Modeling of probabilities by functions

Normal distributions

INTRODUCTION TO INTELLIGENT SYSTEMS 17/18

Classification example

Given are the sets (training data) S1 and S2

S1 = [8.70, 3.31, -13.48, 15.48, -6.17, -6.99, -14.24,
-1.10, -1.03, -3.23]

S2 = [18.21, 1.79, 95.25, 65.02, 27.82, 32.70,
42.18, 34.76, 23.59, 53.68, 12.23, 74.15,
-0.26, 28.53, 52.45]

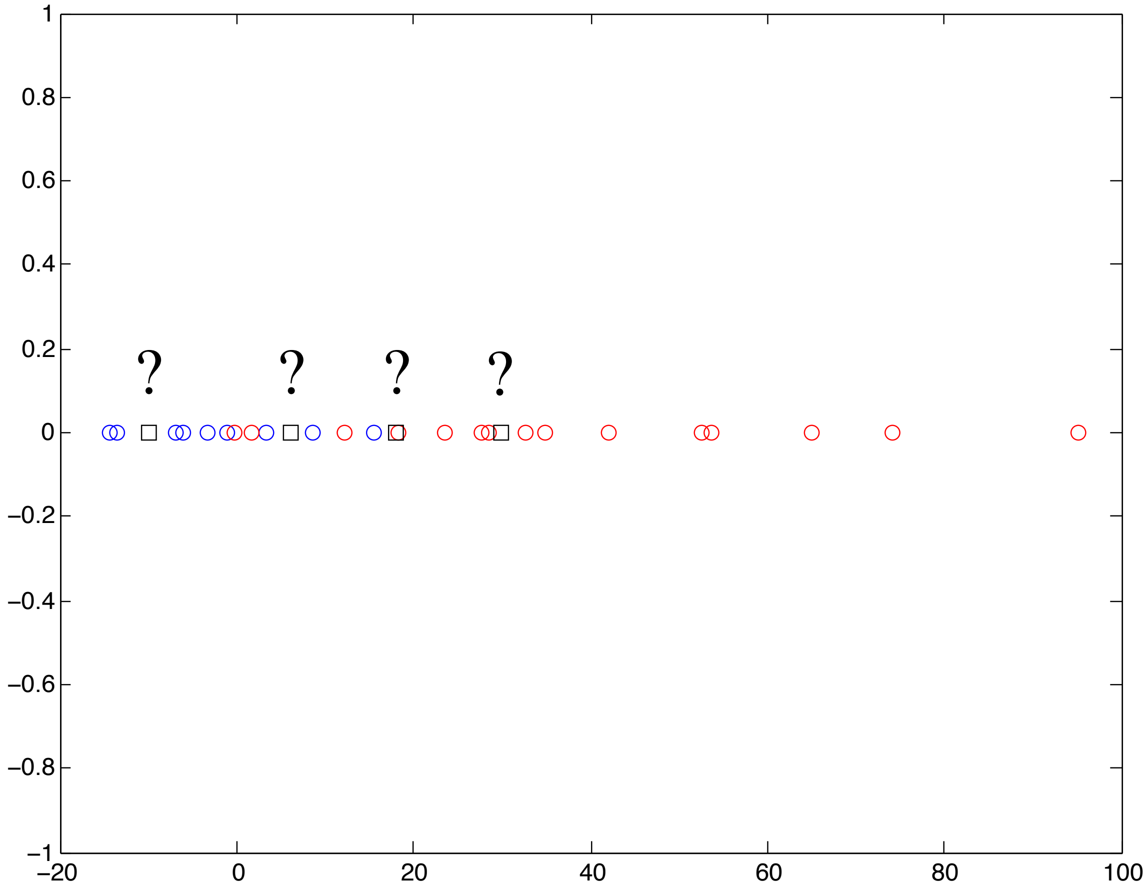
that come from two different classes.

We want to classify the following test points:

-10, 6, 18, 30

Example

Given the observed data of class 1 (blue) and of class 2 (red), what class labels should be assigned to the test points (black)?



knn classifier (k = 5)

Test point x = -10.

The 5 nearest neighbors of this point are:

S1 = [8.70, 3.31, **-13.48**, 15.48, **-6.17, -6.99, -14.24**, -1.10, -1.03, **-3.23**]

S2 = [18.21, 1.79, 95.25, 65.02, 27.82, 32.70, 42.18, 34.76, 23.59, 53.68, 12.23, 74.15, -0.26, 28.53, 52.45]

All 5 nearest neighbors come from S_1.

Hence, point -10 is decided to belong to class 1.

knn classifier (k = 5)

Test point x = 6.

The 5 nearest neighbors of this point are

S1 = [**8.70, 3.31**, -13.48, 15.48, -6.17, -6.99, -14.24, -1.10, -1.03, -3.23]

S2 = [18.21, **1.79**, 95.25, 65.02, 27.82, 32.70, 42.18, 34.76, 23.59, 53.68, **12.23**, 74.15, **-0.26**, 28.53, 52.45]

Most of the 5 nearest neighbors come from S_2.

Hence, point 6 is decided to belong to class 2.

knn classifier (k = 5)

Test point x = 18.

The 5 nearest neighbors of this point are:

S1 = [**8.70,** 3.31, -13.48, **15.48**, -6.17, -6.99, -14.24,

-1.10, -1.03, -3.23]

S2 = [**18.21**, 1.79, 95.25, 65.02, **27.82**, 32.70,
42.18, 34.76, 23.59, 53.68, **12.23**, 74.15,
-0.26, 28.53, 52.45]

Most of the 5 nearest neighbors come from S_2.

Hence, point 18 is decided to belong to class 2.

knn classifier (k = 5)

Test point x = 30.

The 5 nearest neighbors of this point are:

S1 = [8.70, 3.31, -13.48, 15.48, -6.17, -6.99, -14.24,
-1.10, -1.03, -3.23]

S2 = [18.21, 1.79, 95.25, 65.02, **27.82, 32.70**, 42.18, **34.76**, **23.59**, 53.68, 12.23, 74.15, -0.26, **28.53**, 52.45]

All 5 nearest neighbors come from S_2.

Hence, point 30 is decided to belong to class 2.

Some doubts in knn …

Hmmm, this counting of nearest neighbors seems a little bit shaky – if the votes are 3:2 and I move the test point a little bit, they may become 2:3. The probabilities seem to jump and fall from point to point. How reliable is this?

I will model the probabilities by some smooth, reliable and predictably changing functions. For instance, a **Gaussian** function!

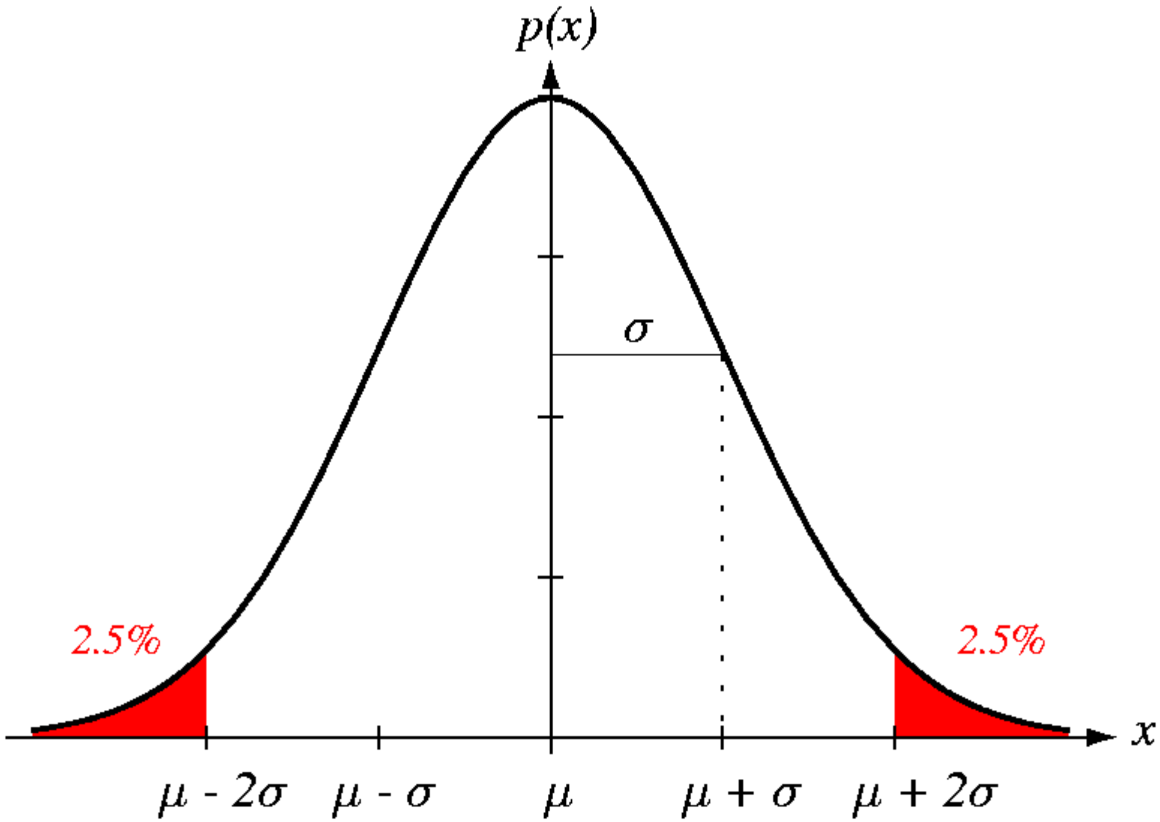
Uni-variate normal density: one-dimensional Gaussian function

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

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

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

Gaussian function – some properties



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

from Duda, Hart, Stork (2001) Pattern classification

Back to our example

S1 = [8.70, 3.31, -13.48, 15.48, -6.17, -6.99, -14.24, -1.10, -1.03, -3.23]

S2 = [18.21, 1.79, 95.25, 65.02, 27.82, 32.70, 42.18, 34.76, 23.59, 53.68, 12.23, 74.15, -0.26, 28.53, 52.45]

We decide to model the two classes that generated this data by two normal distributions.

What are the parameters of these normal distributions?

Maximum likelihood estimation

We do not know what the real values of the parameters of the two distributions are.

We assume as most likely values:

- The mean μ_1 of the normal distribution that generated S_1 is equal to the mean of S_1.
- The std σ_1 of the normal distribution that generated S_1 is equal to the std of S_1.

Maximum likelihood estimation of the parameters of a normal distribution

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu})^2$$

We compute $\mu_1 = 0$, $\sigma_1 = 10$

$\mu_2 = 35$, $\sigma_2 = 20$

Estimated (class-conditional) probability density functions

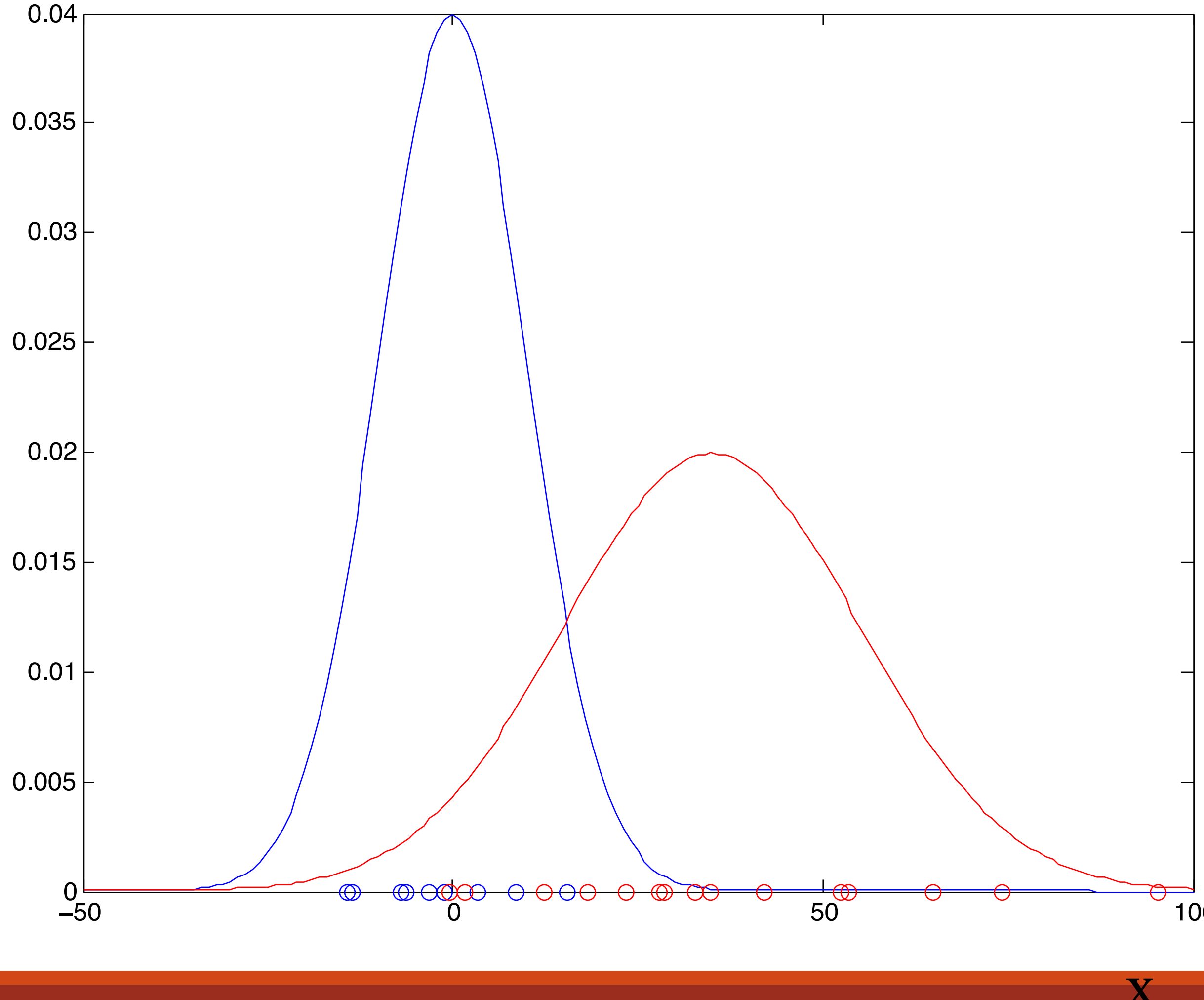
$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} = \frac{1}{10\sqrt{2\pi}} e^{-\frac{(x)^2}{2*10^2}}$$

$$p(x|\omega_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} = \frac{1}{20\sqrt{2\pi}} e^{-\frac{(x-35)^2}{2*20^2}}$$

Formulas are not really attractive… even for the normal distribution

Estimated (class-conditional) probability density functions

$p(x|\omega_1)$
$p(x|\omega_2)$



Estimation of prior probabilities

From class 1, we observe 10 values: card(S_1)=10.

From class 2, we observe 15 values: card(S_2)=15.

We estimate the prior probabilities of class 1 and class 2 by the relative frequencies of occurrence:

$$P(\omega_1) = \frac{|S_1|}{|S_1| + |S_2|} = \frac{10}{10 + 15} = 0.4$$

$$P(\omega_2) = \frac{|S_2|}{|S_1| + |S_2|} = \frac{15}{10 + 15} = 0.6$$

Posterior probabilities

$$P(\omega_1|x) = \frac{P(\omega_1)p(x|\omega_1)}{p(x)} = \frac{0.4\frac{1}{10\sqrt{2\pi}}e^{-\frac{(x)^2}{2*10^2}}}{p(x)}$$

$$P(\omega_2|x) = \frac{P(\omega_2)p(x|\omega_2)}{p(x)} = \frac{0.6\frac{1}{20\sqrt{2\pi}}e^{-\frac{(x-35)^2}{2*20^2}}}{p(x)}$$

where:

$$p(x) = 0.4\frac{1}{10\sqrt{2\pi}}e^{-\frac{(x)^2}{2*10^2}} + 0.6\frac{1}{20\sqrt{2\pi}}e^{-\frac{(x-35)^2}{2*20^2}}$$

Evaluation of posterior probabilities and classification

The formulas are sufficient to compute the posterior probabilities of the classes ω_1 and ω_2 for any point x, such as our test points -10, 6, 18 and 30.

There is even something better

- we can compute the value of a decision criterion that separates the classes on the x-axis.

Decision criterion

For the decision criterion, it holds $P(\omega_1|x) = P(\omega_2|x)$

This leads to the following equation:

$$0.4\frac{1}{10\sqrt{2\pi}}e^{-\frac{(x)^2}{2*10^2}} = 0.6\frac{1}{20\sqrt{2\pi}}e^{-\frac{(x-35)^2}{2*20^2}}$$

that can be simplified to

$$8e^{-\frac{(x)^2}{2*10^2}} = 6e^{-\frac{(x-35)^2}{2*20^2}}$$

Decision criterion

We take the logarithm of both sides of the above equation and obtain:

$$\ln 8 - \frac{(x)^2}{2 * 10^2} = \ln 6 - \frac{(x-35)^2}{2 * 20^2}$$

and simplify it to

-> $3x^2 + 70x - 1455 = 0$

Decision criterion

The quadratic equation (familiar from high school)

$$3x^2 + 70x - 1455 = 0$$

Has the following solutions:

$x_1 = -36.6$, $x_2 = 13.25$

We use them for classification:

If $-36.6 < x < 13.25$ then x belongs to ω_1

If $x < -36.6$ or $x > 13.25$, x belongs to ω_2

Our problem: -10 and 6 are ω_1, 18 and 30 are ω_2

Classification methods

Parametric classification: the probability densities are available (or determined) as functions; these functions have given parameters (e.g. mean and variance of a Gaussian function)

Non-parametric classification: no pdf’s are available; no assumptions are made about the pdf’s, hence no parameters of such pdf’s are used; classification is done using the available training data (what about k in knn?)

Parametric classification

Class conditional probability densities are modeled with functions of a given type, e.g. Gaussian functions

The probability density functions (pdf's) have some parameters, e.g. a mean and a variance for a Gaussian function (normal distribution)

The values of the parameters are estimated using acquired data (called training data)

Parametric classification using normal distributions

The class conditional probability density functions are assumed to be **Gaussian functions** (normal distributions)

A Gaussian pdf is characterized by the values of its parameters: a mean and a co-variance matrix

The values of the parameters are estimated using acquired data (called training data)

Parametric classification

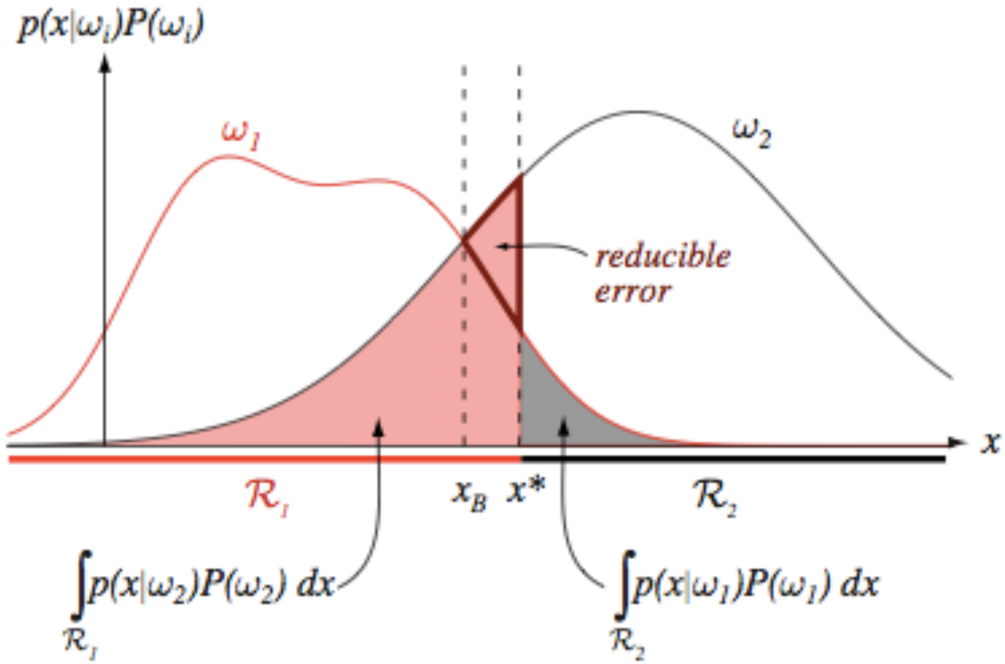
Advantages: once you have the pdf's in analytical form (i.e. mathematical expression),

- the classification of a new point is easy and fast: by evaluation of (pdf) functions or comparing to the decision criterion ☺
- the classification error can be computed (analytically or numerically), for a given point and also for the whole feature space ☺

Disadvantages:

- Assuming a pdf of a given type (e.g. Gaussian) may not be correct ☹

Estimating the classification error of a parametric classifier



from Duda, Hart, Stork (2001) Pattern classification

The grey and pink areas are the classification errors for the two classes. They can be computed for given pdf's.

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 (d-dimensional vector)
- $\sum$ is the covariance matrix ($|\Sigma|$ its determinant and Σ^{-1} its inverse)

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

Covariance matrix

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

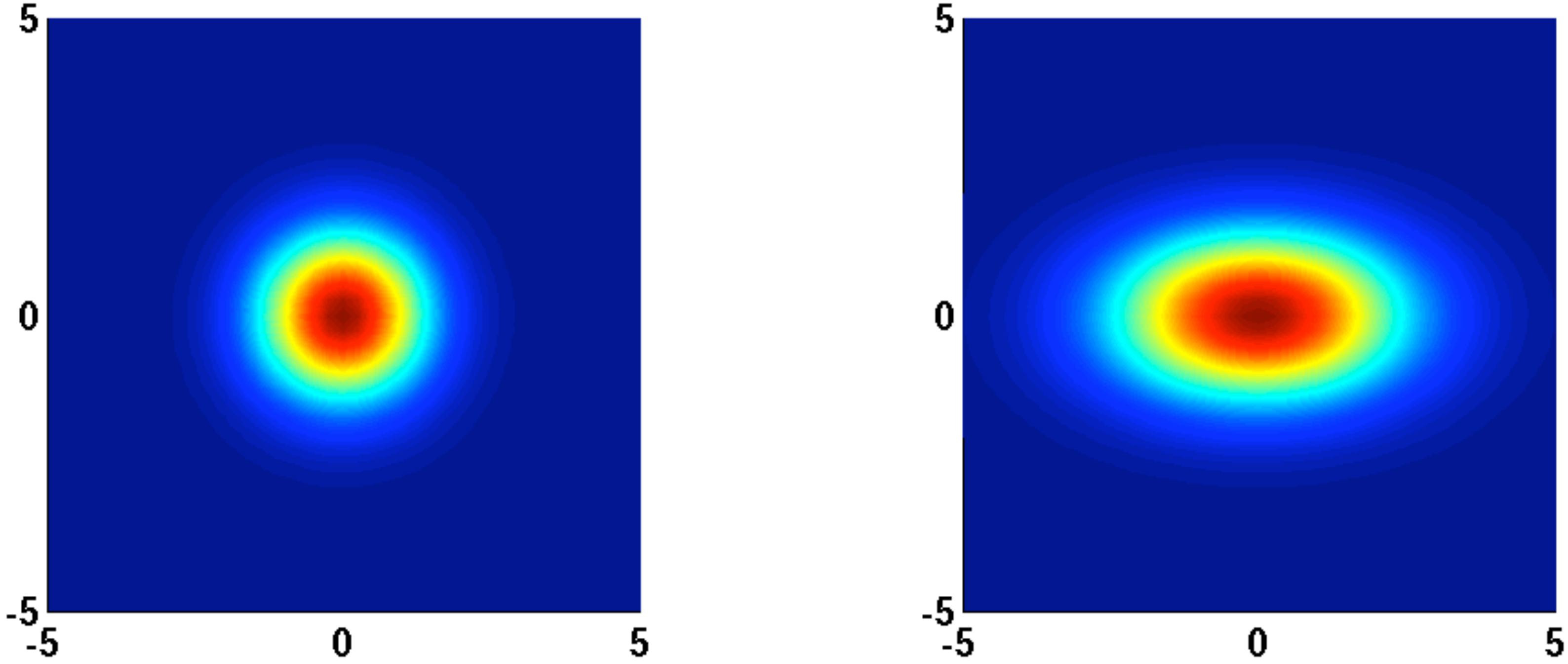
$$\varepsilon[(x_i - \mu_i)(x_j - \mu_j)] = \int (x_i - \mu_i)(x_j - \mu_j)\, p(x)\, dx = \sigma_{i,j}$$

$\sum$ 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.

Multivariate Gaussians

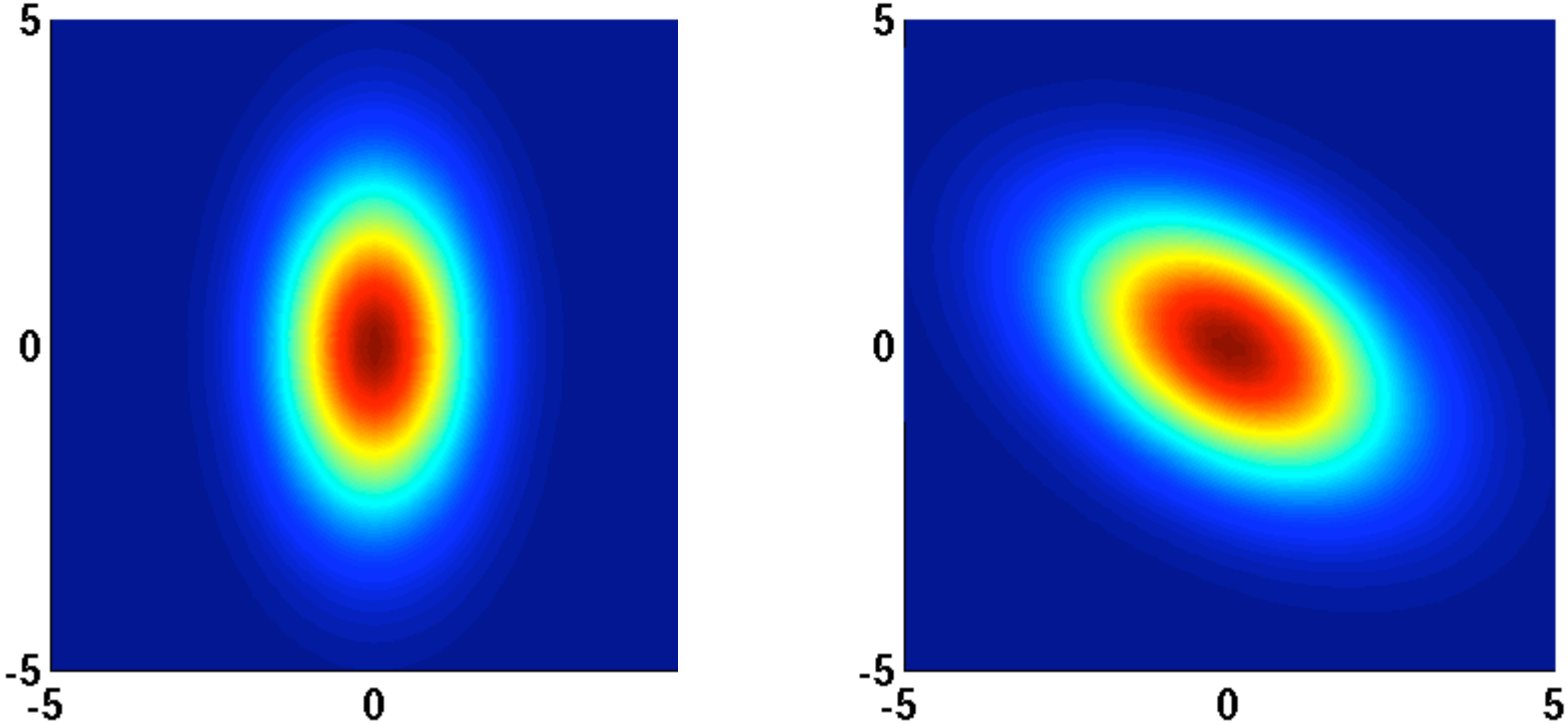


$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

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

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

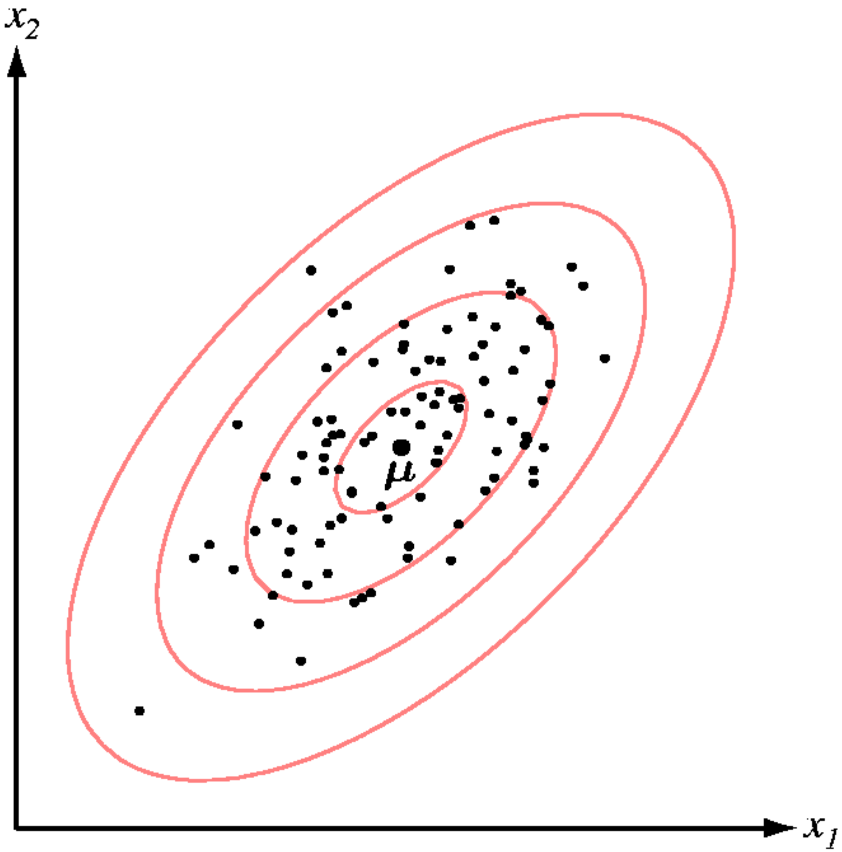
Multivariate Gaussians



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

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

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



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

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.

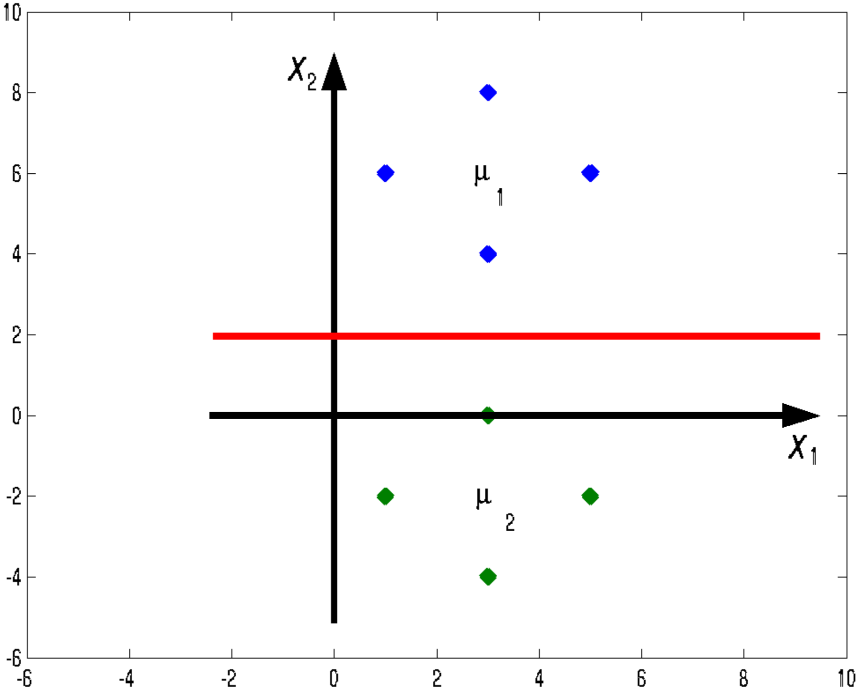
Example of how to determine the decision boundary in analytical form

Given:

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

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

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



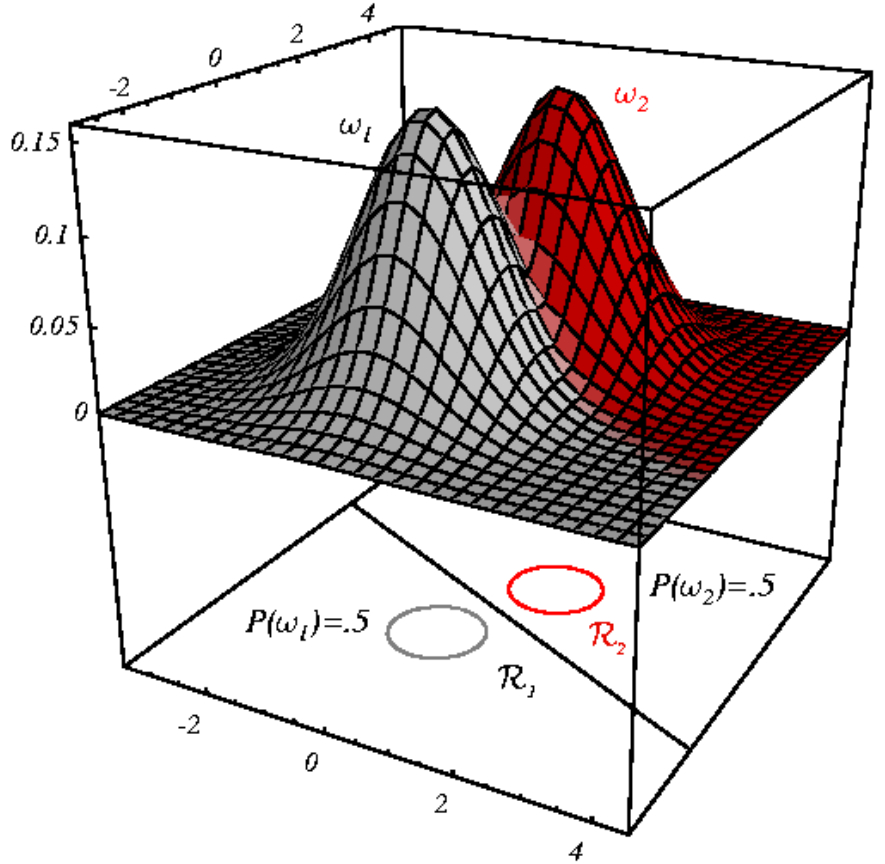
Solution:

$$P(\omega_1 \mid x) = P(\omega_2 \mid x) \;\Rightarrow$$

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

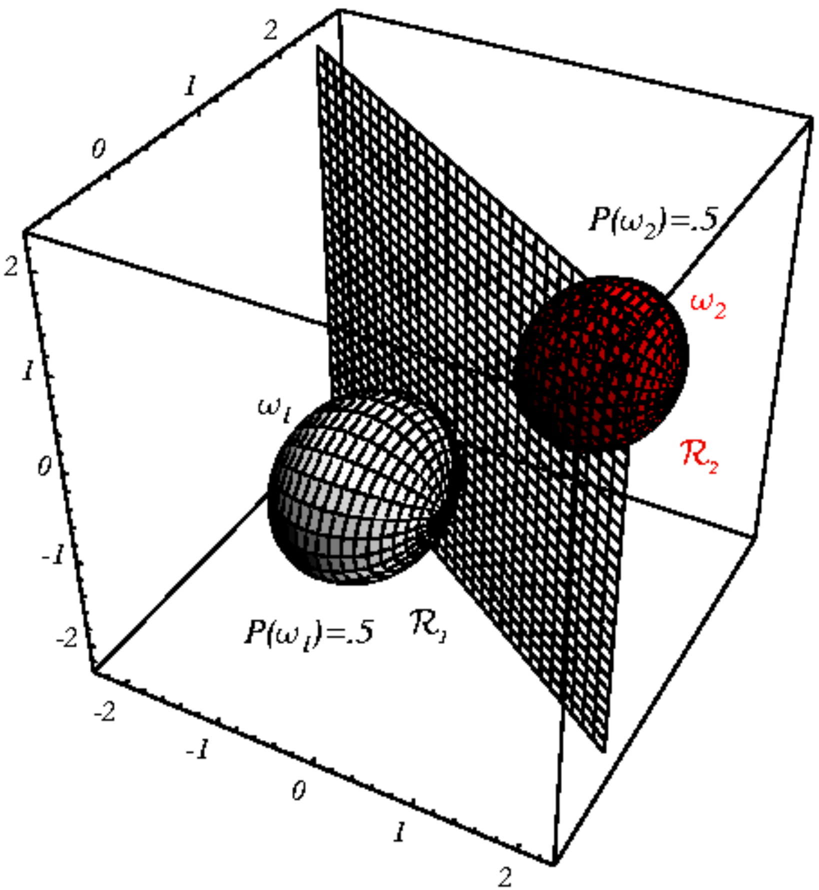
$$\Rightarrow \quad \begin{bmatrix} x_1 - 3 & x_2 - 6 \end{bmatrix} \begin{bmatrix} x_1 - 3 \\ x_2 - 6 \end{bmatrix} = \begin{bmatrix} x_1 - 3 & x_2 + 2 \end{bmatrix} \begin{bmatrix} x_1 - 3 \\ x_2 + 2 \end{bmatrix} \quad \Rightarrow \quad x_2 = 2$$

Two-dimensional case



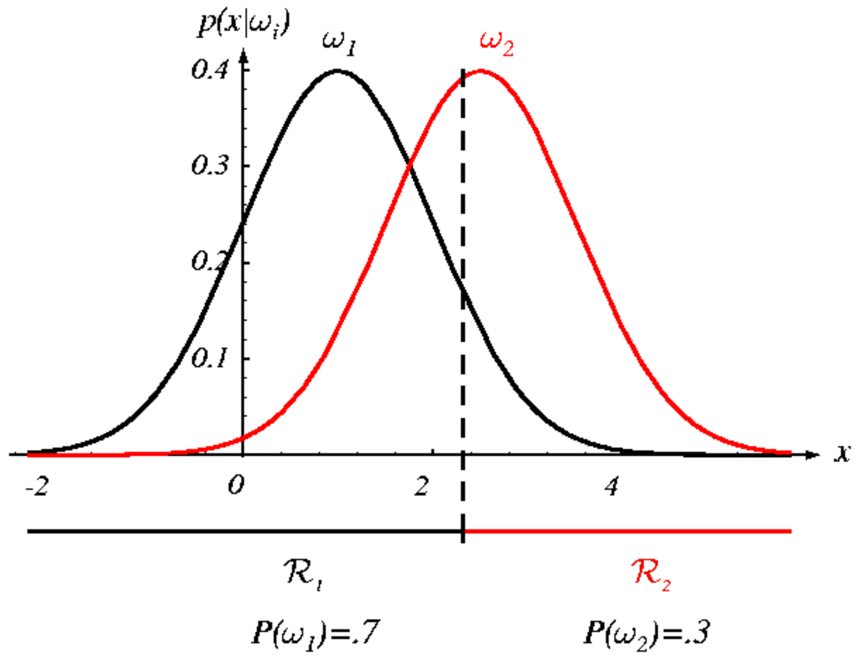
from Duda, Hart, Stork (2001) Pattern classification

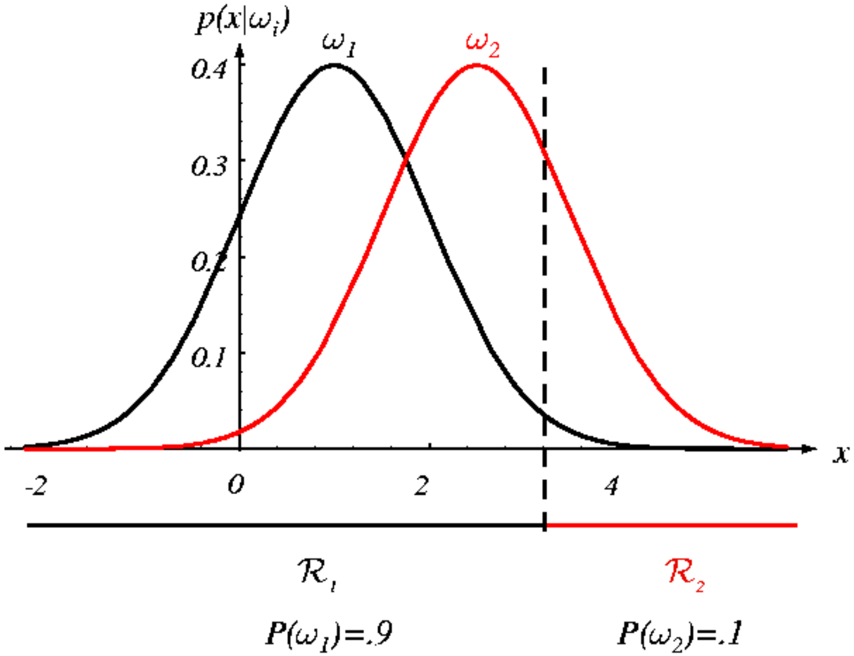
Three-dimensional case



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

Unequal priors

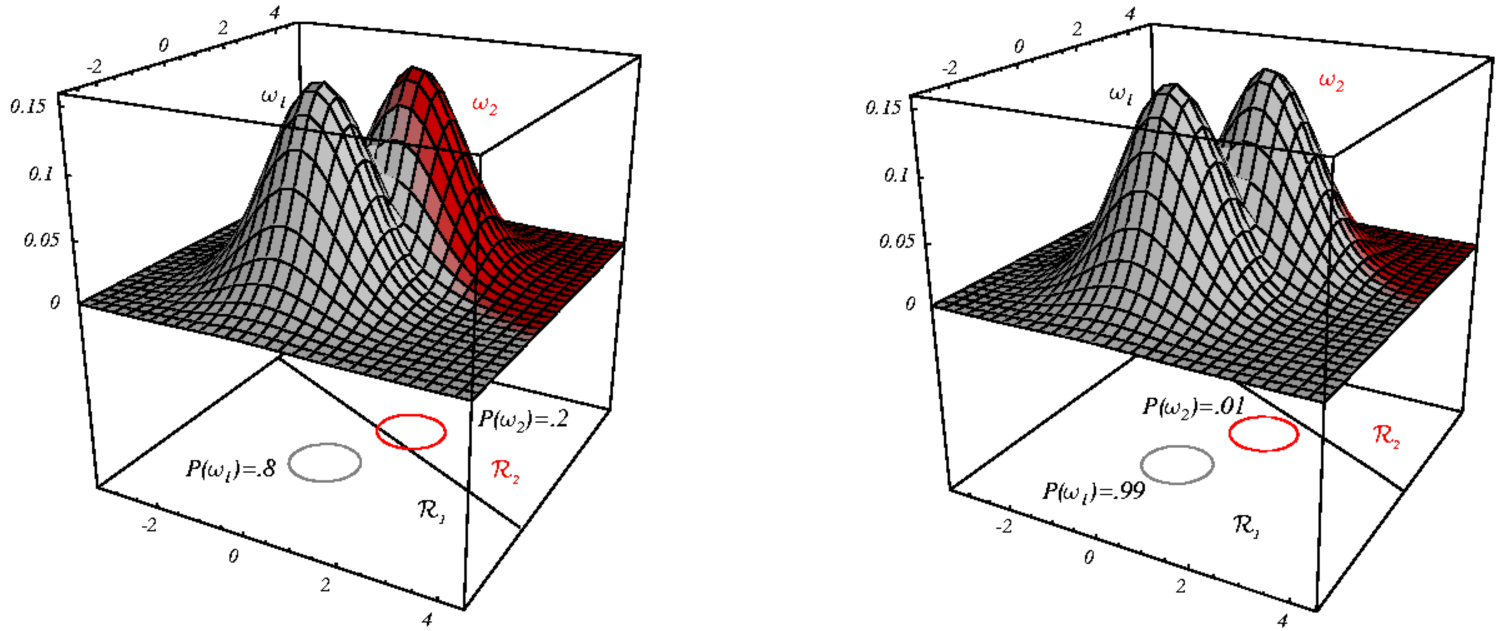




When $P(\omega_i) \neq P(\omega_j)$, the decision boundary is shifted

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

Unequal priors



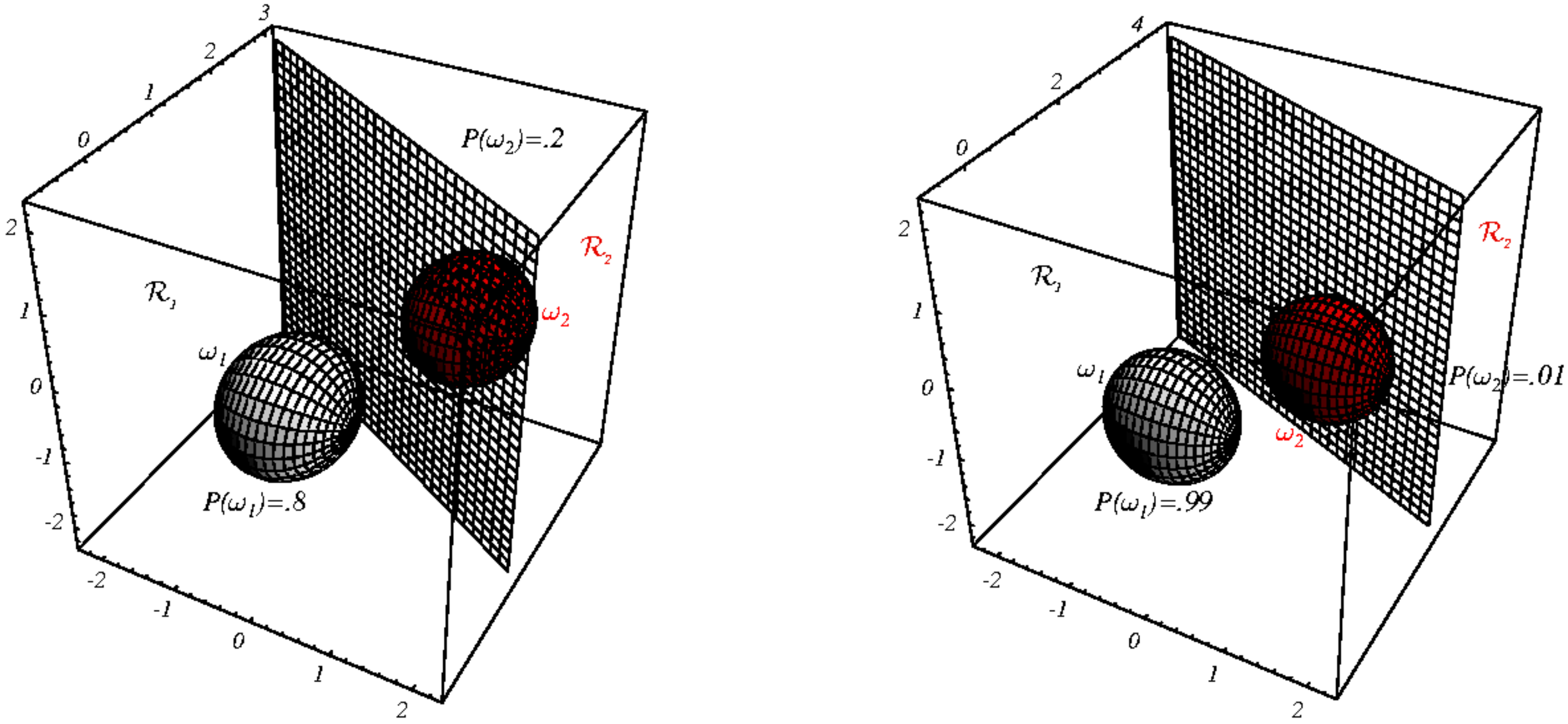
$P(\omega_i) \neq P(\omega_j)$

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

(but still orthogonal to the segment connecting the means).

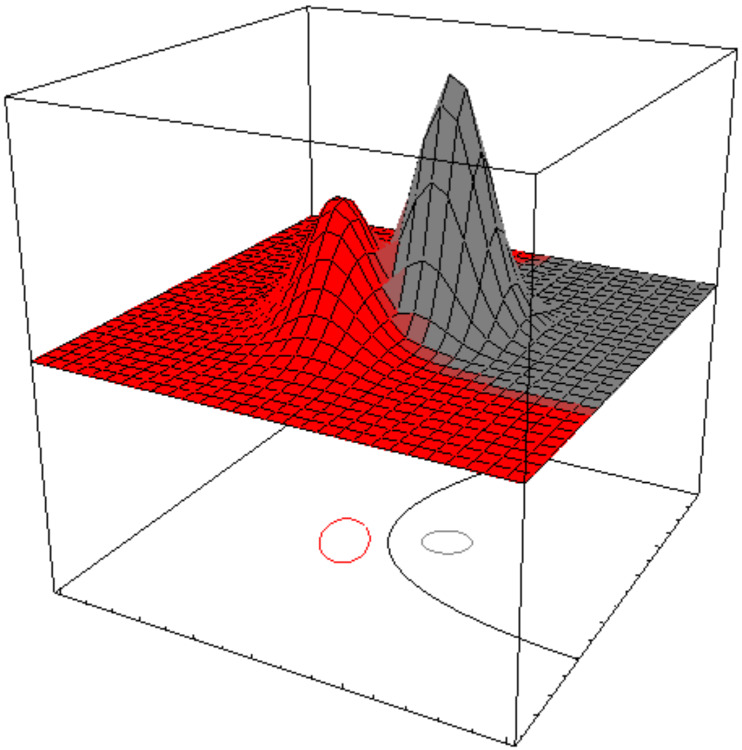
from Duda, Hart, Stork (2001) Pattern classification

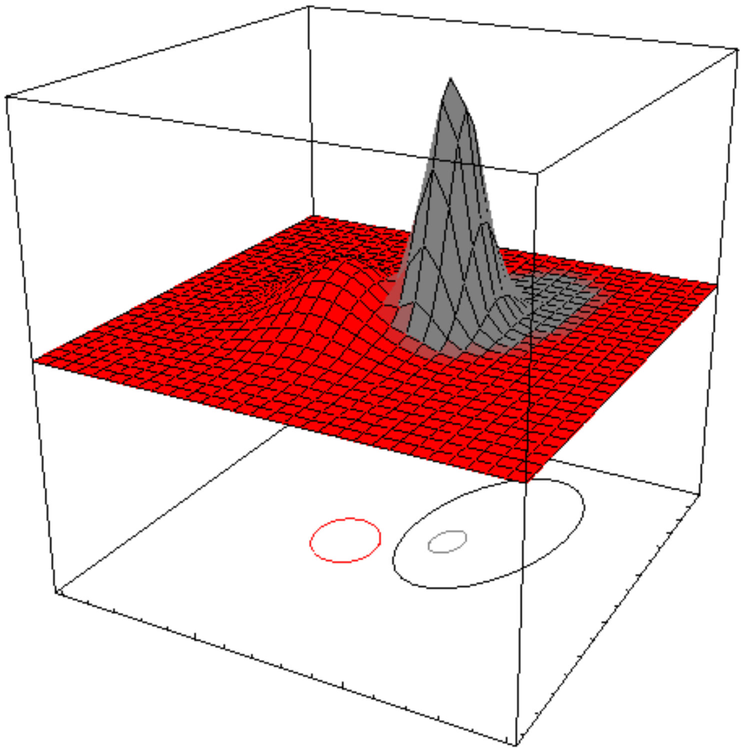
Unequal priors



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

General 2D case





from Duda, Hart, Stork (2001) Pattern classification

ML Estimation

Similarly, in the multidimensional case:

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{x}^{(\mathrm{k})}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (\mathrm{x}^{(\mathrm{k})} - \hat{\mu})(\mathrm{x}^{(\mathrm{k})} - \hat{\mu})^t$$

$$\hat{\Sigma}_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_i^{(k)} - \hat{\mu}_i)(x_j^{(k)} - \hat{\mu}_j)$$

where $x_i^{(k)}$ is the i-th feature of the k-th feature vector $\mathrm{x}^{(\mathrm{k})}$

and $\hat{\mu}_i$ is the i-th feature of the mean $\hat{\mu}$ of all n feature vectors

Summary of concepts and facts

Normal distribution, mean, standard deviation, variance, covariance matrix

Maximum likelihood estimation of the parameters of a normal distribution

How to find an analytical expression for the decision criterion for two normal distributions