Università di Verona

A.Y. 2021-22

Machine Learning & Artificial Intelligence

Bayes Decision theory

Vittorio Murino

Rev. Thomas Bayes, F.R.S (1702-1761)

Introduction

- Fundamental statistical approach to pattern classification
- Hypothesis:
  1. The decision problem is cast in probabilistic terms;
  2. All relevant probabilities are known;

- Goal:

  Discriminate the different **decision rules** using

  the ***probabilities*** and the associated ***costs***

- The classification problem is not different from regression:
    - given $\mathbf{x}$ you have to estimate the relative value of y where y is **continuous** in regression problems, while it is **discrete** (class labels) in classification problems

- Estimating the joint probability $p(\mathbf{x}, \mathrm{y})$ from the training data set is a classic *inference* problem

- Many times it is not required, and the problem is to predict a value of y associated with a certain $\mathbf{x}$, or more generally to make a **decision** (**action**) based on the prediction of the value y.

An easy example

- Let ω be the **state of nature** to be probabilistically described
- There are:
  1. Two classes ω_1 and ω_2 for which are known
     a) $P(\omega = \omega_1) = 0.7$
     b) $P(\omega = \omega_2) = 0.3$

     → **a-priori** o **prior probability**
  2. No measurements/observations.

- Decision rule:
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2

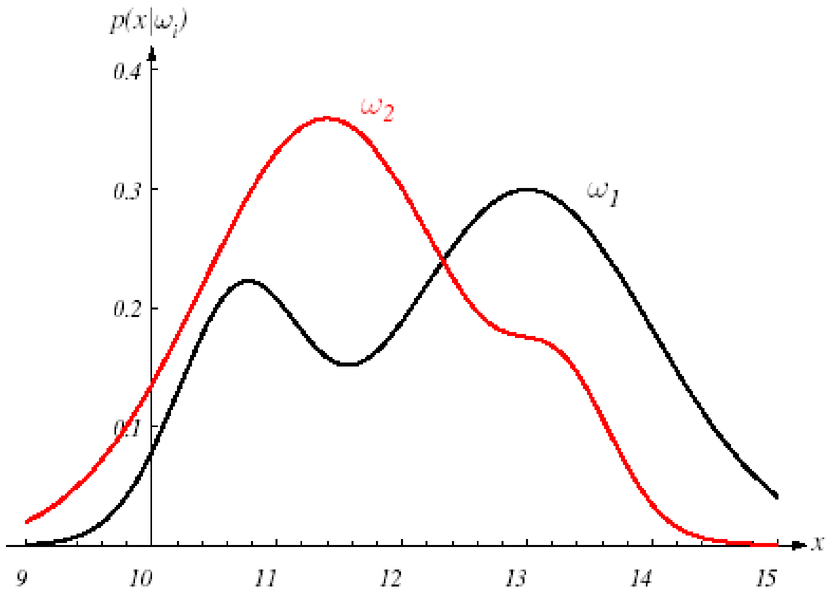
- Rather than decide, I *guess* the state of nature.

Another example – Bayes' formula

- In the previous hypothesis, with in addition the single measurement x, random variable dependent on ω_j, we can get

$$p(x \mid \omega_j)_{j=1,2}$$ = **Likelihood** *or* **class-conditional probability density function**

i.e. *the probability of having the measurement x knowing that the state of nature is ω_j.*



Fixed the measurement x, the higher $p(x|\omega_j)$ is the more likely ω_j the "right" state.

Another example – Bayes' formula (2)

- Assuming known $P(\omega_j)$ and $p(x|\omega_j)$, the decision of the state of nature becomes, for Bayes

$$p(\omega_j, x) = P(\omega_j \mid x)\, p(x) = p(x \mid \omega_j) P(\omega_j)$$

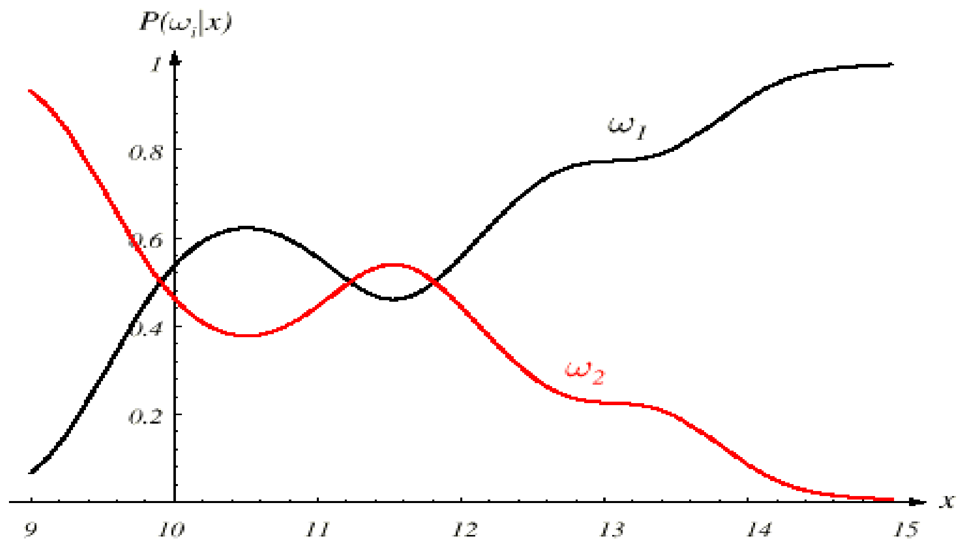
that is

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)} \propto p(x \mid \omega_j) P(\omega_j)$$

where:

- $P(\omega_j)$ = Prior
- $p(x \mid \omega_j)$ = Likelihood
- $P(\omega_j \mid x)$ = ***Posterior***
- $p(x) = \sum_{j=1}^{J} p(x \mid \omega_j) P(\omega_j)$

= ***Evidence***



Bayes decision rule

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)} \Longleftrightarrow posterior = \frac{likelihood \times prior}{evidence}$$

- *The posterior or **a-posteriori probability** is the probability that the state of nature is ω_j given the observation x.*
- The most important factor is the product $likelihood \times prior$; the evidence $p(x)$ is simply a scale factor, which ensures that

$$\sum_j P(\omega_j \mid x) = 1$$

- From the formula of Bayes derives the **Bayes' decision rule**:

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$, ω_2 otherwise

Bayes decision rule (2)

To prove the effectiveness of the Bayes decision rule:

1) We define the ***probability of error*** attached to this decision:

$$P(error \mid x) = \begin{cases} P(\omega_1 \mid x) & \text{if we decide } \omega_2 \\ P(\omega_2 \mid x) & \text{if we decide } \omega_1 \end{cases}$$

2) I prove that the **Bayes decision rule minimizes the probability of error.**
   We decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ and viceversa.

3) So, if I want to **minimize the average probability of error** on all possible observations,

$$P(error) = \int_{-\infty}^{+\infty} P(error, x)dx = \int_{-\infty}^{+\infty} P(error \mid x)\, p(x)dx$$

if, for every x, I take $P(error|x)$ as small as possible I secure the least probability of error (the factor $p(x)$ is irrelevant).

Bayes decision rule (3)

In such a case, the probability of error becomes

$$P(error|x)=min[P(\omega_1|x), P(\omega_2|x)]$$

This ensures that the Bayes decision rule

$$decide\ \ \omega_1\ if\ \ P(\omega_1|x) > P(\omega_2|x)\ ,\ otherwise\ \ \omega_2$$

minimizes the error!

Rule of equivalent decision:

- The shape of the decision rule highlights *the importance of the posterior probability*, and emphasizes the *irrelevance of the evidence*, just a scale factor that shows how frequently you observe a pattern x. By eliminating it, you get the equivalent decision rule:

$$decide\ \ \omega_1\ if\ \ p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)\ ,\ otherwise\ \ \omega_2$$

Extension of Bayes decision theory

- It's possible to extend the Bayesian approach by using:
  - More than one type of observations or **feature** *x,* e.g. weight, height, ...

$$x \rightarrow \mathbf{x} = \{x_1, x_2, \ldots, x_d\}, \mathbf{x} \in \mathrm{R}^d \text{ with } \mathrm{R}^d \;\; \textbf{\textit{feature space}}$$

  - More than two states of nature or **categories**, c

$$\omega_1, \omega_2 \rightarrow \{\omega_1, \omega_2, \ldots, \omega_c\}$$

  - **Different actions**, in addition to the choice of states of nature

$$\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$$

  - A **cost function**, more general than the probability of error, i.e. $\lambda(\alpha_i \mid \omega_j)$ which describes the cost (or loss) of the action α_i when the state is ω_j.

Extension of Bayes decision theory (2)

- This extension does not change the shape of the posterior probability, which remains:

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j)P(\omega_j)}{p(\mathbf{x})} \quad , \mathbf{x} = \{x_1, x_2, ..., x_d\}, \mathbf{x} \in \mathrm{R}^d$$

- Suppose we observe a particular $\mathbf{x}$, and we decide to carry out the action α_i : by definition, we will be subject to loss $\lambda(\alpha_i \mid \omega_j)$.

  Given the indeterminacy of ω_j , the expected loss (or **risk**) associated with this decision will be:

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$$

Conditional risk

- In this case, Bayes decision theory indicates to carry out the action that minimizes the conditional risk or, formally, a decision function $\alpha(\mathbf{x})$ such that:

$$\alpha(\mathbf{x}) \rightarrow \alpha_i \quad , \alpha_i \in \{\alpha_1, \alpha_2, ..., \alpha_a\}, \text{ such that } R(\alpha_i \mid \mathbf{x}) \text{ is minimal.}$$

Extension of Bayes decision theory (3)

- To evaluate such a function, the **overall risk** is introduced, i.e. *the expected loss given a decision rule.*
- Since $R(\alpha_i \mid \mathbf{x})$ is the conditional risk associated to the action and since the decision rule specifies the action, the overall risk is

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

- Clearly, if $\alpha(\mathbf{x})$ is chosen so that $R(\alpha_i \mid \mathbf{x})$ is as little as possible for each $\mathbf{x}$, the overall risk is minimized. So Bayes extended decision rule is:

1) Calculate $$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$$

2) Choose the action $$i^* = arg \min_i R(\alpha_i|\mathbf{x})$$

The resulting minimum overall risk is called **Bayes Risk** R^* and it is *the best performance that can be achieved.*

Two-category classification problems

- Consider Bayes' decision rule applied to binary classification problems, i.e., with two possible states of nature ω_1, ω_2, with α_i → the right state is ω_i. By definition, $\lambda_{ij}=\lambda(\alpha_i \mid \omega_j)$.
- Conditional risk becomes

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x})$$

$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

- There are many **equivalent** ways of expressing the minimum risk decision rule, each with its own advantages:
- Fundamental form: choose ω_1 if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$
  - In terms of *a posteriori probability* choose ω_1 if

$$(\lambda_{21} - \lambda_{11}) P(\omega_1 \mid \mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(\omega_2 \mid \mathbf{x}).$$

Two-category classification problems (2)

- Ordinarily, the loss for a wrong decision is greater than the loss for a right decision, therefore

$$(\lambda_{21}-\lambda_{11}),(\lambda_{12}-\lambda_{22})>0$$

- So, in practice, our decision is determined by the most likely state of nature (indicated by probability a posteriori), although scaled by the difference factor (however positive) given by the losses.
- Using Bayes, we replace the a-posteriori probability with

$$(\lambda_{21}-\lambda_{11})P(\omega_1\,|\,\mathbf{x})>(\lambda_{12}-\lambda_{22})P(\omega_2\,|\,\mathbf{x}).$$

$$(\lambda_{21}-\lambda_{11})p(\mathbf{x}\,|\,\omega_1)P(\omega_1)>(\lambda_{12}-\lambda_{22})p(\mathbf{x}\,|\,\omega_2)P(\omega_2).$$

obtaining the equivalent form *dependent upon prior* and *conditional densities*

Two-category classification problems (3)

- Another alternative form, valid for the reasonable assumption that $\lambda_{21} > \lambda_{11}$ is to decide ω_1 if

$$(\lambda_{21}-\lambda_{11})p(\mathbf{x}\,|\,\omega_1)P(\omega_1) > (\lambda_{12}-\lambda_{22})p(\mathbf{x}\,|\,\omega_2)P(\omega_2).$$

$$\frac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\frac{P(\omega_2)}{P(\omega_1)}$$

This form of decision rule focuses on x's dependence on probability densities. Consider $p(\mathbf{x}|\omega_j)$ a function of ω_j that is, the likelihood function, and we calculate the **likelihood ratio**, which translates Bayes rule as *the choice of* ω_1 *if the likelihood ratio exceeds a certain threshold*, <u>choice independent of observation $\mathbf{x}$.</u>

Minimum Error Rate Classification

- In classification problems, each state is associated with one of the c classes ω_j , and actions α_i mean that " the right state is ω_i ”.
- The loss function associated with this case is referred to as **0-1 loss** or **symmetric loss**
- $$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & \text{se} \quad \mathrm{i} = \mathrm{j} \\ 1 & \text{se} \quad \mathrm{i} \neq \mathrm{j} \end{cases}$$
- The risk corresponding to this loss function is **the average probability of error**, since the conditional risk is
- $$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x}) =$$

$$= \sum_{j \neq i}^{c} P(\omega_j \mid \mathbf{x}) = 1 - P(\omega_i \mid \mathbf{x})$$

and $P(\omega_i \mid \mathbf{x})$ is the probability that action α_i is correct.

Classification *Minimum Error Rate* (2)

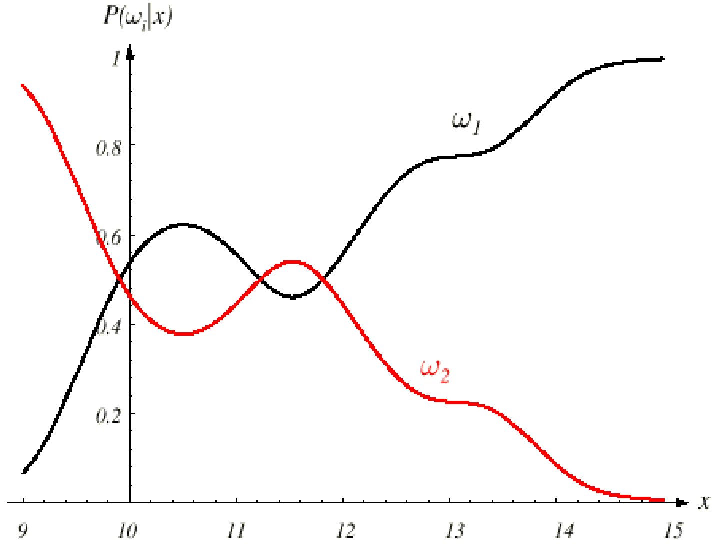
- To minimize the total risk, i.e. in this case to minimize the average probability of error, we must choose i that <u>maximizes</u> the *probability a posteriori* $P(\omega_i \mid \mathbf{x})$, that is, for the **Minimum Error Rate**:

$$\textit{decide} \quad \omega_i \quad \textit{if} \;\; P(\omega_i \mid \boldsymbol{x}) > P(\omega_j \mid \boldsymbol{x}) \quad \textit{for each } j \neq i$$

Recap

Bayes' formula

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j) P(\omega_j)}{p(\mathbf{x})}$$



Bayes decision rule:

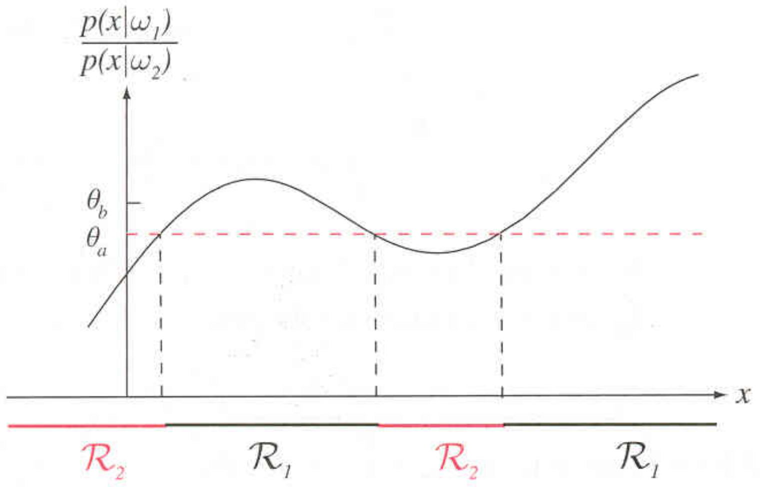
decide ω_1 *if* $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$, ω_2 *otherwise,*

equiv. $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$

With the loss function, the rule does not change:

decide ω_1 *if* $(\lambda_{21} - \lambda_{11}) p(\mathbf{x} \mid \omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) p(\mathbf{x} \mid \omega_2) P(\omega_2)$, *otherwise* ω_2

<u>*and allows you to minimize the risk!*</u>



By rearranging the likelihoods we have

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

where (Minimum Error Rate)

$$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & \text{se i} = \text{j} \\ 1 & \text{se i} \neq \text{j} \end{cases}$$

from which I reconnect to the initial rule!

Decision Theory

- In practice, the problem can be split into an ***inference*** phase where data are used to train a model $p(\omega_k|\mathbf{x})$ and a subsequent ***decision*** step, in which the *posterior* is used to make the choice of the class.
- An alternative is to solve the 2 problems at the same time and train a function that maps the input $\mathbf{x}$ directly in the space of decisions, that is, of the classes

  → **discriminating functions**

- There are 3 approaches to solve the decision problem (in descending order of complexity):
  1) Solve for the inference problem first to determine the *class-conditional* densities for each individual class, also infer the *priors* and then use Bayes to find the *posterior* and then determine the class (based on decision theory.

- Alternatively, the joint $p(\mathbf{x},\omega_k)$ can be modeled directly and then normalize to obtain the posterior

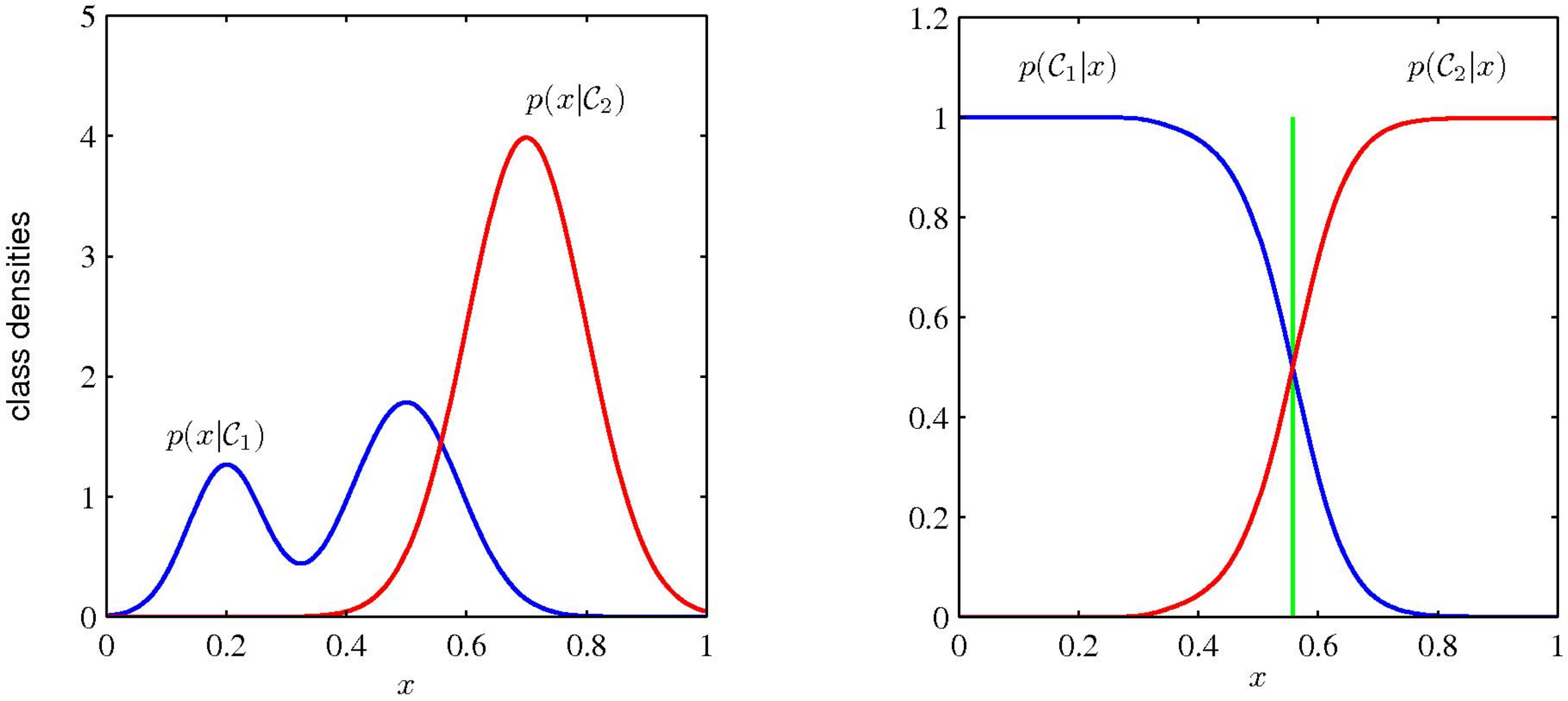
  ➔ **Generative models**

2. Solve the inference problem first to determine the *posterior* *directly* and then use decision theory to decide the class

  ➔ **Discriminatory models**

3. Find an $f(\mathbf{x})$ function, called a discriminating function, that maps $\mathbf{x}$ input directly to a class label

- Each approach has some advantages and disadvantages.
- Generative methods are more complex, requiring "good" training sets, but have the advantage of being able to manipulate all the variables at play.
- But when the problem is classification (decision, action), then the discriminative methods are more efficient, also because sometimes the *class-conditional* probabilities have a complex profile but that does not affect the *posterior.*
- Even better would be to use the discriminating functions, that is, to find directly the separation surface between the classes.



class densities
$p(x|\mathcal{C}_1)$
$p(x|\mathcal{C}_2)$
x
$p(\mathcal{C}_1|x)$
$p(\mathcal{C}_2|x)$
x

- However, estimating the posterior is often times useful because:
  - the risk is minimized when the loss matrix changes over time;
  - the *priors* of the classes can be balanced when the training set is unbalanced;
  - one can combine the models in case a complex problem needs to be subdivided into simpler problems, and then "merge" the results (*naive Bayes* under the conditional independence hypothesis)

$$\begin{aligned} p(\omega_j \mid \mathbf{x}_A, \mathbf{x}_B) &\propto p(\mathbf{x}_A, \mathbf{x}_B \mid \omega_j)\, p(\omega_j) \\ &\propto p(\mathbf{x}_A \mid \omega_j)\, p(\mathbf{x}_B \mid \omega_j)\, p(\omega_j) \quad \textit{naive Bayes} \\ &\propto \frac{p(\omega_j \mid \mathbf{x}_A)\, p(\omega_j \mid \mathbf{x}_B)}{p(\omega_j)} \end{aligned}$$

Classifiers, discriminating functions and separation surfaces

- One of the various methods of representing pattern classifiers is a set of ***discriminating functions*** $g_i(\mathbf{x}),\ i=1,\ldots,c$
- The classifier assigns the feature vector $\mathbf{x}$ to the class ω_i if

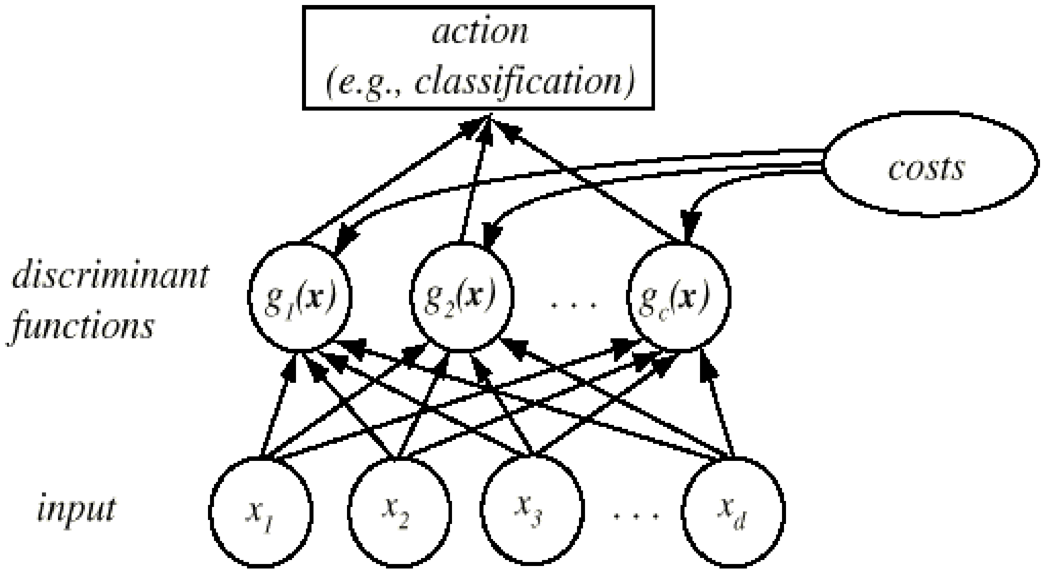
$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \text{ for each } j \neq i$$

- Such a classifier can be considered as a network that calculates c discriminating functions and chooses the function that discriminates the most.

- A Bayes classifier lends itself easily to this representation:

Generic Risk $\qquad g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$

Minimum Error Rate $\qquad g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$



Classifiers, discriminating functions and separation surfaces (2)

- There are many equivalent discriminating functions. For example, all those for which the classification results are the same
- For example, if f è una funzione monotona crescente, then

$$g_i(\mathbf{x}) \Leftrightarrow f(g_i(\mathbf{x}))$$

Some forms of discriminating functions are easier to understand or to be calculated

Minimum Error Rate

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i),$$

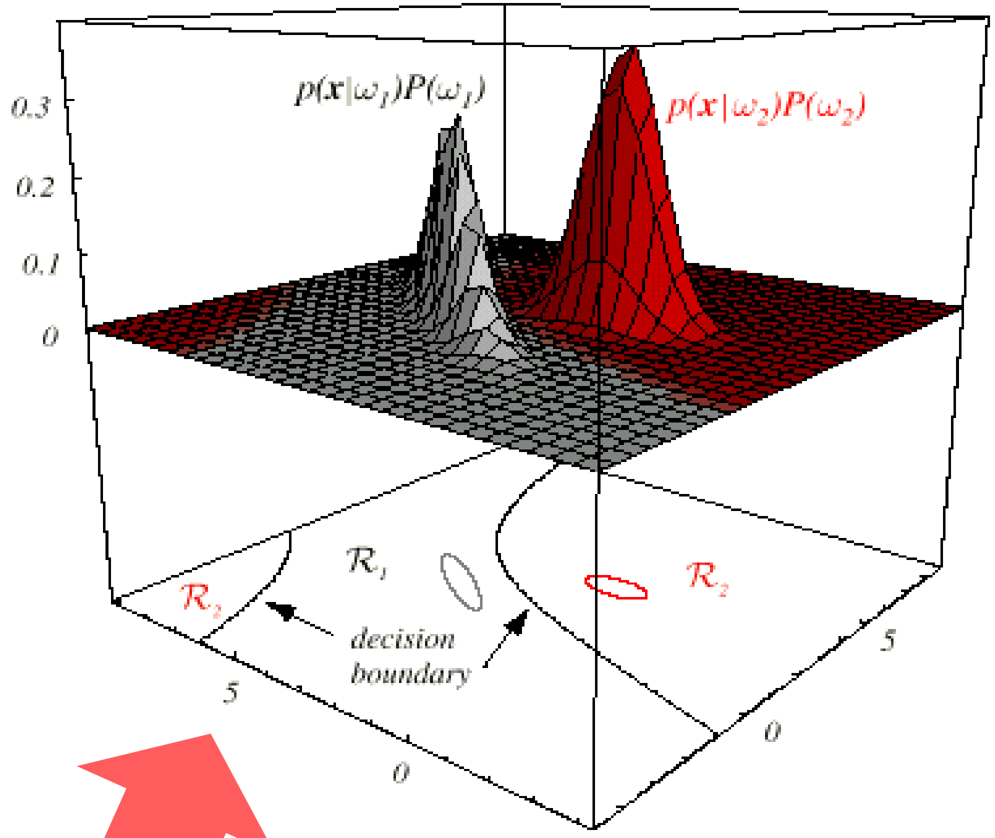
Classifiers, discriminating functions and separation surfaces (3)

- The effect of each decision is to *divide the feature space* into c ***separation or decision surfaces*** $R_1, \ldots, R_c$
  - The regions are separated with ***decision boundaries***, lines described by the maximum discriminating functions.
  - In the case of *two categories* we have two discriminating functions, g_1 e g_2, for which assign $\mathbf{x}$ to ω_1 if $g_1 > g_2$ or $g_1 - g_2 > 0$
  - Using

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

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

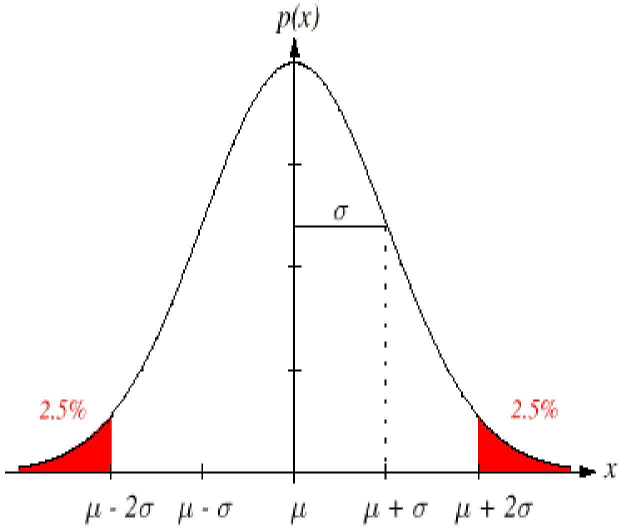
$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$



Just one discriminating function!

The Normal density

- The structure of a Bayes classifier is determined by:
  - Conditional densities $p(\mathbf{x} \mid \omega_i)$
  - A-priori probabilities $P(\omega_i)$
- One of the most important densities is the **Normal density** or **Gaussian multivariate**, in fact:
  - is analytically manageable;
  - most important, provides the best modeling of both theoretical and practical problems
    - o the Central Limit theorem asserts that "under various conditions, the distribution of the sum of *d* independent random variables tends to a particular limit known as the Normal distribution”.



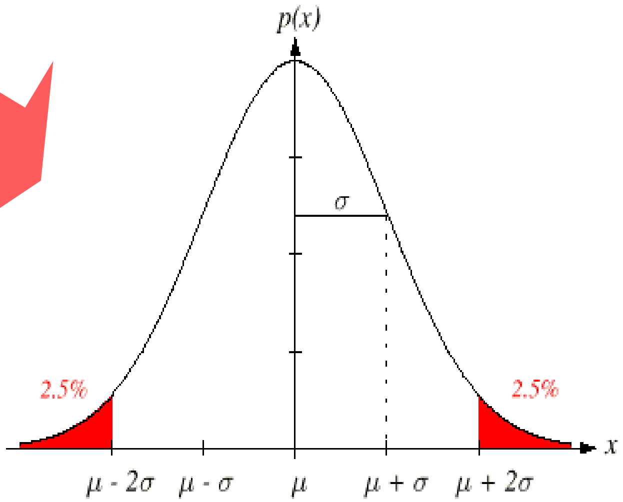
The Normal density (2)

- The Gaussian function has other properties
  - The Fourier transform of a Gaussian function is a Gaussian function;
  - It is optimal for localization over time or frequency
    - The uncertainty principle states that localization cannot occur simultaneously in time and frequency

The univariate Normal density

- Let's start with the univariate Normal density. It is fully specified by two parameters, *mean* μ and variance σ^2, it is indicated by $N(\mu, \sigma^2)$ in the form:

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



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

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

- With fixed mean and variance, the Normal density is the one with maximum entropy
  - Entropy measures the uncertainty of a distribution or the amount of information needed on average to describe the associated random variable, and is given by

$$H(p(x)) = -\int p(x) \ln p(x)\, dx$$

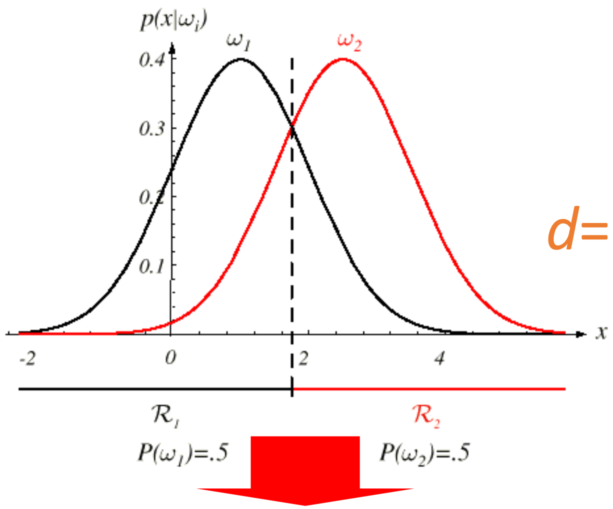
Multivariate Normal density

- The generic multivariate Normal density at *d* dimensions is:

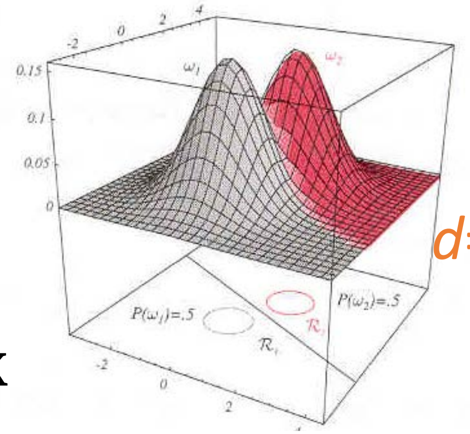
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\}$$

where

- $\boldsymbol{\mu}$ = ***mean*** vector of *d* components
- Σ = **covariance** matrix *d x d*, where
    - $|\Sigma|$ = matrix determinant
    - Σ^{-1} = inverse matrix



d=1



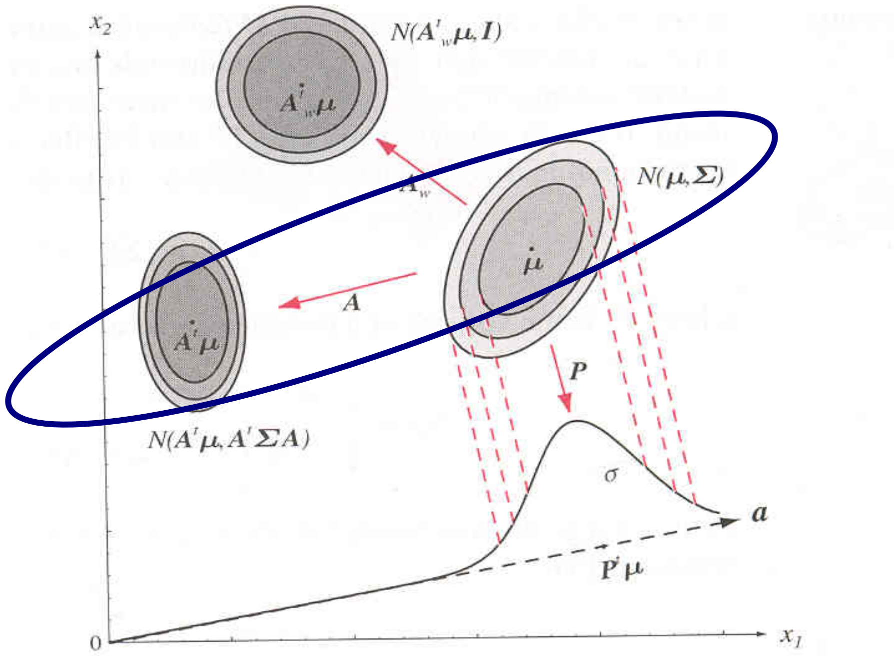
d=2

- o *Analytically* $\Sigma = \mathrm{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\, p(\mathbf{x})d\mathbf{x}$
- o *Item by element* $\sigma_{ij} = \mathrm{E}\left[(x_i-\mu_i)(x_j-\mu_j)\right]$

Multivariate Normal density (2)

- Characteristics of the covariance matrix
  - Symmetric
  - Semi-defined positive ($|\Sigma| \geq 0$)
  - σ_{ii}= variance of x_i (= σ_i^2)
  - σ_{ij}= covariance between x_i and x_j (if x_i and x_j are **statistically independent** σ_{ij}= 0)
  - If $\sigma_{ij} = 0 \quad \forall i \neq j$ $p(\mathbf{x})$ is the product of the univariate density for $\mathbf{x}$ component by component.
  - If
    - $p(\mathbf{x}) \approx N(\boldsymbol{\mu}, \Sigma)$
    - A matrix d x k
    - $\mathbf{y} = \mathrm{A}^t \mathbf{x}$

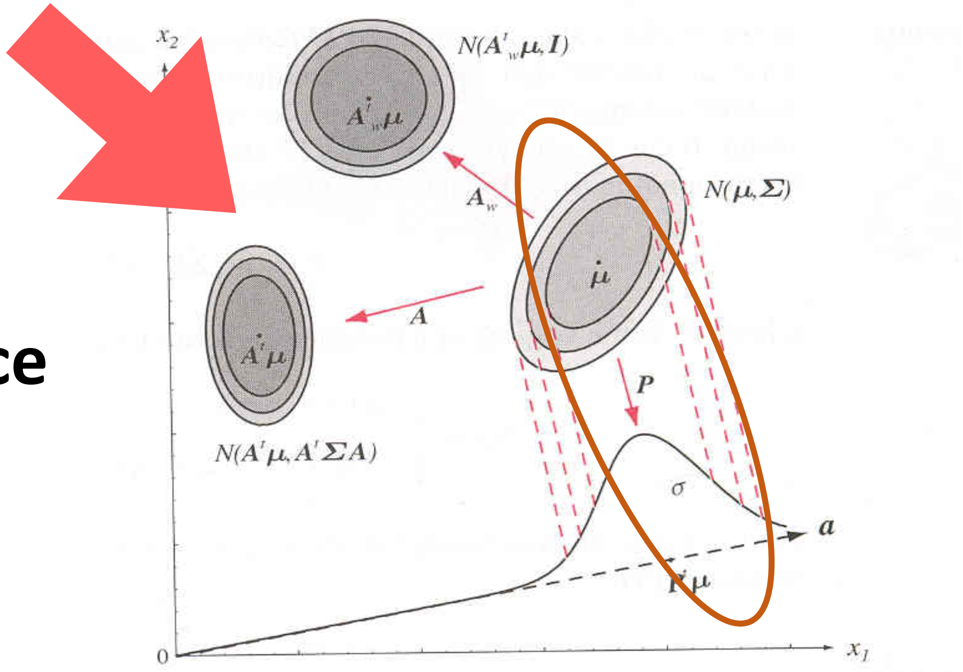
$$\rightarrow p(\mathbf{y}) \approx N(A^t\boldsymbol{\mu}, A^t\Sigma A)$$



Multivariate Normal density (3)

- SPECIAL CASE: $k = 1$
  - $p(\mathbf{x}) \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\boldsymbol{a}$ vector d x 1 of unit length
  - $y = \boldsymbol{a}^t \mathbf{x}$
  - y is a scalar that represents the projection of $\mathbf{x}$ on a line in the direction defined by $\boldsymbol{a}$
  - $\boldsymbol{a}^t \Sigma\, \boldsymbol{a}$ is the variance of $\mathbf{x}$ on $\boldsymbol{a}$

- **Generally, Σ allows you to calculate the *dispersion* of data in each surface or subspace.**

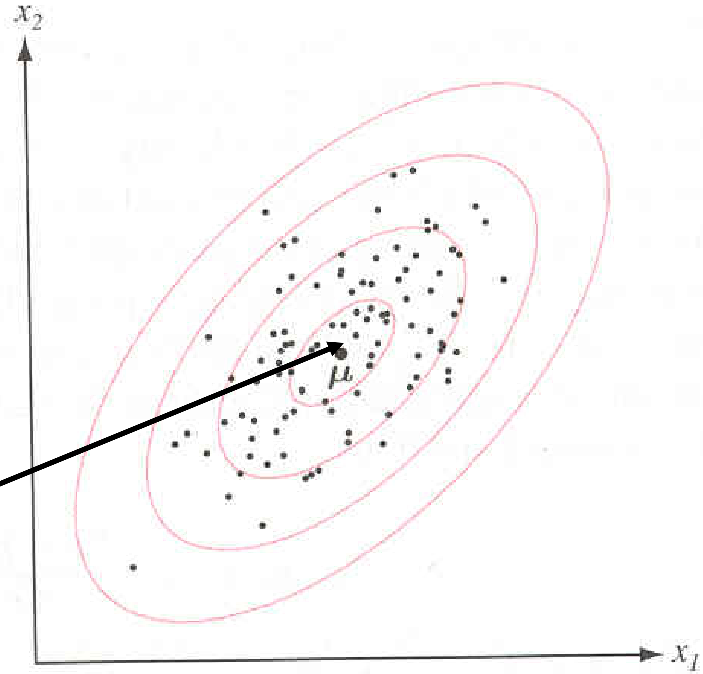


Multivariate Normal density (4)

- Let's define the *whitening transform*, being:
  - Φ the matrix of orthonormal eigenvectors of Σ in column
  - Λ the diagonal matrix of the corresponding eigenvalues
- The transformation $A_w = \Phi\Lambda^{-1/2}$, applied to feature space coordinates, provides a covariance matrix distribution = I (identity matrix)
- The *d-dimensional* density $N(\mu, \Sigma)$ needs $d + d(d+1)/2$ parameters to be defined

But what Φ and Λ represent graphically?

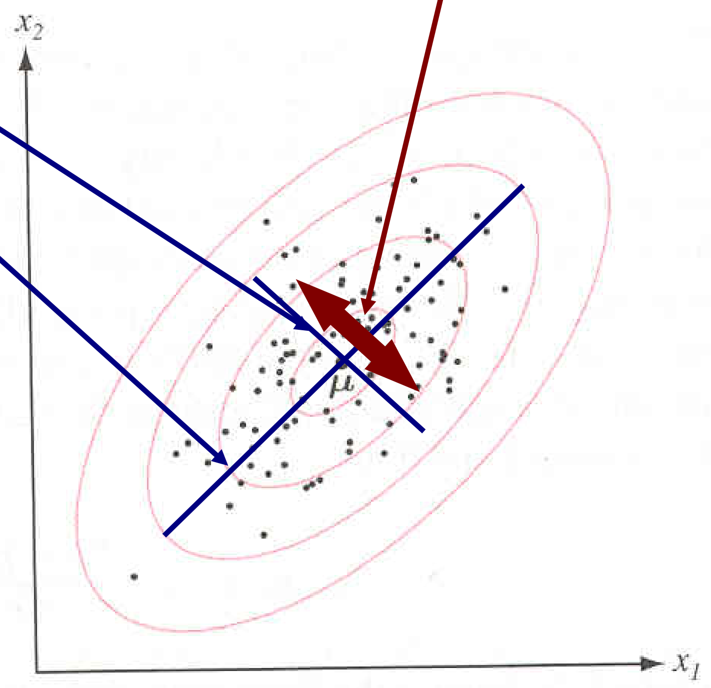
Mean identified by the coordinates of μ



Multivariate Normal density (5)

The lengths of the main axes of the hyper-ellipsoids are given by the eigenvalues of $\mathbf{\Sigma}$ (described by $\mathbf{\Lambda}$)

The main axes of the hyper-ellipsoids are given by the eigenvectors of $\mathbf{\Sigma}$ (described by $\mathbf{\Phi}$)



Hyper-ellipsoids are those places of the points for which the distance of $\mathbf{x}$ from $\mathbf{m}$

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

also called ***Mahalanobis* distance**, is constant

Discriminating Functions - Normal Density

- Back to the Bayesian classifiers, and in particular to the discriminating functions, we analyze the discriminating function as it translates into the case of Normal density and *minimum error rate*

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

$$\Downarrow$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

- Depending on the nature of Σ, the formula above can be simplified. Let's see some examples ...

Discriminating Functions - Normal Density $\Sigma_i=\sigma^2 I$

- This is the simplest case where features are statistically independent ($\sigma_{ij}= 0$, $i \neq j$), and <u>each class has the same variance</u> *(1-D case):*

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i\right] + \ln P(\omega_i)$$

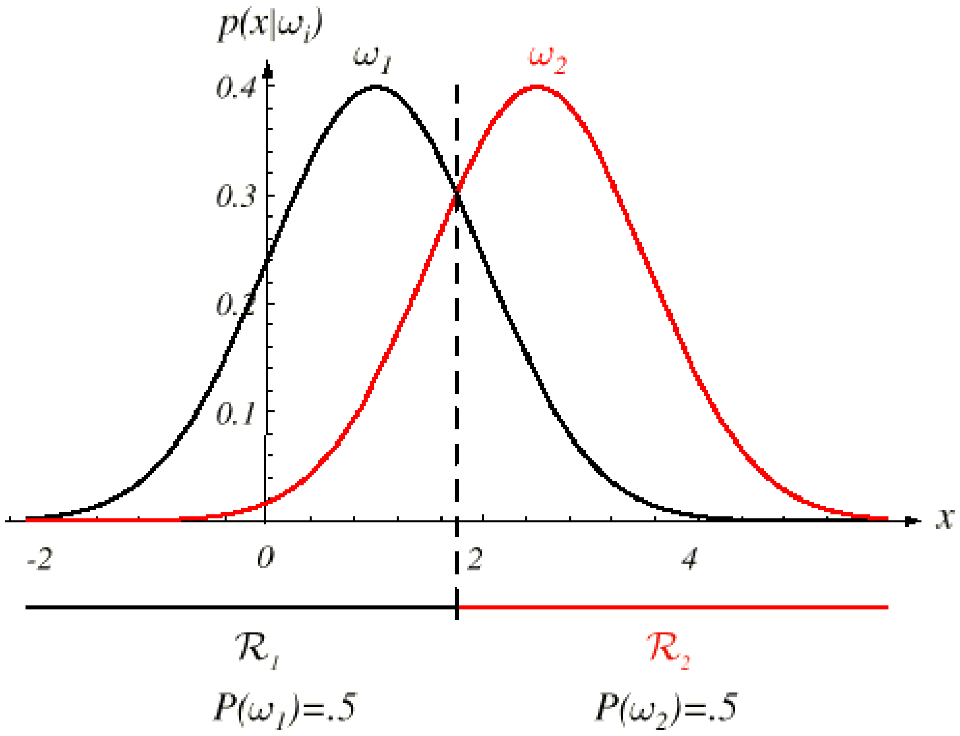
where the term $\mathbf{x}^t\mathbf{x}$, equal for every $\mathbf{x}$, can be ignored, leading to the form:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0},$$

where

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \quad \text{e} \quad w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$$

= THRESHOLD for the i-th class



Discriminating Functions - Normal Density $\Sigma_i=\sigma^2 I$ (2)

- The above functions are called ***linear discriminant functions*** (or ***linear machines***)
  - The **decision boundaries** are given by $g_i(\mathbf{x}) = g_j(\mathbf{x})$ for the two classes with the highest probability a posteriori
  - In this particular case, we have:

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$

where

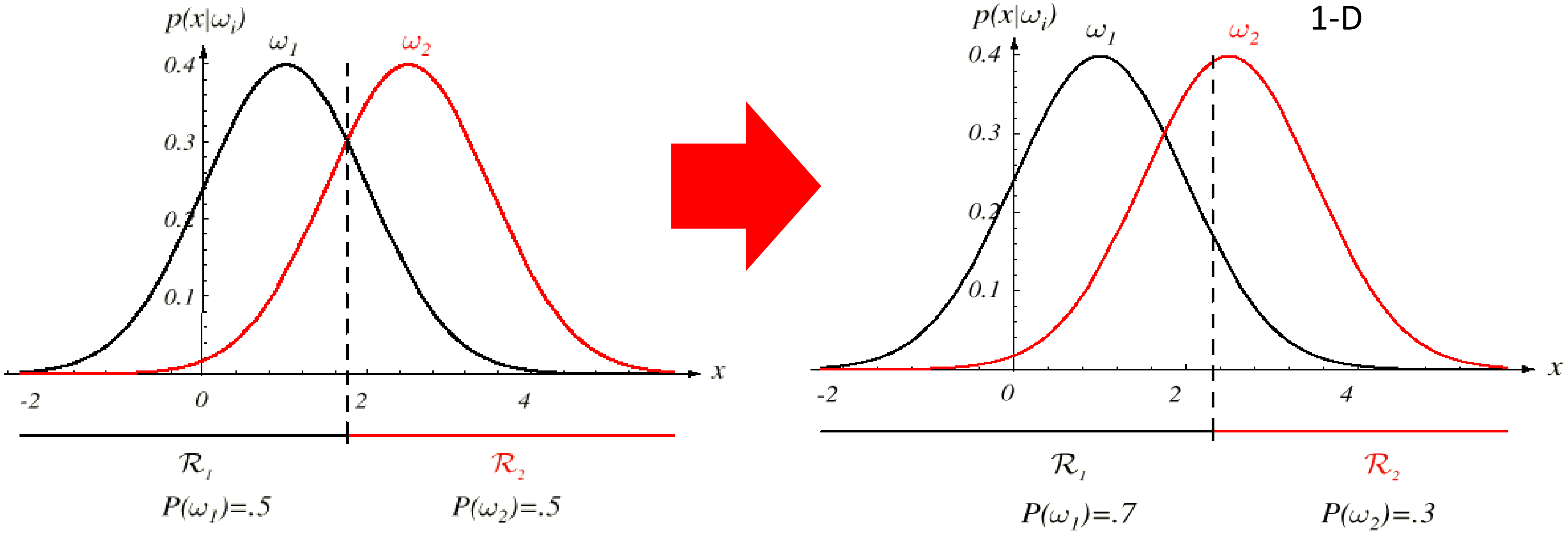
$$\mathbf{w}=\boldsymbol{\mu}_i-\boldsymbol{\mu}_j$$

$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\sigma^2}{\left\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

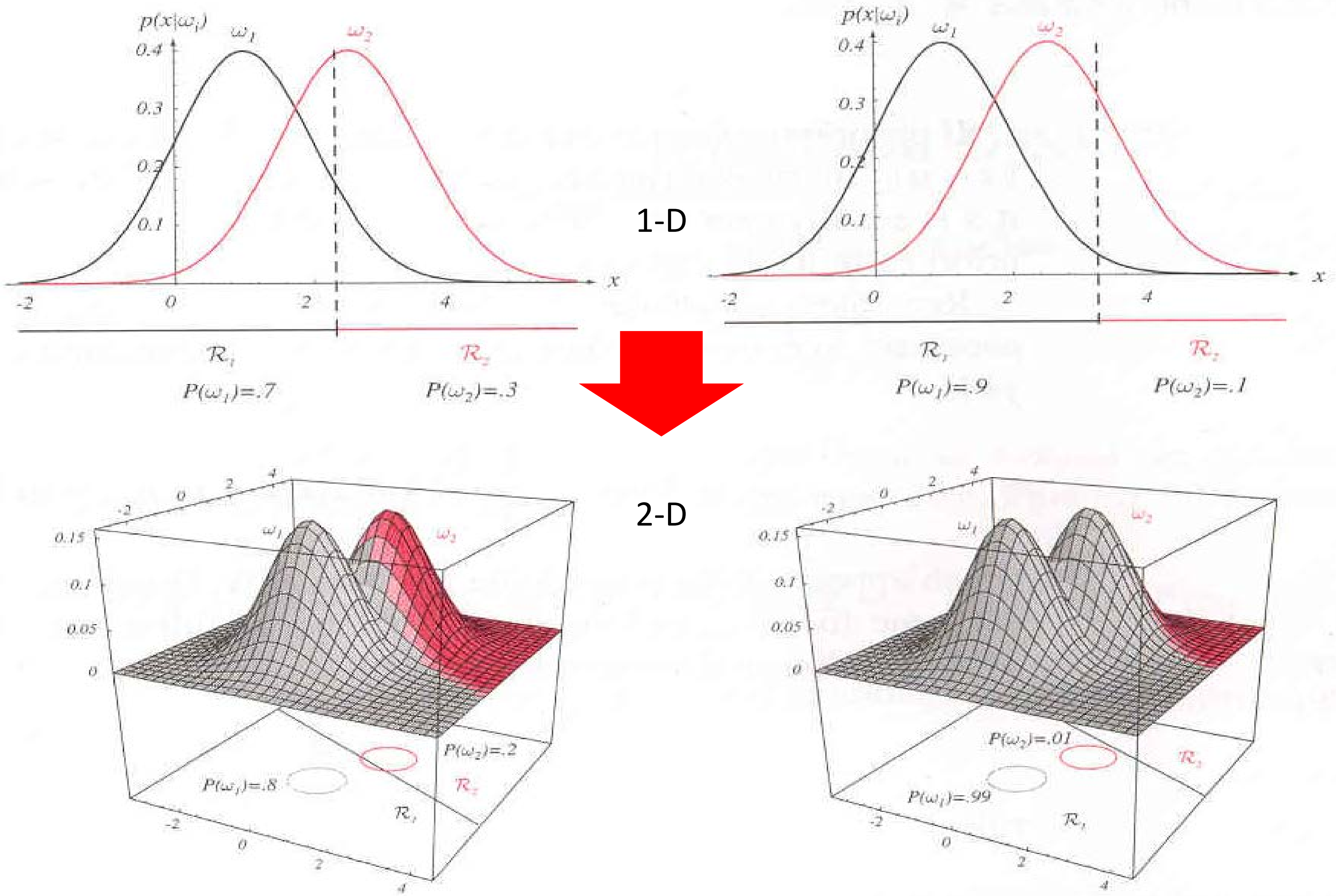
NB: if $\sigma^2 << \left\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right\|^2$ the location of the decision boundary is insensitive to the priors!

Discriminating Functions - Normal Density $\Sigma_i=\sigma^2 I$ (3)

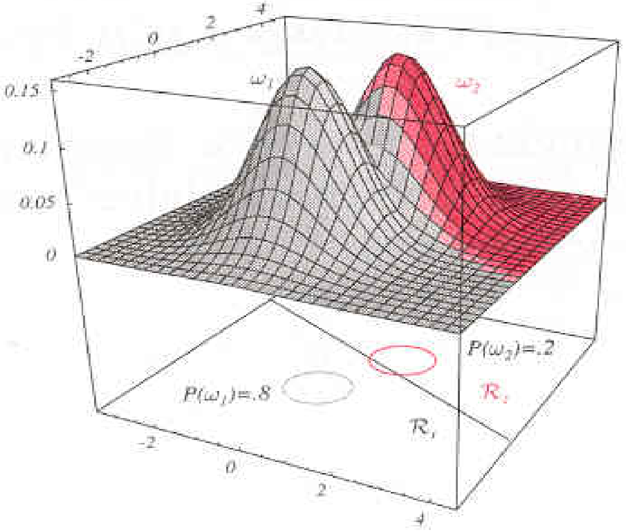
- Linear discriminant functions define a hyper-plane through $\mathbf{x}_0$ and orthogonal to $\mathbf{w}$:
  - o since $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$, the hyperplane that separates R_i from R_j is *orthogonal* to the line joining the means.
- From the previous formula, it can be noted that, with the same variance, the larger prior determines the classification result

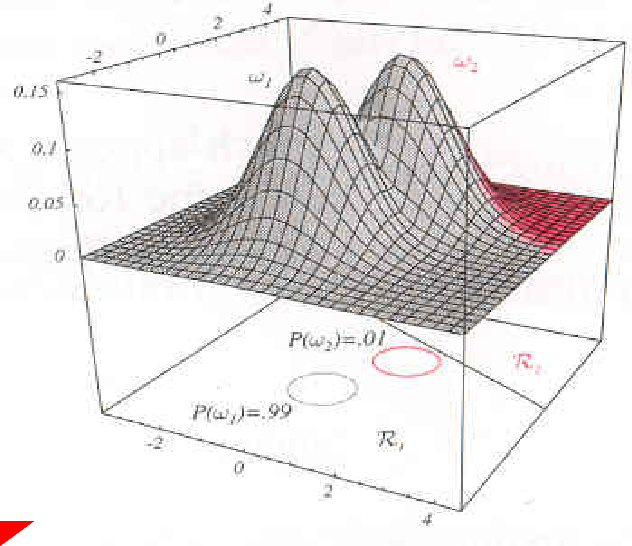


Discriminating Functions - Normal Density $\Sigma_i=\sigma^2 I$ (4)



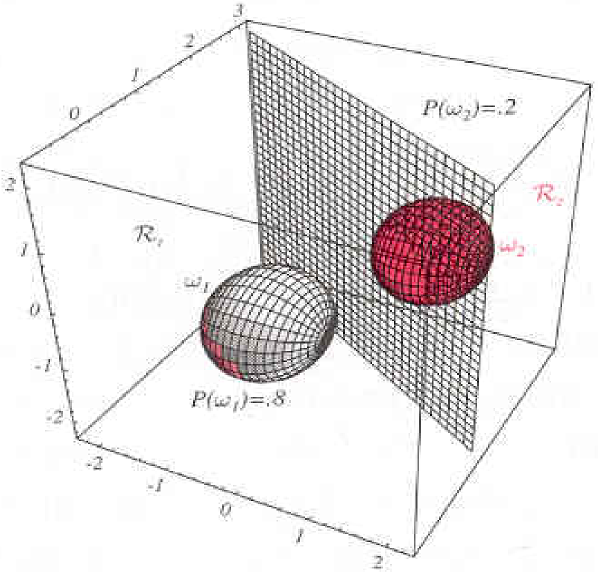
Discriminating Functions - Normal Density $\Sigma_i = \sigma^2 I$ (5)

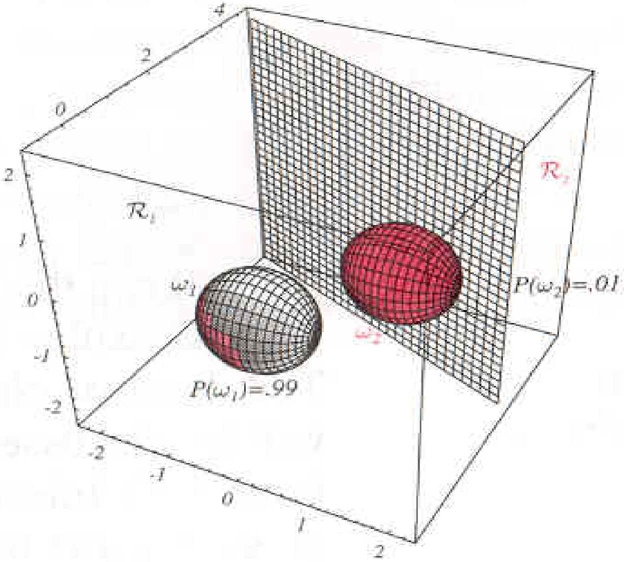




2-D

3-D





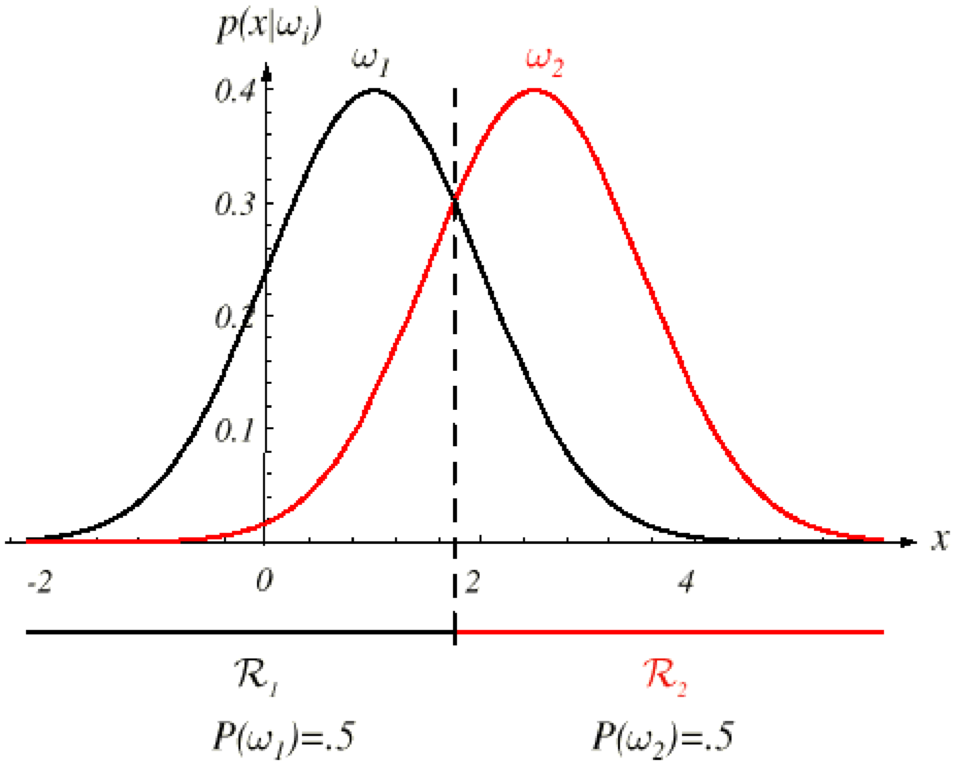
Discriminating Functions - Normal Density $\Sigma_i=\sigma^2 I$ (6)

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

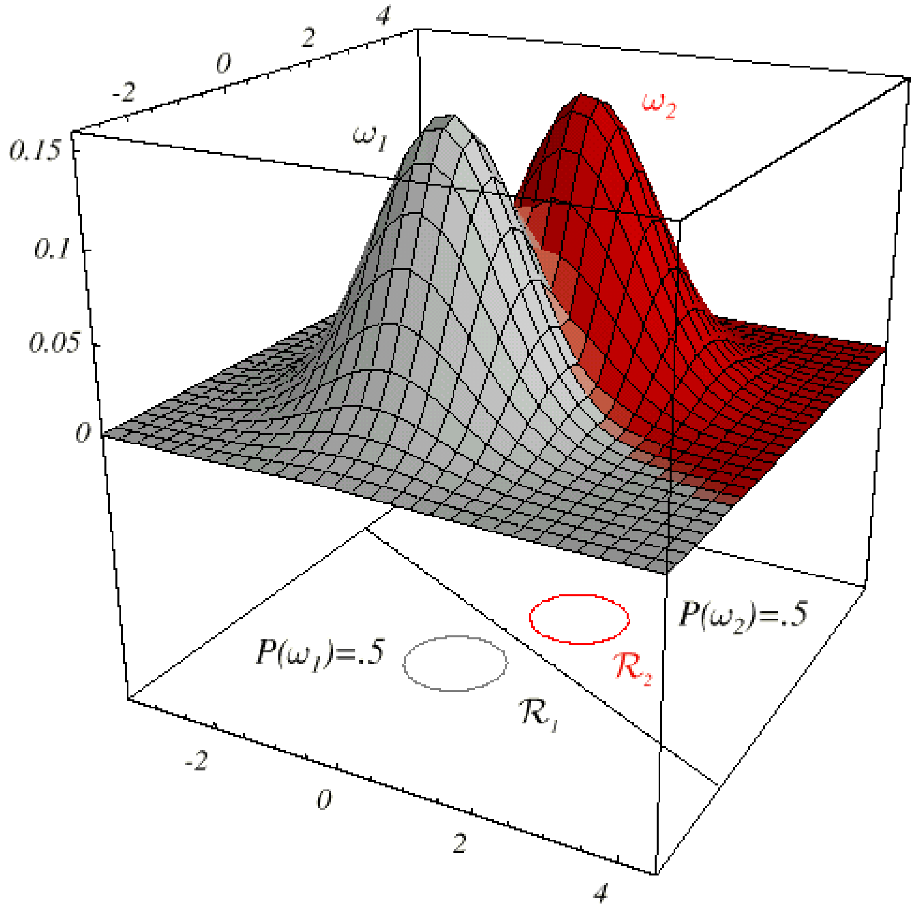
- PLEASE NOTE: If the prior probabilities $P(\omega_i)$, $i=1,\ldots,c$ are *equal*, then the term with the logarithm is equal to 0 and can be ignored, reducing the classifier to a **minimum distance classifier**.
- In practice, the optimal decision rule has a simple geometric interpretation
  - Assigns $\mathbf{x}$ to the class whose mean $\boldsymbol{\mu}$ is closer

Discriminating Functions - Normal Density $\Sigma_i=\sigma^2 I$ (7)

1-D

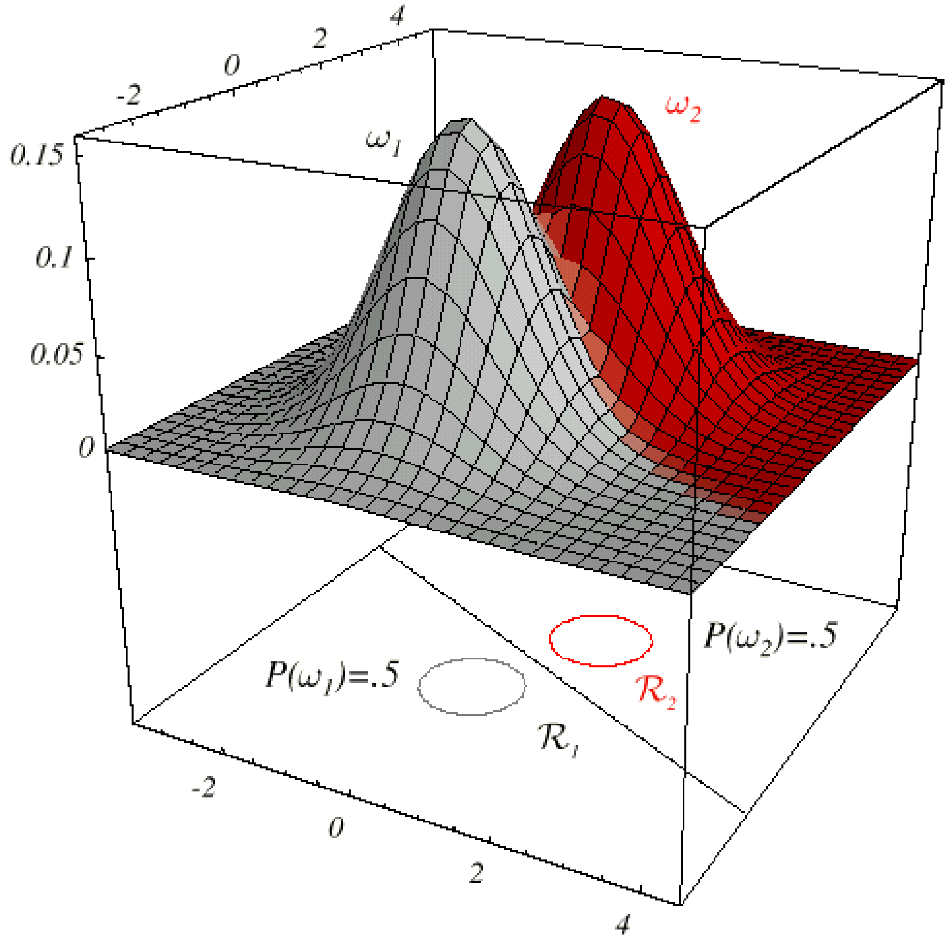


2-D

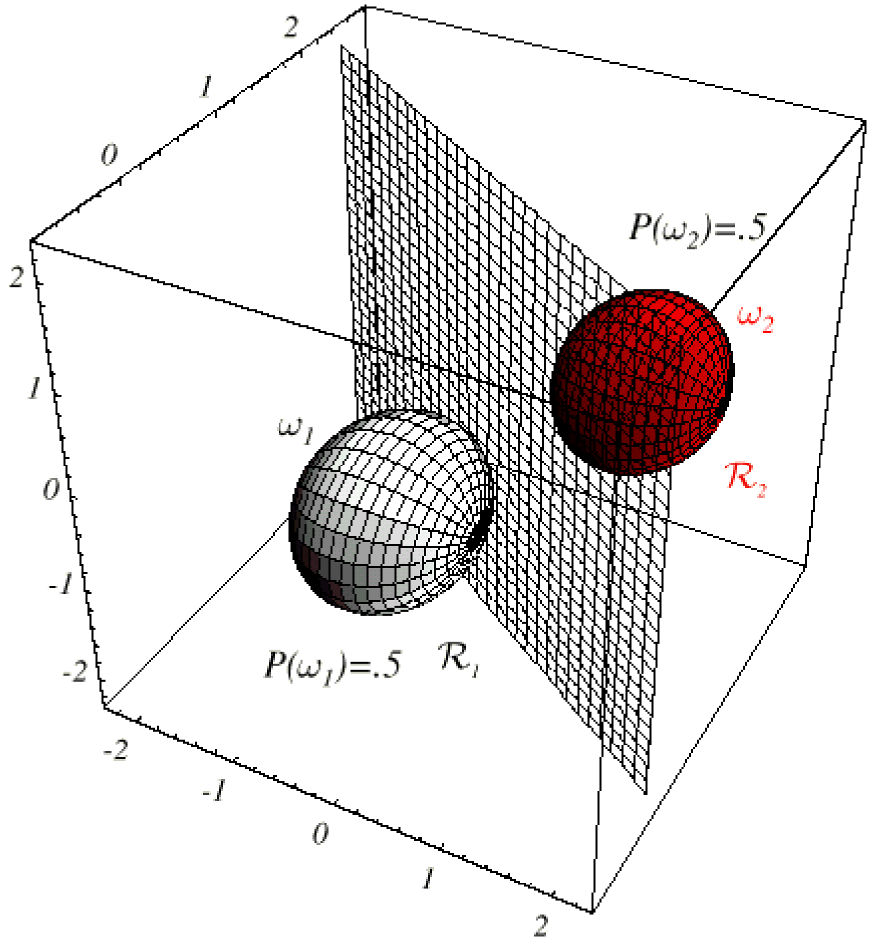


Discriminating Functions - Normal Density $\Sigma_i=\sigma^2 I$ (8)

2-D



3-D



Discriminating Functions - Normal Density $\Sigma_i = \Sigma$

- Another simple case occurs when the covariance matrices for all classes are equal, but arbitrary.
- In this case, the ordinary formula:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

can be simplified to

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

which is further manageable, with a process similar to the previous case (developing the product and eliminating the term $\mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}$).

Discriminating Functions - Normal Density $\Sigma_i = \Sigma$ (2)

- This, we still obtain linear discriminating functions, in the form:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

where

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

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

- Since the discriminants are linear, the *decision boundaries* are still hyper-planes.

Discriminating Functions - Normal Density $\Sigma_i = \Sigma$ (3)

If the decision-making regions R_i and R_j are contiguous, the boundary between them becomes:

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0,$$

where

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

and

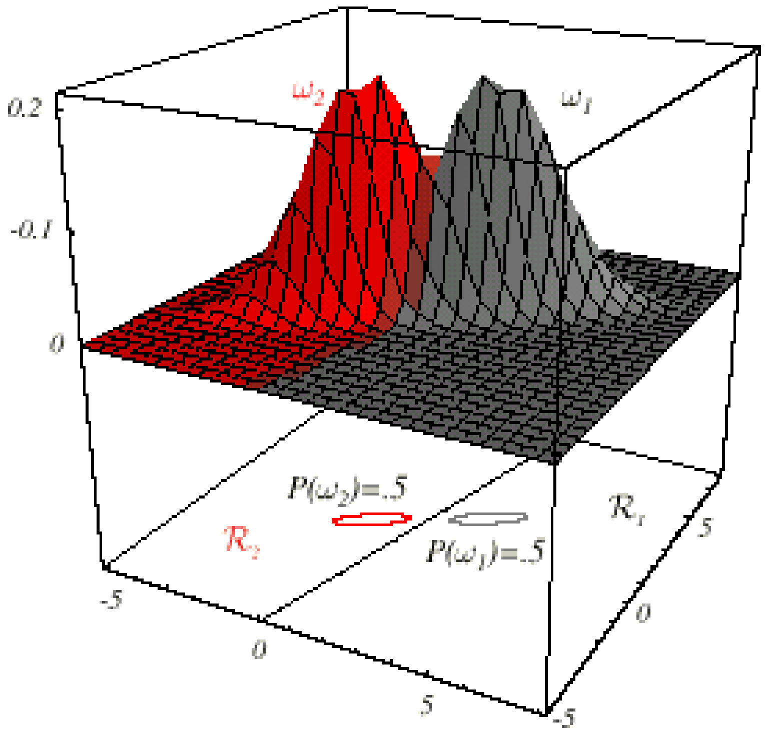
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

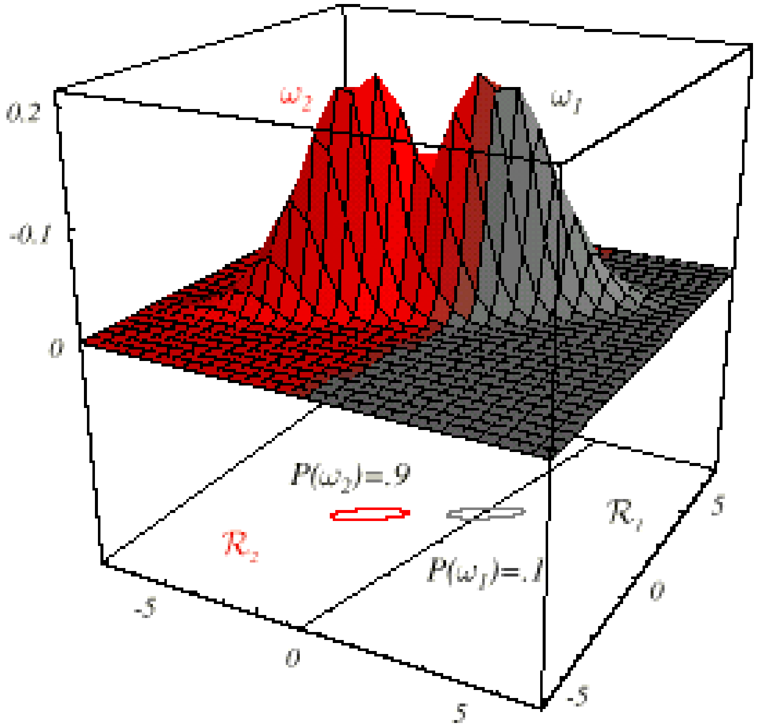
Discriminating Functions - Normal Density $\Sigma_i = \Sigma$ (4)

- Since $\mathbf{w}$ in general (differently from before) it is not the vector that joins the 2 means ($\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$), the hyperplane that divides R_i from R_j it is not orthogonal to the line between the means. However, it intersects this line in $\mathbf{x}_0$

- If the *priors* are equal, then $\mathbf{x}_0$ is in the middle of the means, otherwise the optimal hyper-plane of separation will be shifted towards the average of the least likely class.

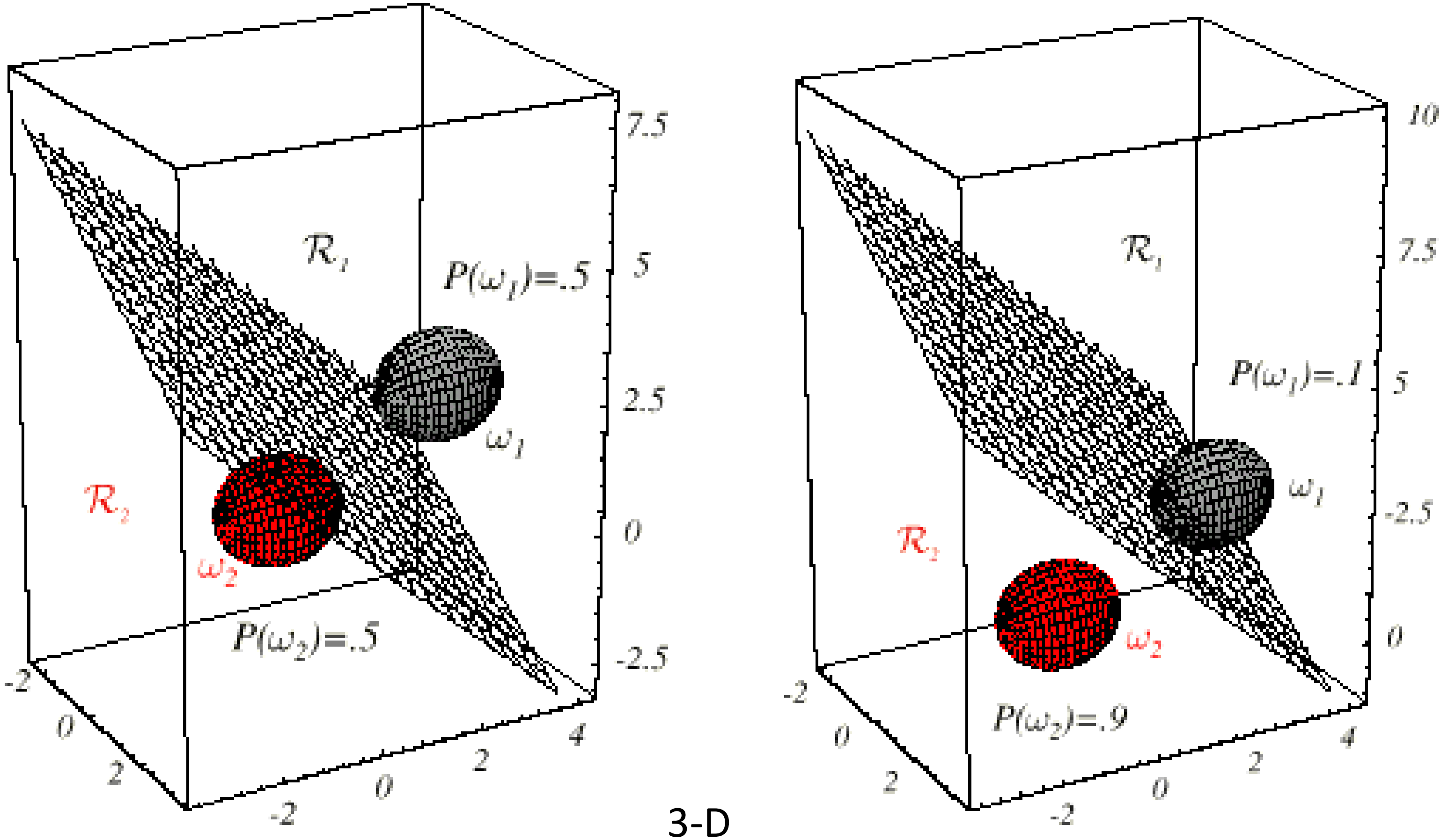
Discriminating Functions - Normal Density $\Sigma_i = \Sigma$ (5)





2-D

Discriminating Functions - Normal Density $\Sigma_i = \Sigma$ (6)



3-D

Discriminating Functions - Normal Density Σ_i arbitrary

- Covariance matrices are different for each category;
- Discriminating functions are inherently quadratic;

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0},$$

where

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

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

and

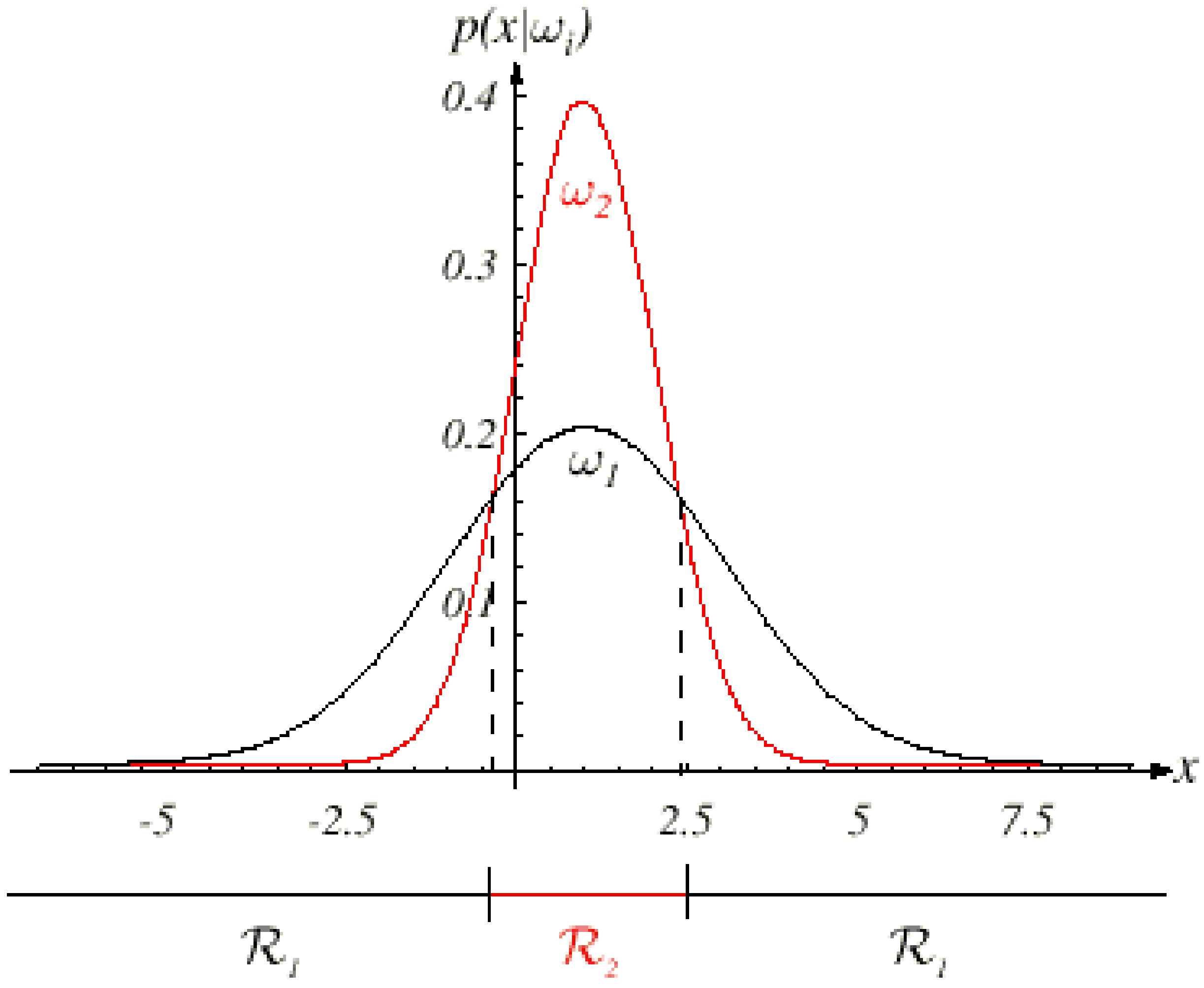
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$$

Discriminating Functions - Normal Density Σ_i arbitrary (2)

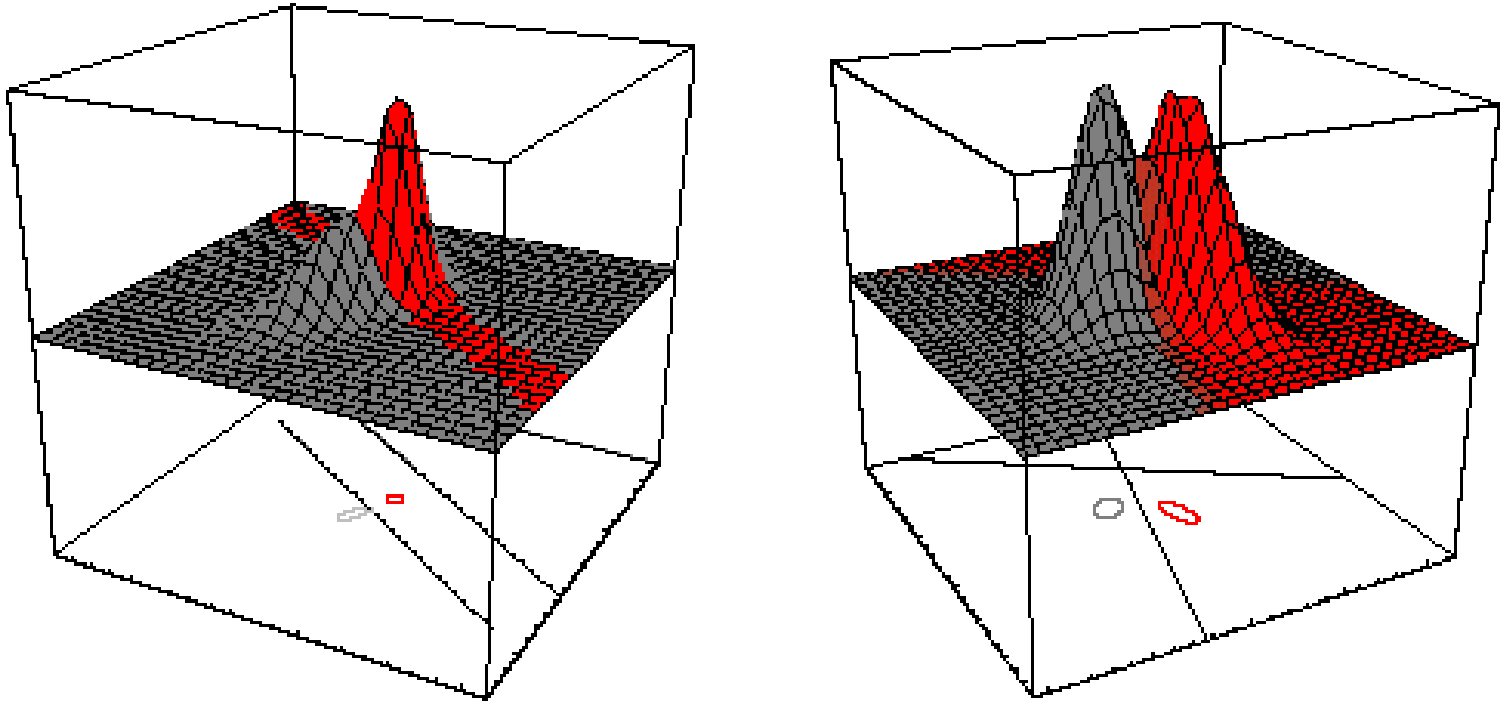
- In case 2-D the decision surfaces are hyper-quadric:
  – Hyper-planes
  – Pair of hyper-planes
  – Hyper-spheres
  – Hyper-paraboloids
  – Hyperboloids of various types

- Even in the 1-D case, due to arbitrary variance, the decision regions are usually unconnected.

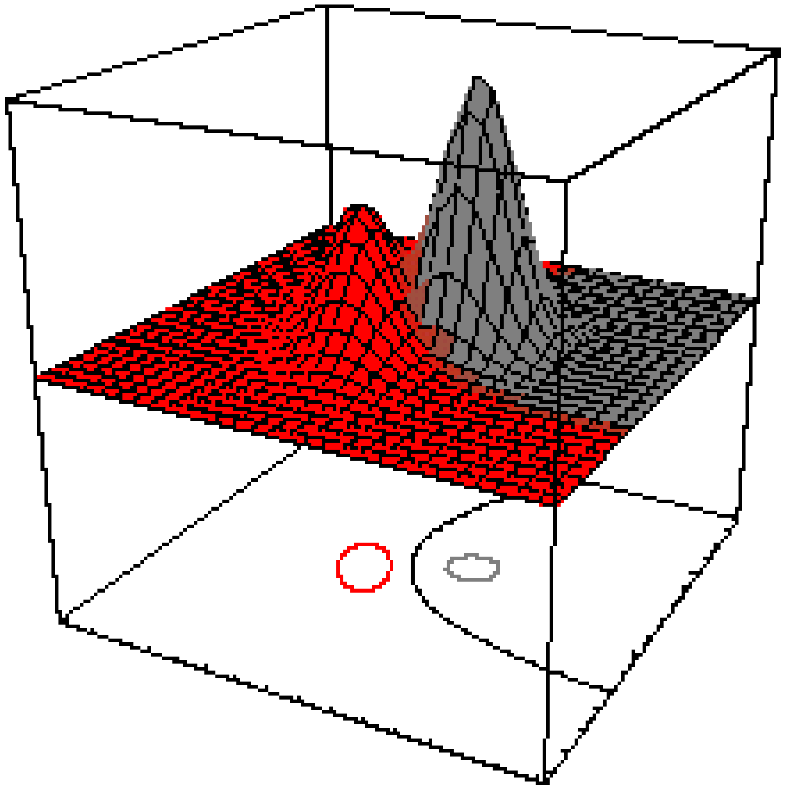
Discriminating Functions - Normal Density Σ_i arbitrary (3)

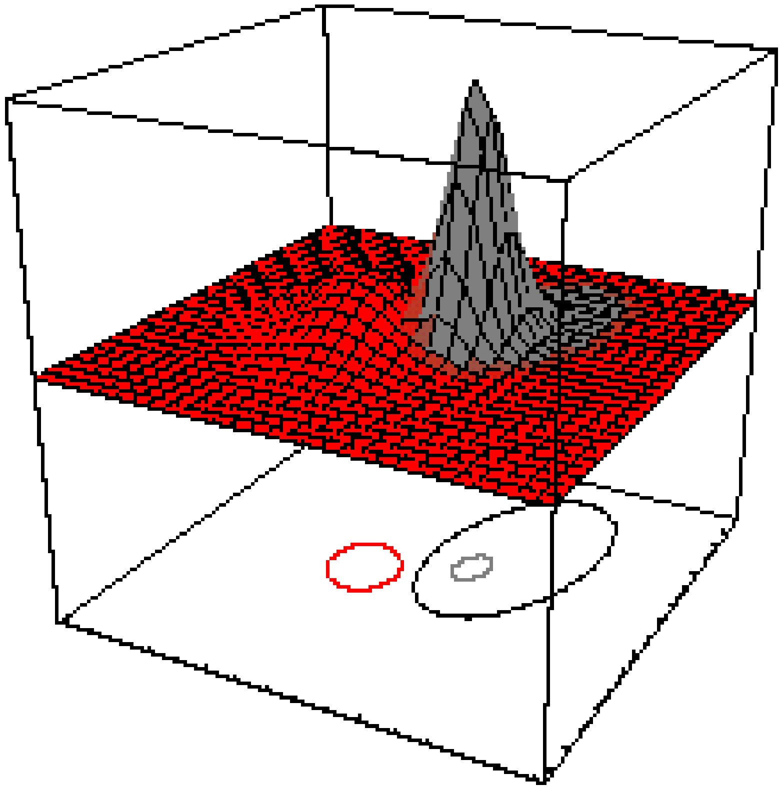


Discriminating Functions – Normal Density Σ_i arbitrary (4)

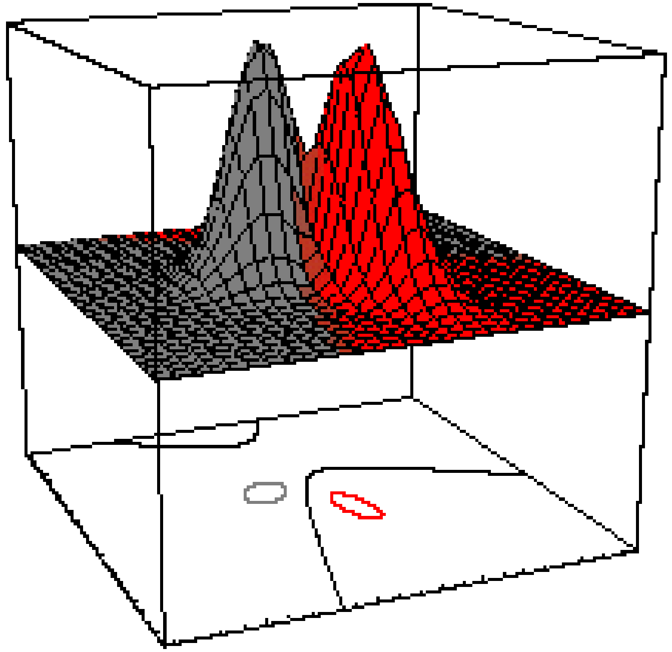


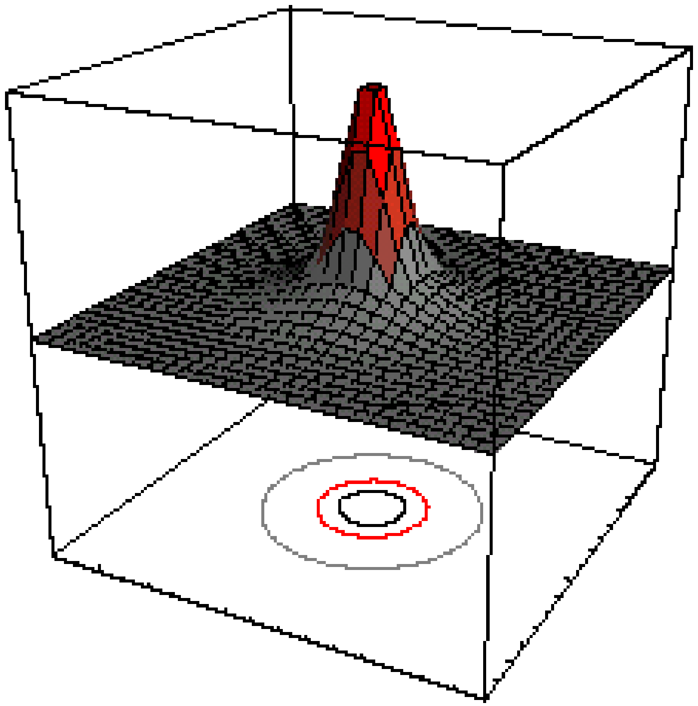
Discriminating Functions – Normal Density Σ_i arbitrary (5)



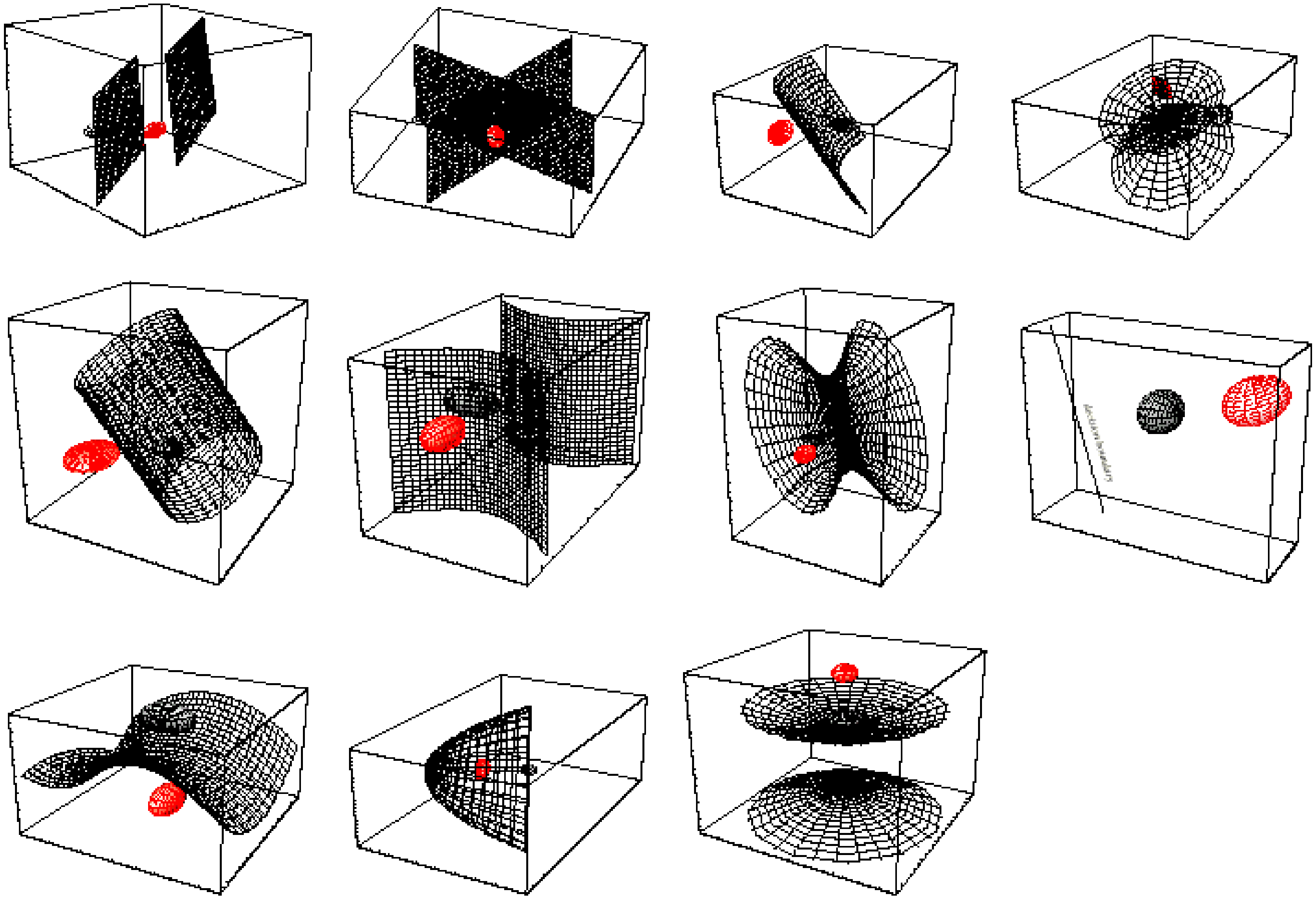


Discriminating Functions – Normal Density Σ_i arbitrary (6)





Discriminating Functions – Normal Density Σ_i arbitrary (7)



Discriminating Functions – Normal Density Σ_i arbitrary (8)