)

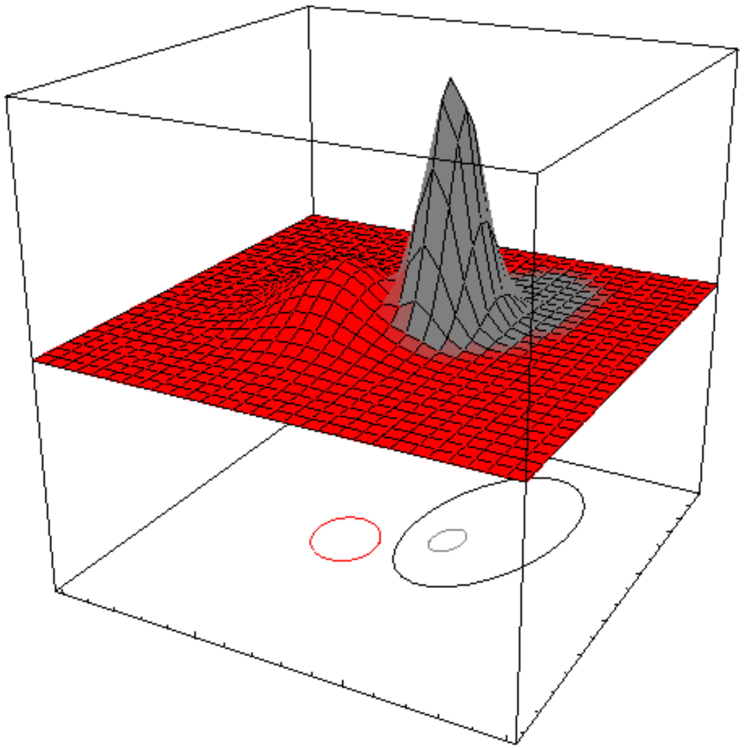
Two dimensions



(from Duda, Hart, Stork (2001) Pattern classification)

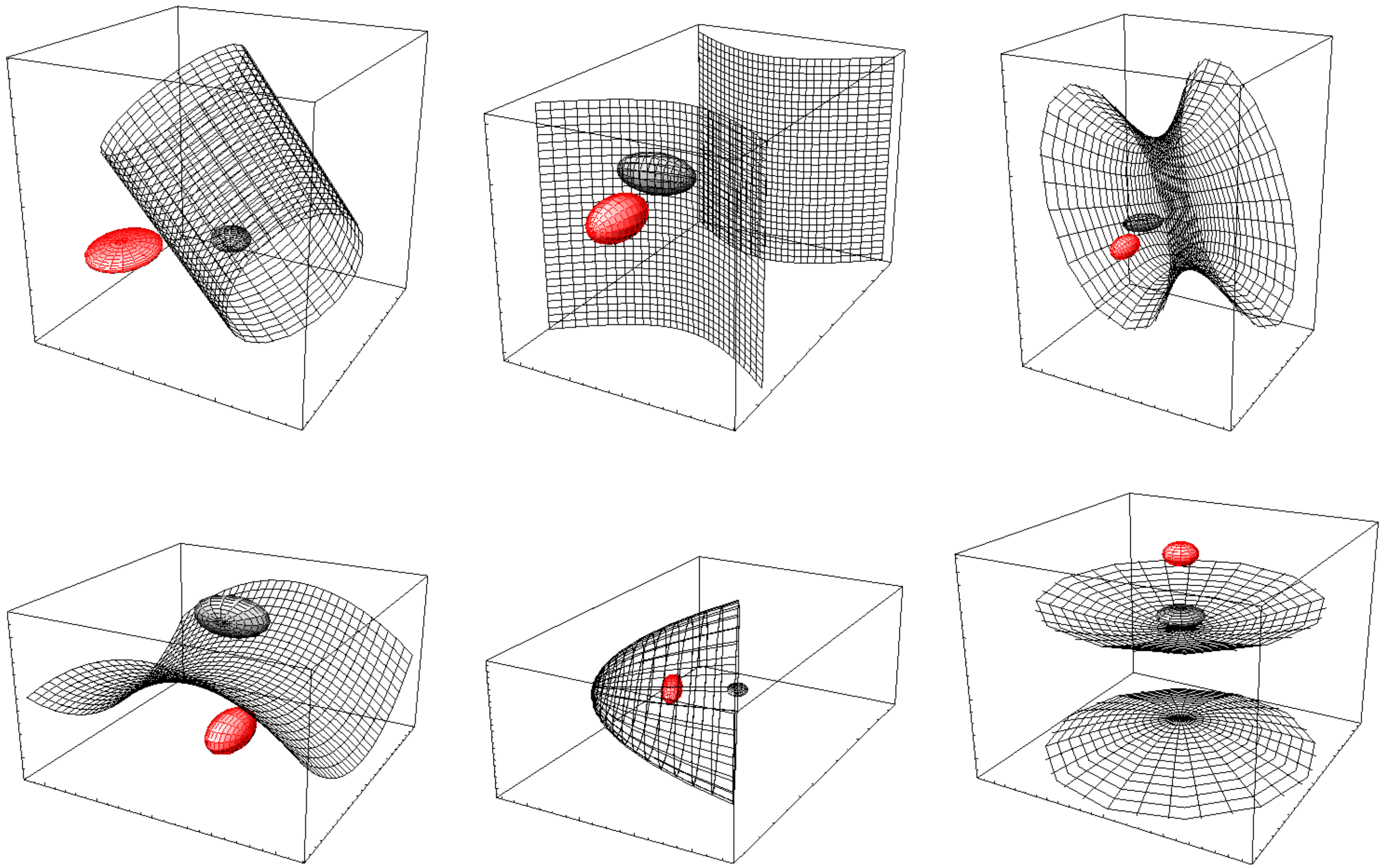
Two dimensions





(from Duda, Hart, Stork (2001) Pattern classification)

Three dimensions



(from Duda, Hart, Stork (2001) Pattern classification)

Conclusion

For multidimensional normal distributions, the discriminant functions can be computed analytically.

Classification Error

Classification Error

Example: Two-class problem

Suppose that a dichotomizer has divided the space into two regions $\mathcal{R}_1$ and $\mathcal{R}_2$. The probability of error is:

$$P(error) = P(\mathrm{x} \in \mathcal{R}_2, \omega_1) + P(\mathrm{x} \in \mathcal{R}_1, \omega_2)$$

$$= P(\mathrm{x} \in \mathcal{R}_2 \mid \omega_1) P(\omega_1) + P(\mathrm{x} \in \mathcal{R}_1 \mid \omega_2) P(\omega_2)$$

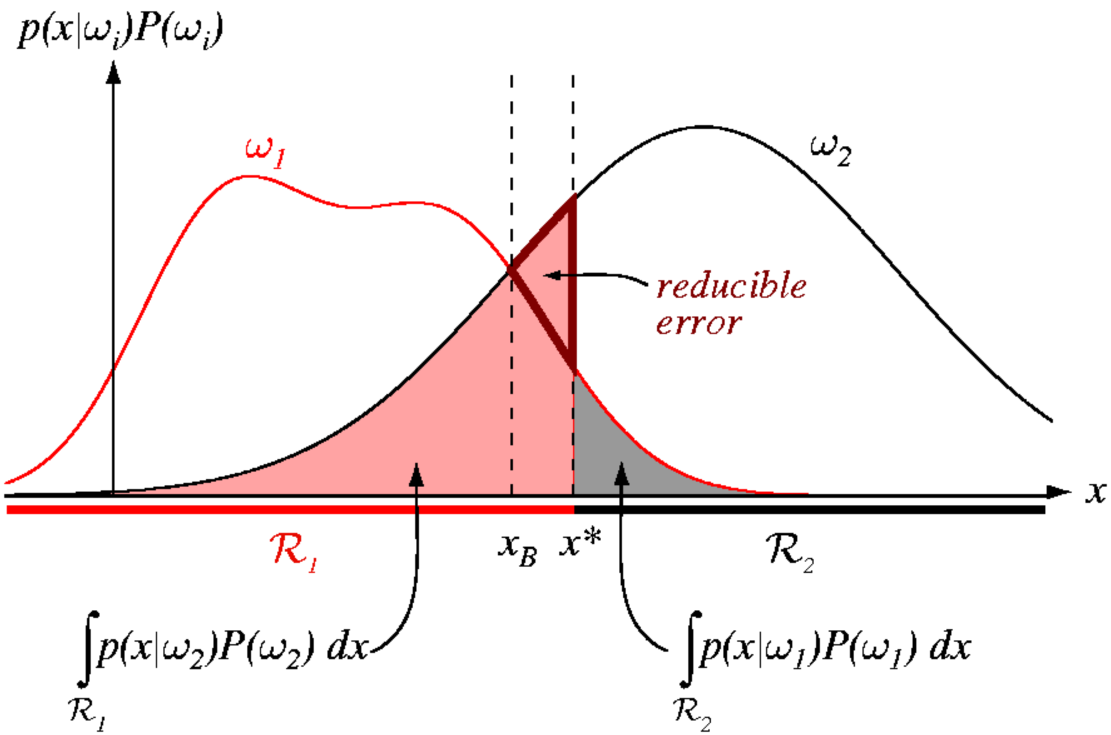
$$= \int_{\mathcal{R}_2} p(\mathrm{x} \mid \omega_1) P(\omega_1) d\mathrm{x} + \int_{\mathcal{R}_1} p(\mathrm{x} \mid \omega_2) P(\omega_2) d\mathrm{x}$$

When the class conditional pdf's are available, the above integrals can be evaluated analytically or numerically.

Classification Error - 1D case

The total error is the sum of the grey and pink areas.

Bayes optimal decision boundary gives the lowest probability of total error.



Classification Error

In the multi-category case, it is simpler to compute the probability of being correct

$$P(correct) = \sum_{i=1}^{c} P(\mathrm{x} \in \mathcal{R}_i, \omega_i)$$

$$= \sum_{i=1}^{c} P(\mathrm{x} \in \mathcal{R}_i \mid \omega_i) P(\omega_i)$$

$$= \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathrm{x} \mid \omega_i)\, P(\omega_i)\, d\mathrm{x}$$

Summary of concepts

Discriminant functions and classifier. Transformations of discriminant functions. Dichotomizer

Discriminant functions and decision boundaries for normal distributions

Univariate/Multivariate normal density

Covariance matrix. Mahalanobis distance

Linear combinations of normally distributed random variables. Whitening transform

Discriminant functions for normal densities

Statistically independent features of equal variance, linear discriminant functions, linear machine, minimum-distance classifier

Arbitrary covariance matrices, quadratic forms. Total classification error and its evaluation.