Part 2

Bayesian Decision Theory and Gaussian Pattern Classifiers

Introduction

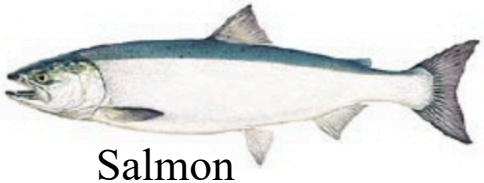
- Statistical machine learning is grounded into **Bayesian decision theory**, therefore, knowing the elements of this theory is a must for anybody wants to work in this field.
- Bayesian theory assumes that decision making problems are formulated in probabilistic terms. Probabilities are known or should be estimated.
- Bayes decision theory allows to take into account both *probability* and *"risk"* of decisions. Making a rational decision means to take into account both the probability and the risk (or the utility) associated to the decision.

- We present elements of Bayesian decision theory assuming that all the probabilities involved in the problem considered are known.

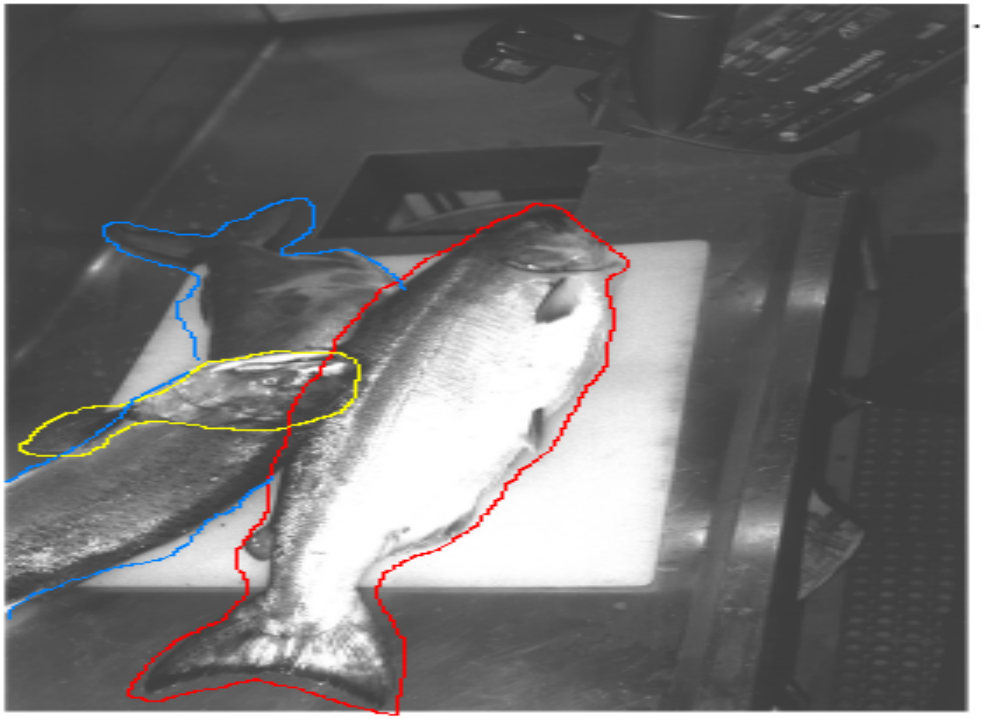
First thing to know: the MAP decision rule

➢We present the fundamental decision rule of Bayesian decision theory using the example of the **salmon/sea bass classification** introduced in Part 1.

Sea bass



Salmon



Let us assume that an image segmentation module has already extracted the shape of the fishes as shown in the figure, and a feature extraction module has characterized each pattern with **one feature**: the **length** of the shape. Decision problem: we want to assign each shape/pattern to one of the two classes considered (salmon, sea bass).

First thing to know: the MAP decision rule

- We assume that we cannot know deterministically which is the "class" (salmon or sea bass) of the next fish that we see. So the problem must be formulated in probabilistic terms.
- Next incoming fish can be a salmon or a sea bass with a given probability. Bayes decision theory formalizes this situation with the concept of "state of nature" (usually called "class" in pattern recognition). In our example, we have two states-of-nature/classes: ω_1 and ω_2
- Let $\omega = \omega_1$ or $\omega = \omega_2$ be the variable that identifies the class, where ω is a *random variable*.

➢ The two classes could have the same *prior probability*:

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

$$P(\omega_1) + P(\omega_2) = 1 \text{ (we have just two species of fish)}$$

The MAP decision rule

•If we should make a decision without being able to see the incoming fish, the only rational decision would be:

Assign the fish to ω_1 *if* $P(\omega_1) > P(\omega_2)$*, else assign the fish to* ω_2

➢In general, we must "see" the pattern to make a rational decision according to Bayesian theory.

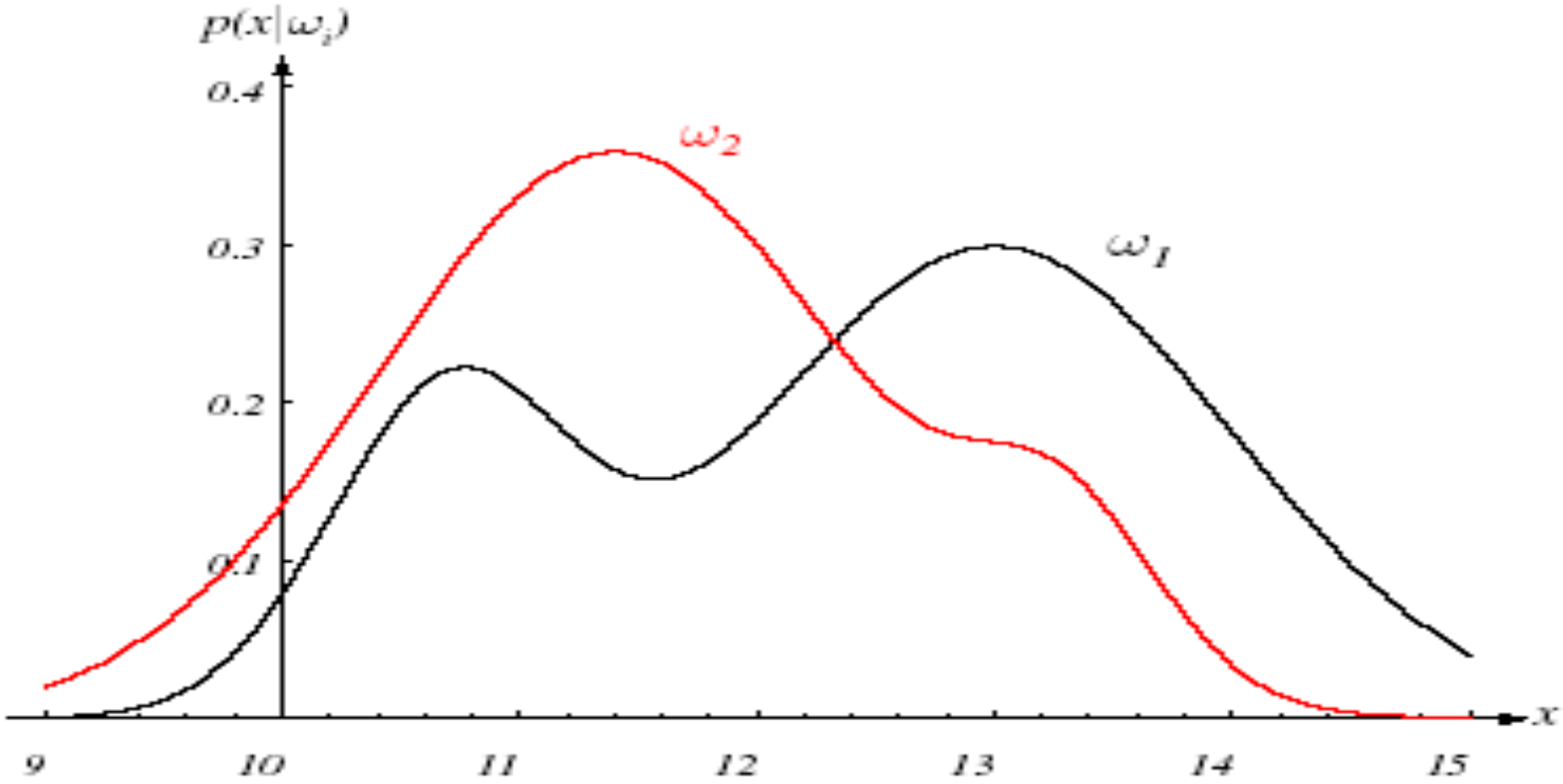
•We must see the fish and characterize it with some features.

For example, the average lightness of the pattern.

•As fishes incoming on the belt will have "random" **length** values, the length feature **x** should be treated as a random variable with *conditional distribution* $p(x \mid \omega_i)$.

An example of a mono-dimensional $p(x \mid \omega_i)$

- $p(x \mid \omega_i)$ is the *class-conditional probability density function*



If x is the average length of the image region associated to a given fish of class ω_i, then the difference between the functions $p(x \mid \omega_i)$ characterizes the expected length difference between the two fish species.

Bayes decision rule

- Let us assume to know the two priors $P(\omega_j)$ and the two class-conditional density functions $p(x \mid \omega_j)$, $j=1,2$.
- If we measure the length x of the fish, the most rationale decision rule is based on the probability:

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

- That we can rewrite as the **Bayes decision rule (MAP, maximum a posteriori, decision rule)** :

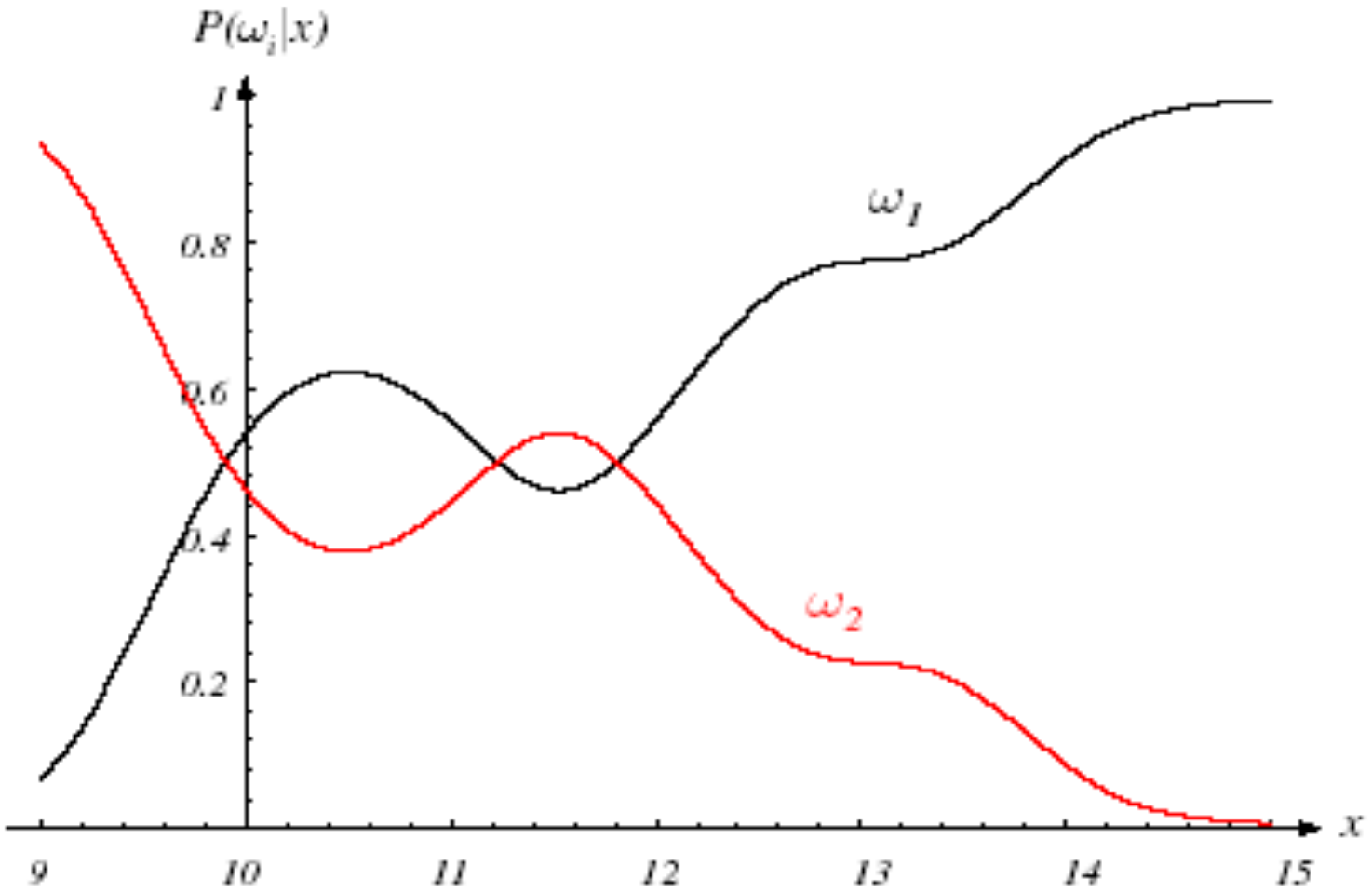
$$P(\omega_j \mid x) = p(x \mid \omega_j)\ \ P(\omega_j) / p(x)$$

*Posterior = (Likelihood * Prior) / Evidence*

Note that: $$p(x) = \sum_{j=1}^{2} p(x \mid \omega_j) P(\omega_j)$$

An example of mono-dimensional $P(\omega_i|\, x)$

$P(\omega_i|\, x)$ with $P(\omega_1)=2/3$ e $P(\omega_2)=1/3$



The MAP decision rule

- The MAP, maximum a posteriori probability, criterion is the most rationale decision rule for the considered probabilistic setting:

If $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ then is most rationale to assign x to ω_1

If $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ then is most rationale to assign x to ω_2

- This rule is the most rationale because it minimizes the error probability for any given x:

$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we assign } x \text{ to } \omega_2$$
$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we assign } x \text{ to } \omega_1$$

- We can prove that MAP rule also minimizes the average error:

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

Likelihood ratio test and ML rule

- We can reformulate the MAP rule as follows:

  If $p(x \mid \omega_1)\, P(\omega_1) > p(x \mid \omega_2)\, P(\omega_2)$ then assign x to ω_1

  else assign x to ω_2

Likelihood ratio test

$$l(x) = \frac{p(x/\omega_1)}{p(x/\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)} = \theta$$

Two special cases:

- If $p(x/\omega_1)=p(x/\omega_2)$, then the decision depends only on priors
- If $P(\omega_1)=P(\omega_2)$, then the decision depends only on likelihoods (**ML**, **Maximum Likelihood, decision rule**)

MAP decision rule with more than two classes

The MAP decision rule with more than two classes is:

$$\mathbf{x} \rightarrow \omega_i \Leftrightarrow P(\omega_i \mid \mathbf{x}) > P(\omega_j \mid \mathbf{x}) \ \forall i \neq j,\ i=1,\ldots,c$$

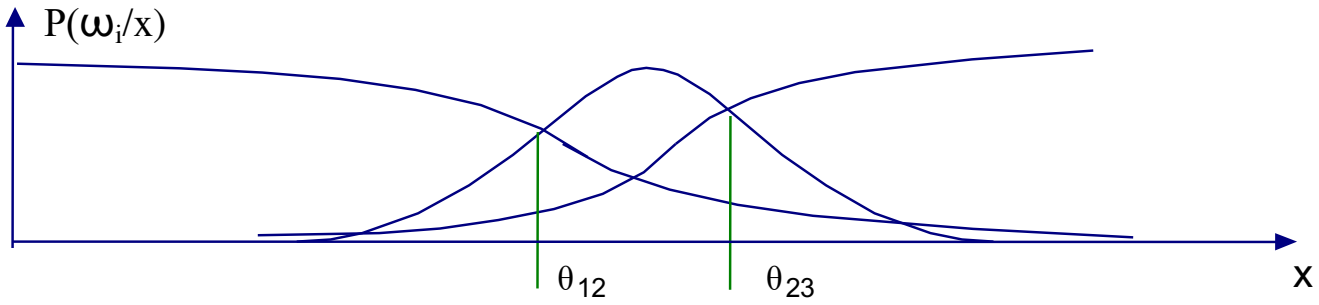
- The Likelihood ratio test is defined accordingly.

- It is easy to see that we should have multiple thresholds θ_{st} defined according to the following rule :

$$P(\omega_s \mid \mathbf{x}) > P(\omega_i \mid \mathbf{x}) \ \forall s,t \neq i,\ s \neq t \ \ i=1,\ldots,c$$

$$P(\omega_t \mid \mathbf{x}) > P(\omega_i \mid \mathbf{x})$$

- In this example, we have three classes and two thresholds θ_{12} e θ_{23}.



Basic concepts of error probability

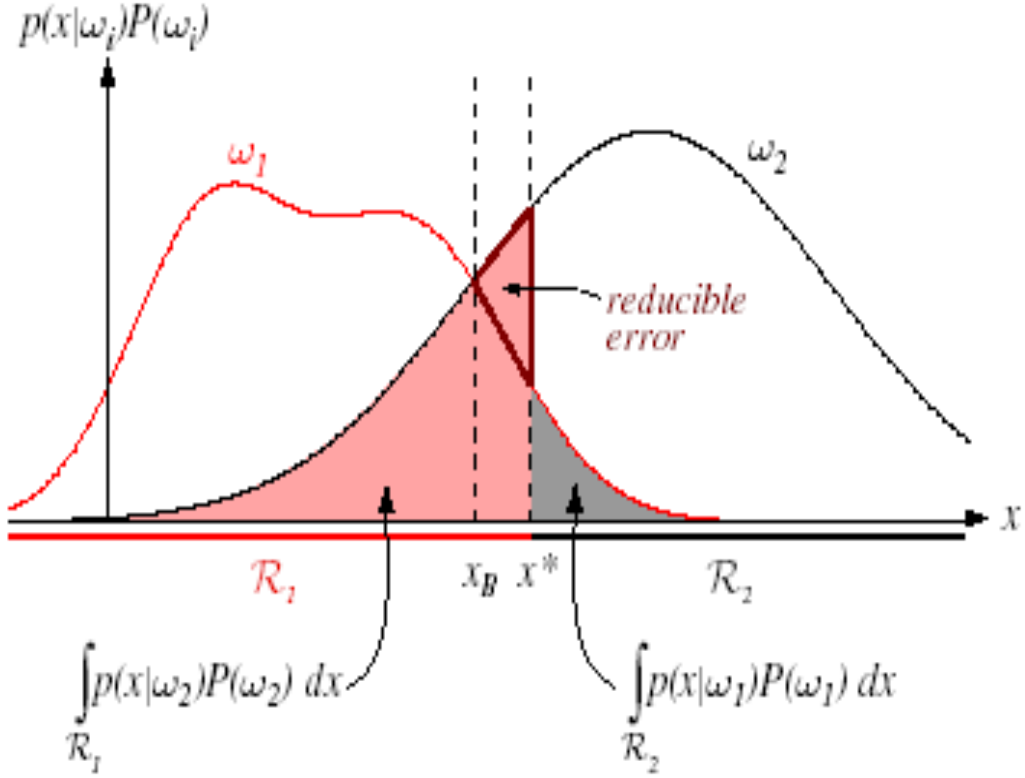
- For the two class case:

$$
\begin{aligned}
P(error) &= P\{x \in R_2, \omega_1\} + P\{x \in R_1, \omega_2\} = \\
&= P(\omega_1)P\{x \in R_2 \mid \omega_1\} + P(\omega_2)P\{x \in R_1 \mid \omega_2\} = \\
&= P(\omega_1)\int_{R_2} p(x \mid \omega_1)\,dx + P(\omega_2)\int_{R_1} p(x \mid \omega_2)\,dx
\end{aligned}
$$

The optimal threshold $x=x_B$ is the Bayesian threshold providing the minimum error, called **Bayes error**.

In the figure $x=x^*$ is a suboptimal threshold, that brings to an *added* (reducible) error over the Bayes error.

In practical cases we usually have an **added error** because the optimal threshold providing the Bayes error is nearly impossible to estimate.



Basic concepts of error probability

- With more than two classes (c>2), it is more convenient to compute the error probability by the probability of correct classification:

$$P(correct) = \sum_{i=1}^{c} P\{\mathbf{x} \in R_i, \omega_i\} = \sum_{i=1}^{c} P_i P\{\mathbf{x} \in R_i / \omega_i\} = \sum_{i=1}^{c} P_i \int_{R_i} p(\mathbf{x} \mid \omega_i) d\mathbf{x}$$

$$P(error) = 1 - P(correct)$$

In general, it is easy to see that the above computation can be very difficult, as it requires multidimensional integrals and involves density functions with very complicated analytical forms.

The computation is easy only for Gaussian probability densities.

Homework 2 - send me your answer to roli@unica.it

In a city of 1 million inhabitants there are 100 known terrorists and 999.900 non-terrorists.

The prior probability of one random inhabitant of the city being a terrorist is thus 0.0001 and the prior probability of a random inhabitant being a non-terrorist is 0.9999. In an attempt to catch the terrorists, the city installs a surveillance camera with automatic facial recognition software. The software has two failure rates of 1%:

- if the camera sees a terrorist, it will ring a bell 99% of the time, and mistakenly fail to ring it 1% of the time (in other words, the false-negative rate is 1%).
- if the camera sees a non-terrorist, it will not ring the bell 99% of the time, but it will mistakenly ring it 1% of the time (the false-positive rate is 1%).

So, the failure rate of the camera is always 1%.

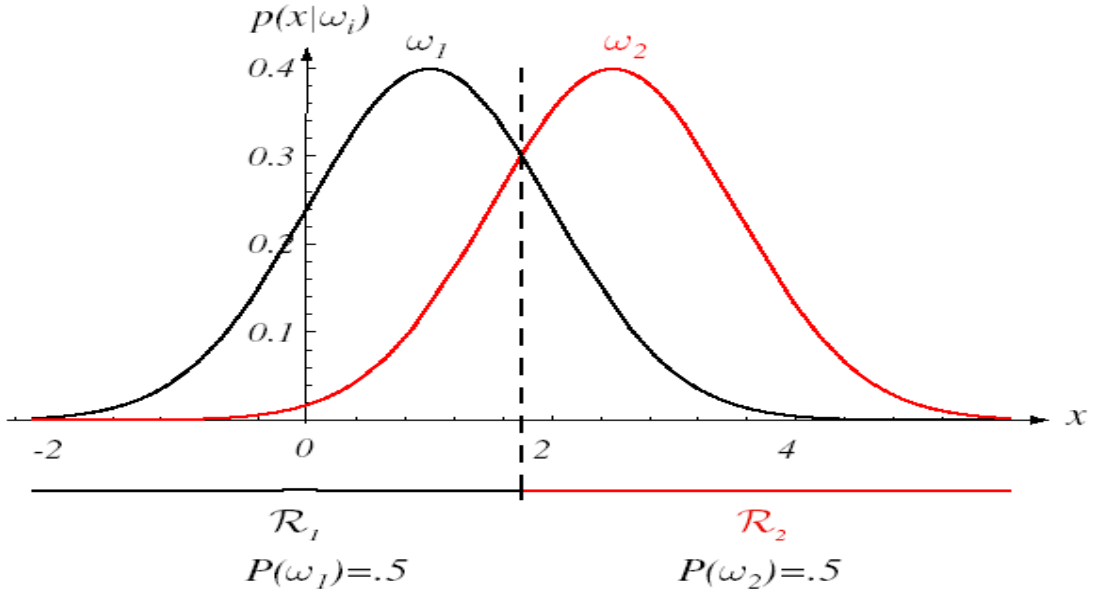
Suppose somebody triggers the alarm. What is the chance that is a terrorist?

The concept of "decision regions"

The likelihood ratio test is defined by $l(x)$ and the threshold θ. This test identifies two decision regions R_1 e R_2 in the feature space R (here we consider a single feature x).

$$l(x) = \frac{p(x/\omega_1)}{p(x/\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)} = \theta$$

➢ $R_1 = \{\mathbf{x} \in R: l(\boldsymbol{x}) > \theta\}$ and $R_2 = \{\mathbf{x} \in R: l(\boldsymbol{x}) < \theta\}$ (if $l(\boldsymbol{x}) = \theta$ then x can be assigned randomly to R_1 or R_2).



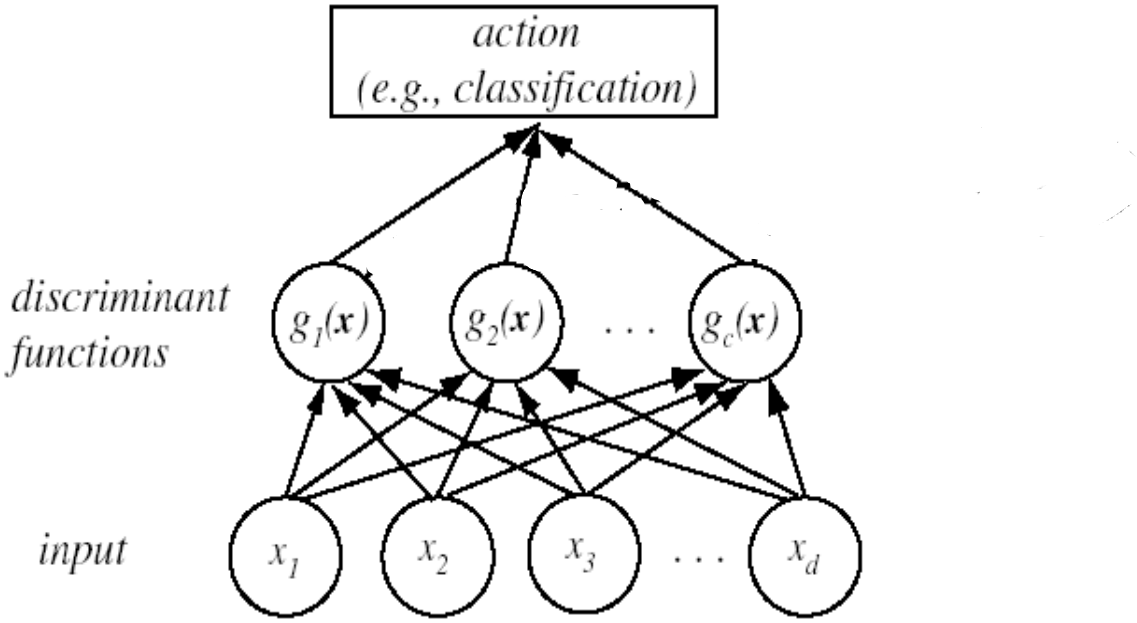
A Gaussian example

$$\begin{cases} R_1 = R_1(\theta) \\ R_2 = R_2(\theta) \end{cases}$$

Discriminant functions and decision regions

- An alternative representation of a pattern classifier is in terms of a set of ***discriminant functions*** $g_i(\mathbf{x})$, i=1,..,c.

- We assign the pattern **x** to the class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})$, $j \neq i$



Discriminant functions, decision regions

- Possible choices for discriminant functions:

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

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

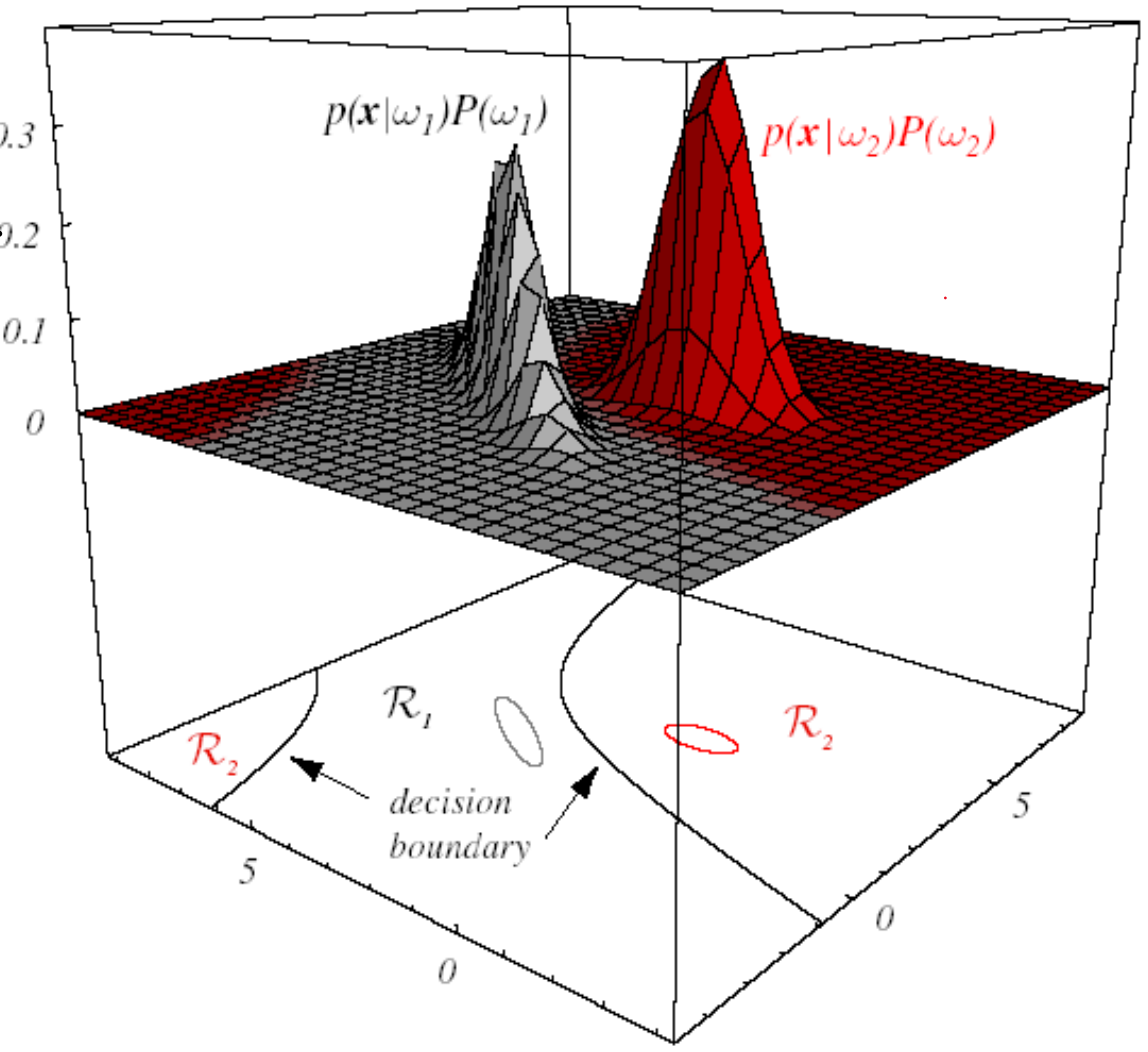
$$g_i(\mathbf{x}) = \ln\big(p(\mathbf{x}/\omega_i)\big) + \ln\big(P(\omega_i)\big)$$

Discriminant functions, decision surfaces/regions

•Discriminat functions subdivide the "feature space" into *c decision regions* R_1 ... R_c

•If $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for any $j \neq i$, then $\mathbf{x} \in R_i$, and it is assigned to the class ω_i.

•**"Decision boundaries"** among regions are specified by $g_i(\mathbf{x}) = g_j(\mathbf{x})$, considering the two discriminant functions exhibiting maximum values



•Bi-dimensional example. Two classes with Gaussian distributions. Quadratic decision surfaces. Region R_2 is not simply connected.

From error probability to risk

It is easy to see that the expression of error probability assumes that all the errors are "equal", that is, all the terms related to probabilities of error ($j \neq i$) have the same "cost" equal to 1.

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

From error probability to risk

However, for some applications, the costs of different actions (classifications) need to be different.

For example, assume that you built a classifier and trained it to predict if a mushroom is poisonous based on the following photograph [https://d2l.ai, Chapter 1].

Death cap—do not eat!

From error probability to risk

Assume that your classifier outputs a probability P(y=deathcap / image)=0.2, namely, P(y=edible mushroom / image)=0.8.

However, it is easy to see that we need to take into account the **risk** that we incur if we eat a poisonous mushroom !

From error probability to risk

We need to compute the **expected risk** that we incur by multiplying the probability of the outcome with the benefit (or harm) associated with it.
In this case, we have two possible actions α_i (α_1=eat, α_2=discard) given the photograph **x**. As an example, we can write the expected risk as follows:

$$R(\alpha_1=\text{eat} / x) = 0.2 * \infty + 0.8 * 0 = \infty$$

$$R(\alpha_2=\text{discard} / x) = 0.2 * 0 + 0.8 * 1 = 0.8$$

Minimum risk theory

- MAP rule does not consider the different costs associated to the different errors

- In some applications this is not a valid choice because errors can bring different losses, and, therefore, they should have different costs.

- The minimum risk theory (also called **utility theory** in economics) takes into account both probabilities and costs of actions/decisions.

Problem formulation:

- Data classes: $\Omega = \{\omega_1, \omega_2, ..., \omega_c\}$;
- Actions/Decisions: $A = \{\alpha_1, \alpha_2, ..., \alpha_a\}$;

➢ In pattern classification, we consider action=classification, that is, the action is the decision about the class of the pattern.

Minimum risk theory

The costs (losses) associated to the different actions given possible classifications are defined by the loss matrix Λ:

$$\Lambda = \begin{bmatrix} \lambda(\alpha_1 \mid \omega_1) & \lambda(\alpha_1 \mid \omega_2) & \cdots & \lambda(\alpha_1 \mid \omega_c) \\ \lambda(\alpha_2 \mid \omega_1) & \lambda(\alpha_2 \mid \omega_2) & \cdots & \lambda(\alpha_2 \mid \omega_c) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda(\alpha_a \mid \omega_1) & \lambda(\alpha_a \mid \omega_1) & \cdots & \lambda(\alpha_a \mid \omega_c) \end{bmatrix}$$

The function $\lambda(\alpha_i|\omega_j)$ is a loss function denoting the "loss/cost" associated to the action/decision α_i when the true data class is ω_j

Minimum risk theory

An example of loss matrix for intrusion detection in computer networks

Ω={ω_1= malicious traffic, ω_2=normal traffic}; *A*= {α_1= server off, α_2= server on};

$$\Lambda = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$$

Bank computer network: $\lambda_{12} << \lambda_{21}$

An example of loss matrix for intrusion detection systems

	Normal Traffic	User to Root Attack	Remote to Local Attack	Probing Attack	Denial of Service Attack
Normal Traffic	**0**	**2**	**2**		**2**
User to Root Attack	**3**	**0**	**2**	**2**	**2**
Remote to Local Attack	**4**	**2**	**0**	**2**	**2**
Probing Attack	**1**	**2**	**2**	**0**	**2**
Denial of Service Attack	**3**	**2**	**2**	**1**	**0**

Minimum risk decision rule

•Let us assume that the action α_i is candidate for execution given that the pattern **x** has been observed. We don't know the true class of the pattern **x**, but let us assume that we know $P(\omega_j|\mathbf{x})$. We can evaluate the ***conditional risk*** associated to the action α_i:

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

➢The conditional risk can be regarded as an average loss/cost.

Minimum risk decision rule

$$\mathbf{x} \rightarrow \alpha_i \Leftrightarrow R(\alpha_i \mid \mathbf{x}) < R(\alpha_j \mid \mathbf{x}) \ \forall i \neq j,\ i=1,\ldots,a$$

Given the pattern **x**, we choose the action α_i with the minimum risk. This is the optimal decision rule for any pattern **x**.

Minimum risk for binary classification

Consider a **two class** problem and the case where **action=classification**

- Therefore, α_i correspond to assign the pattern to the class ω_i
- Let be $\lambda_{ij} = \lambda(\omega_i \,|\omega_j)$ the loss we incur assigning the pattern to the class ω_i when the true class is ω_j
- The conditional risk can be written as follows:

$$R(\omega_1/\mathbf{x}) = \lambda_{11} P(\omega_1/\mathbf{x}) + \lambda_{12} P(\omega_2/\mathbf{x})$$
$$R(\omega_2/\mathbf{x}) = \lambda_{21} P(\omega_1/\mathbf{x}) + \lambda_{22} P(\omega_2/\mathbf{x})$$

- The minimum risk decision rule is:
  - $\mathbf{x} \in \omega_1$ if $R(\omega_1 \,|\mathbf{x}) < R(\omega_2 \,|\mathbf{x})$, else $\mathbf{x} \in \omega_2$

Minimum risk for binary classification

- In terms of posterior probabilities:

$$x \in \omega_1 \text{ if } (\lambda_{21} - \lambda_{11}) P(\omega_1/\mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(\omega_2/\mathbf{x})$$

- According to the Bayes rule:

$$x \in \omega_1 \text{ if } (\lambda_{21} - \lambda_{11}) p(\mathbf{x}/\omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) p(\mathbf{x}/\omega_2) P(\omega_2)$$

- It is reasonable to assume that $\lambda_{21} > \lambda_{11}$. If we make explicit the ratio likelihood $p(\mathbf{x}/\omega_1)/p(\mathbf{x}/\omega_2)$, the above rule can be rewritten as:

The true class is ω_1 if the likelihhod ratio is higher than a threshold θ that does not depend on **x**

$$\mathrm{x} \in \omega_1 \text{ if } l(\mathrm{x}) = \frac{p(x/\omega_1)}{p(x/\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)} = \theta$$

Minimum error and loss matrix 0-1

The action α_i corresponds to the assignment of the "pattern" $\boldsymbol{x}$ to the class ω_i.

In some cases, a simple loss function can be appropriate:

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

Loss matrix 0-1 or "zero-one loss function"

All the errors have the same cost equal to 1. The risk is exactly equal to the error probability:

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

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

Minimum error classification

Using a 0-1 loss function, the minimum risk decision rule become the classical MAP (maximum a posteriori probability):

Assign $\boldsymbol{x}$ to ω_i se $P(\omega_i \mid x) > P(\omega_j \mid x) \quad \forall j \neq i$

Rewriting in terms of the likelihood ratio:

$$\textit{Given} \quad \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \;\; ; \mathrm{x} \in \omega_1 \;\; \text{if} \;\; \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \theta_\lambda$$

Minimum error classification

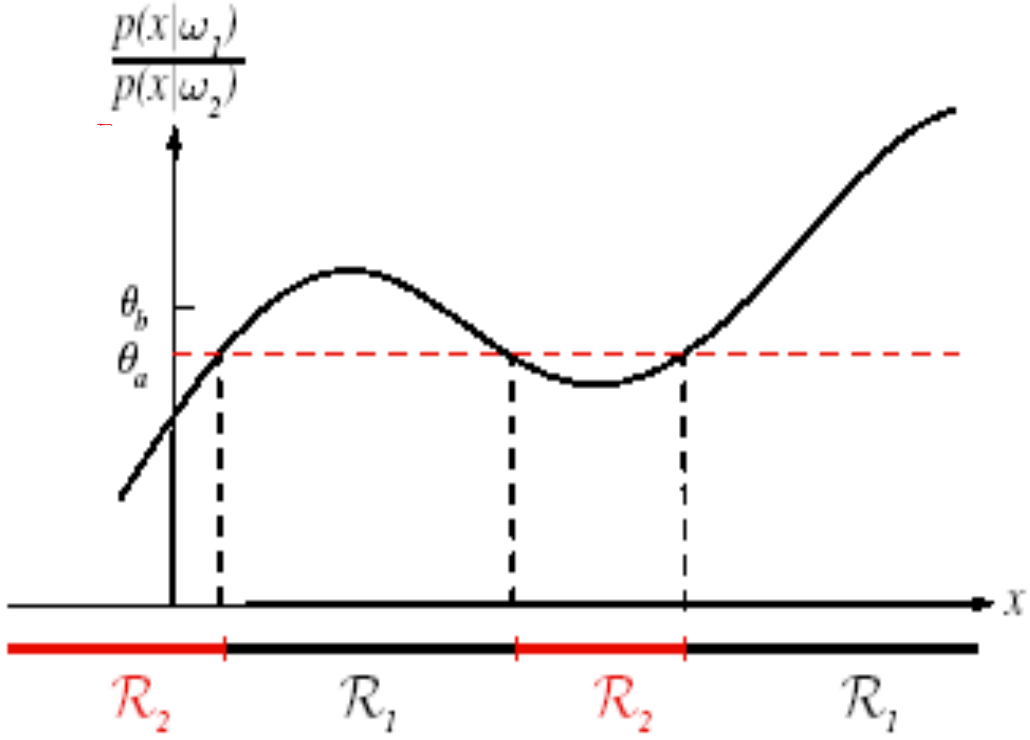
$$Given \quad \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \;\; ; \mathrm{x} \in \omega_1 \;\; \text{if} \;\; \frac{p(x|\omega_1)}{p(x|\omega_2)} > \theta_\lambda$$

Examples

$$If\ \Lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} then\ \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

$$If\ \Lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} then\ \theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$$

Minimum error classification and decision regions



$$\theta = \frac{\lambda_{12}}{\lambda_{21}}$$

If errors for class ω_1 are more costly the threshold is more tight and the decision region R_1 becomes smaller

Class ω_1 should have a higher likelihood if
$\lambda_{12} > \lambda_{21}$

We have the threshold θ_a when $P(\omega_1)=P(\omega_2)$ and $\lambda_{12} = \lambda_{21} = 1$

We have θ_b when $\lambda_{12} > \lambda_{21}$

The region R_1 decreases when $\lambda_{12} > \lambda_{21}$

Example (1)

Let us suppose that we want to discriminate between normal and intrusive network traffic, namely, **two** data classes $\boldsymbol{\omega_N}$, normal traffic, and $\boldsymbol{\omega_{INT}}$, intrusive network traffic. We suppose to use a single *feature* **x** to characterize traffic data (one-dimensional feature space), and we assume that the model of the network traffic is the following:

$$P(\omega_N) = \frac{1}{2}; P(\omega_{INTR}) = \frac{1}{2}$$

$$p(x/\omega_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma}\right)^2\right];$$

$$\mu_N = 0; \mu_{INTR} = 4; \;\; \sigma_N = \sigma_{INTR} = 1;$$

- Let the cost of missing the detection of intrusion be ten times higher than the opposite error (a normal traffic is wrongly recognized as an intrusion).

Example (2)

a) Find the decision regions using the likelihood ratio test, without considering the costs of errors, and compute the total error probability.

$$l(x) = \frac{p(x/\omega_N)}{p(x/\omega_{INTR})} \begin{matrix} \omega_N \\ > \\ < \\ \omega_{INTR} \end{matrix} \frac{P(\omega_{INTR})}{P(\omega_N)} = \theta \qquad \theta = \frac{P(\omega_{INTR})}{P(\omega_N)} = 1$$

$$l(x) = exp\left[\frac{1}{2}\left(\left(\frac{x-4}{1}\right)^2 - \left(\frac{x-0}{1}\right)^2\right)\right] =$$

$$exp\left[\frac{1}{2}(x^2 + 16 - 8x - x^2)\right] =$$

$$= exp[8 - 4x]$$

$$l(x) = \theta \Longrightarrow exp[8 - 4x] = 1 \Longrightarrow$$

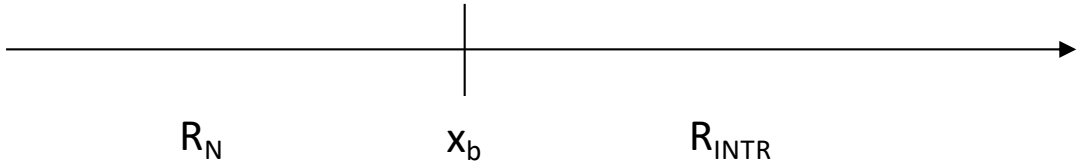
$$8 - 4x = \ln(1) \Longrightarrow \; x_b = 2$$

Example (3)

$$x_b = 2$$

$$l(x) > \theta \Rightarrow x < x_b$$

Let R_N e R_{INTR} be the two decision regions. If x belongs to R_N the traffic is labelled as 'normal'. If x belongs to R_{INTR} the incoming traffic is labelled as 'intrusion'.



Example (4)

Total error probability:

Two components of error probability: (intrusions wrongly labeled as normal traffic) + (normal traffic wrongly labeled as intrusion)

(the actual value of the integrals can be found by looking at the table of values of the erf function)

$$P\{x \in R_N, x \in \omega_{INTR}\} + P\{x \in R_{INTR}, x \in \omega_N\} =$$

$$P\{x \in R_N / \omega_{INTR}\} P(\omega_{INTR}) + P\{x \in R_{INTR} / \omega_N\} P(\omega_N) =$$

$$\int_{-\infty}^{x^*} p(x \mid \omega_{INTR}) P(\omega_{INTR}) dx + \int_{x^*}^{\infty} p(x \mid \omega_N) P(\omega_N) dx =$$

$$\frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{2} \exp\left[-\frac{1}{2}(x-4)^2\right] dx + \int_{2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x)^2\right] dx\right] =$$

$$\frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2} \exp\left[-\frac{1}{2}(y)^2\right] dy + \int_{2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x)^2\right] dx\right] =$$

$$\frac{1}{2}[0.0228 + 0.0228] = 0.0228$$

Example (5)

b) Specify the loss (cost) matrix that satisfies the assumption:
the cost of missing the detection of intrusion be ten times higher than the opposite error (a normal traffic is wrongly recognized as an intrusion).

We can indicate with $\lambda_{N,Intr}$ the cost when the traffic is *intrusive* but it is classified as *normal* (and vice versa for $\lambda_{Intr,N}$). Therefore, we can set $\lambda_{N,Intr} = 10 \cdot \lambda_{Intr,N}$. A possible loss matrix is therefore:

$$\Lambda = \begin{bmatrix} \lambda_{N,N} & \lambda_{N,Intr} \\ \lambda_{Intr,N} & \lambda_{Intr,Intr} \end{bmatrix} = \begin{bmatrix} 0 & \lambda_{N,Intr} \\ \lambda_{Intr,N} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 10 \\ 1 & 0 \end{bmatrix}$$

Homework 3 - send me your answer to roli@unica.it

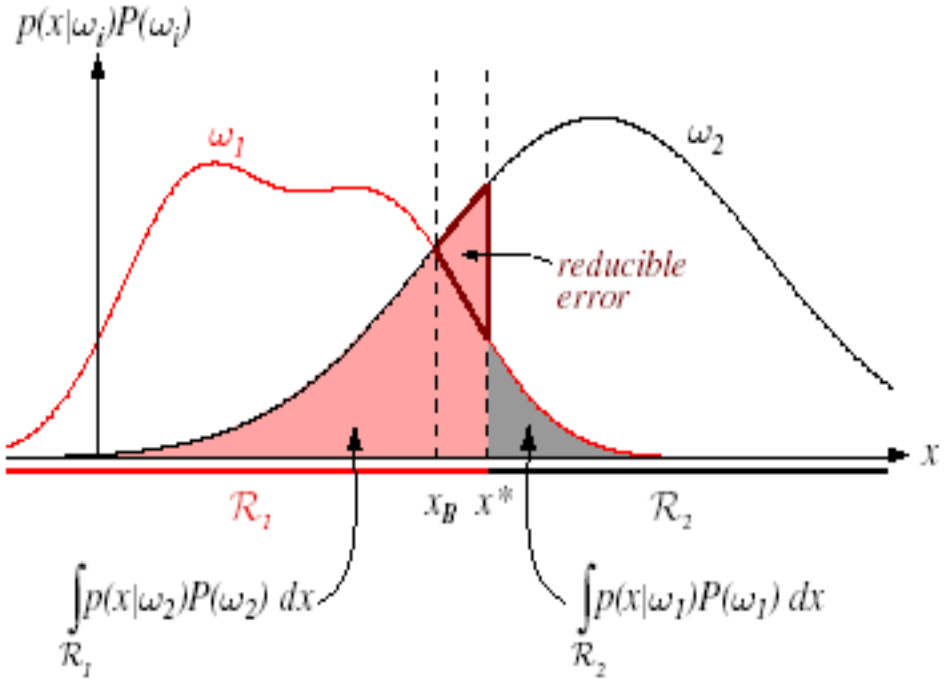
With reference to the previous example:

1)find the decision regions that minimize the risk, and compute the related classification error.
2)Explain why the decision regions are changed with respect to the use of the the likelihood ratio test, without considering the costs of errors. Why this change?
3)Explain why the two components of the total error are changed with respect to the use of the the likelihood ratio test, without considering the costs of errors. Why this change?

Decision with reject option

•Even if one would be able to achieve the minimum Bayes error (threshold x_B in figure), this error rate could be not acceptable for a given application

•Example: "screening" for medical diagnosis. I could demand for a "false negative" rate equal to zero.



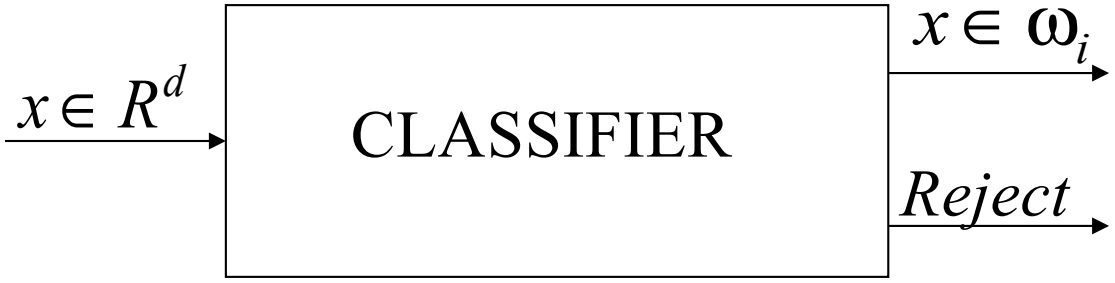
Decision with reject option

✓An obvious way to limit decision errors is not making a decision or postponing decisions

➢To reduce error probability one can omit or defer decisions (**reject option**)

➢Omitting decisions is a rationale and doable option supposed that decisions are taken by other ways (e.g., by humans)

Classification with reject option



It is easy to see that rejection option demands for an additional class with respect to the standard formulation of the classification problem:

- Set of classes: $\Omega = \{\omega_1, \omega_2, ..., \omega_c\}$;
- Set of actions/decisions: $A = \{\alpha_o, \alpha_1, \alpha_2, ..., \alpha_a\}$;
- If our action is a classification: $A = \{\omega_o, \omega_1, \omega_2, ..., \omega_c\}$;

➢ We have an additional class: ω_o, the class containing the rejected samples

Loss matrix and minimum risk with reject option

The loss matrix Λ, with size *(c+1)x c*, is:

$$\Lambda = \begin{bmatrix} \lambda(\omega_0 \mid \omega_1) & \lambda(\omega_0 \mid \omega_2) & \cdots & \lambda(\omega_0 \mid \omega_c) \\ \lambda(\omega_1 \mid \omega_1) & \lambda(\omega_1 \mid \omega_2) & \cdots & \lambda(\omega_1 \mid \omega_c) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda(\omega_c \mid \omega_1) & \lambda(\omega_c \mid \omega_1) & \cdots & \lambda(\omega_c \mid \omega_c) \end{bmatrix}$$

The minimum risk decision criterion is:

$$\mathbf{x} \rightarrow \omega_i \Leftrightarrow R(\omega_i \mid \mathbf{x}) < R(\omega_j \mid \mathbf{x}) \ \forall \mathrm{i} \neq \mathrm{j},\ \mathrm{i}=0,1,..,\mathrm{c}$$

The main difference w.r.t. the case without reject option is that the minimum risk decision could be a “rejection”, if:

$$R(\omega_0 \mid \mathbf{x}) < R(\omega_j \mid \mathbf{x}) \ \forall j \neq 0$$

Binary classification with equal costs

Let us consider a binary classification with equal costs:

$$\Lambda = \begin{pmatrix} \lambda_r & \lambda_r \\ \lambda_c & \lambda_e \\ \\ \lambda_c & \lambda_e \end{pmatrix} = \begin{pmatrix} \lambda(\omega_0 \mid \omega_1) & \lambda(\omega_0 \mid \omega_2) \\ \lambda(\omega_1 \mid \omega_1) & \lambda(\omega_1 \mid \omega_2) \\ \\ \lambda(\omega_2 \mid \omega_2) & \lambda(\omega_2 \mid \omega_1) \end{pmatrix}$$

Reject cost = λ_r Error cost = λ_e Cost of correct classification = λ_c (usually λ_c=0)

According to the minumum risk criterion, we have three decision regions:

$$R_0 = \left\{ x \in R : R(\omega_0 \mid \mathbf{x}) < R(\omega_j \mid \mathbf{x}) \ \forall j \neq 0 \right\}$$

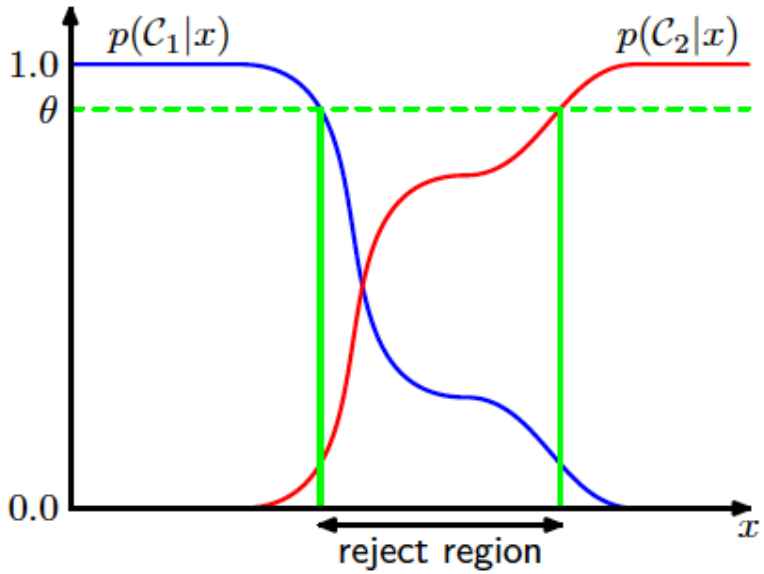
$$R_1 = \left\{ x \in R : R(\omega_1 \mid \mathbf{x}) < R(\omega_j \mid \mathbf{x}) \ \forall j \neq 1 \right\}$$

$$R_2 = \left\{ x \in R : R(\omega_2 \mid \mathbf{x}) < R(\omega_j \mid \mathbf{x}) \ \forall j \neq 2 \right\}$$

Illustration of the reject option

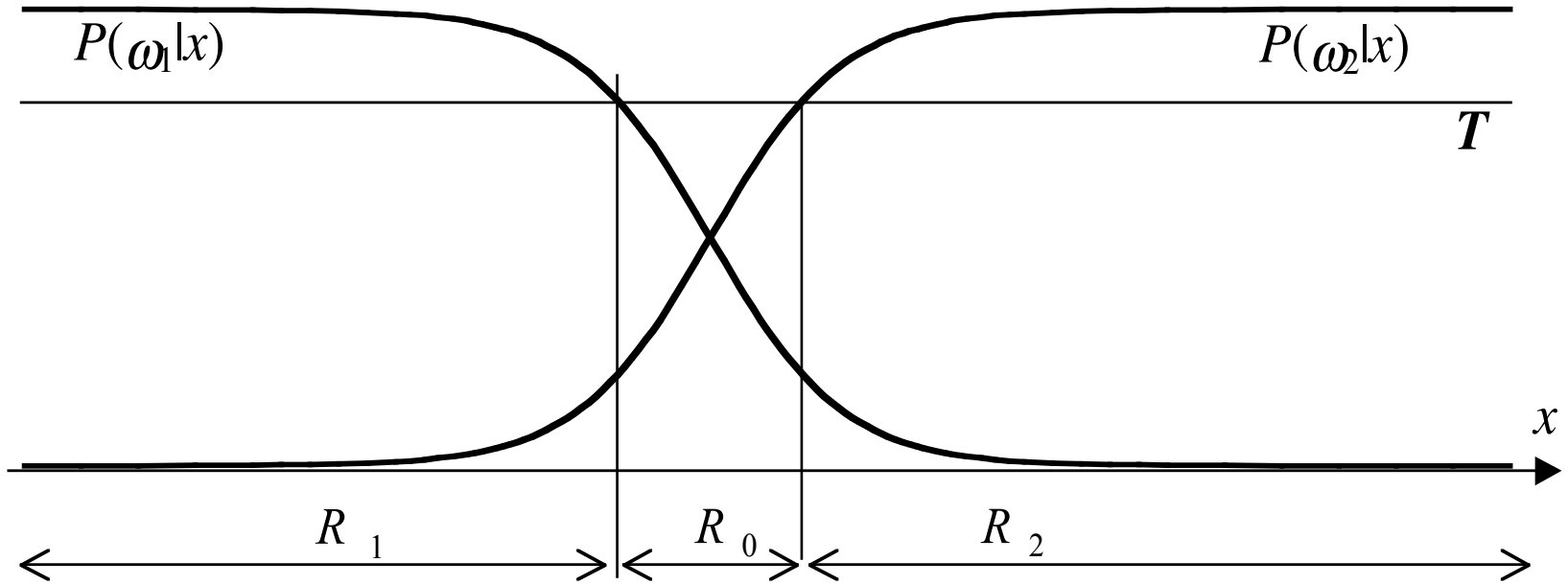
Illustration of the reject option. Inputs x such that the larger of the two posterior probabilities is less than or equal to some threshold θ will be rejected.

[C.Bishop, Pattern Recognition and Machine Learning, 2006]



- Classification errors arise from the regions of feature space where the largest of the posterior probabilities $p(\omega_k|x)$ is significantly less than unity, or equivalently where the joint distributions $p(x, \omega_k)$ have comparable values.
- These are the regions where we are relatively uncertain about class membership.

Simple example of reject option



- Two classes with Gaussian distribution
- The reject threshold *T* identifies the reject region R_0
- This example clearly shows that error probability can be reduced by increasing the reject threshold *T*. Error becomes zero when the region R_0 contains all the patterns which are misclassified.

Error-reject trade-off and Chow's rule (equal costs)

The reject threshold **T** an be found by the **Chow's rule** (C.K. Chow, 1970), that also provides the optimal rule with reject option:

$$if \quad \max_i P(\omega_i / x) \geq T \rightarrow x \in \omega_i$$

$$otherwise \text{ reject } x$$

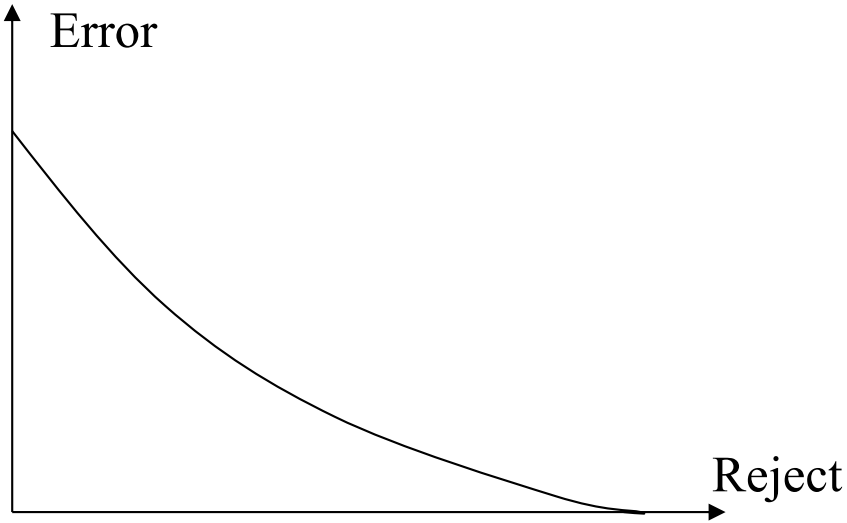
$$with \ T = \frac{\lambda_e - \lambda_r}{\lambda_e - \lambda_c}$$

- T is the **reject threshold**
- T ε [0..1] because $\lambda_c \leq \lambda_r$
- For T=0 ($\lambda_e=\lambda_r$) we have the classical MAP rule

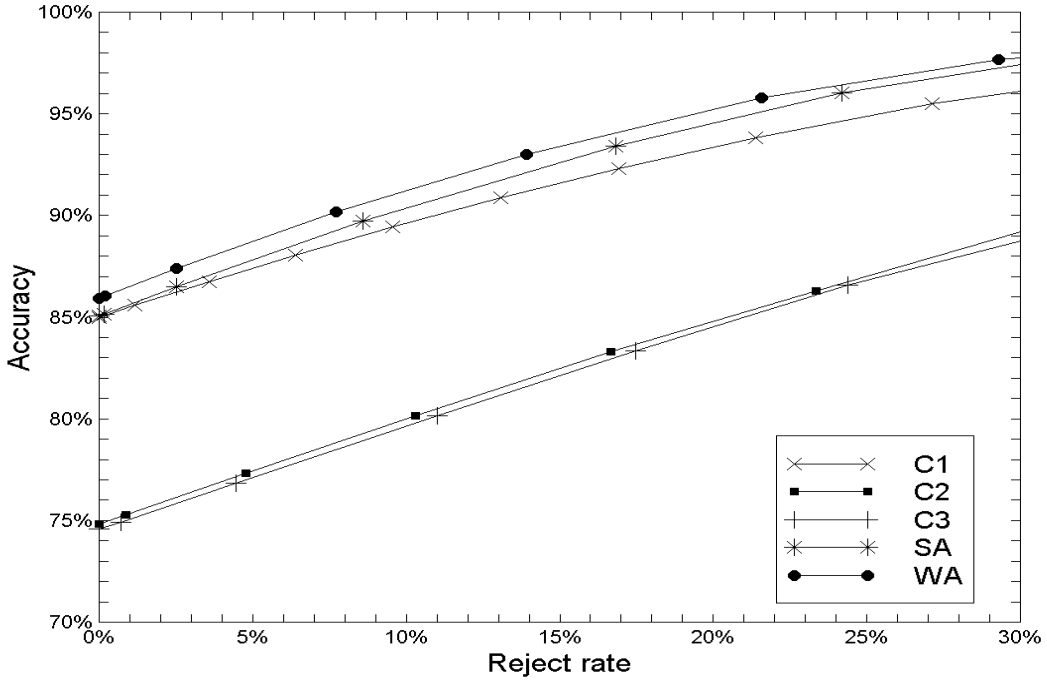
➢ We can show that Chow's rule minimizes error probability (that is, maximize classification accuracy) for any value of the reject probability.

➢ It is easy to see that Chow's rule minimizes error by rejecting patterns for which the classification is not reliable enough.

Examples of error-reject trade-off



- Hypothetical trade-off curve
- *T* increases, then the rejection incresaes and error decreases (error-reject trade-off)



Examples of accuracy-rejection for different OCR (Optical Character Recognition) algorithms

References

- Sections 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, Pattern Classification, R.O. Duda, P. E. Hart, and D. G. Stork, John Wiley & Sons, 2000
- Chapter 1, Statistical Pattern Recognition, Andrew Webb, John Wiley & Sons, 2002
- C.K. Chow, On optimum error and reject trade-off, IEEE Trans. on Information Theory 16 (1970) 41-46

Example of classification with reject option (1)

Let us suppose that we want to diagnose a disease of which we know the prior probability:
$P(\omega_{healthy})=0.85$, $P(\omega_{sick}) =0.15$
$P(\omega_{sick})$ is the prior probability that a person within a given population is affected by this disease.

The disease can be diagnosed by the amount of a certain substance in the blood. The amount of this substance is higher for people affected by the disease.

Let $\mu_h=4$ and $\mu_s=8$ be the average amount of this substance, respectively, for **healthy** people and **sick** people . The amount of substance in the two cases is Gaussian distributed around the average value, with $\sigma=1$

$p(x|\omega_i) = N(\mu_i, \sigma^2)$; i=1 healthy, i=2 sick

Example of classification with reject option (2)

Let us suppose that we want to reduce the minimum Bayes error using the Chow reject option.

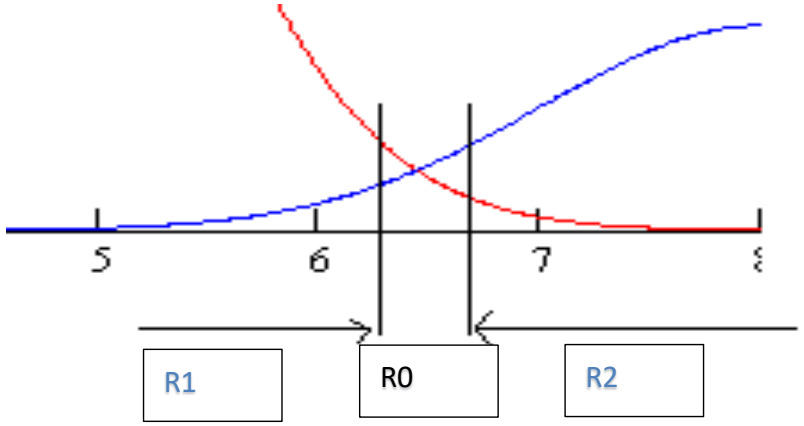
First, we should define a proper loss matrix. Let's consider the following matrix.

$$\Lambda = \begin{pmatrix} \lambda_R & \lambda_R \\ \lambda_{SS} & \lambda_{SH} \\ \lambda_{HS} & \lambda_{HH} \end{pmatrix} = \begin{pmatrix} 0.3 & 0.3 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The threshold *T* of the Chow's rule is:

$$\mathrm{T} = \frac{\lambda_E - \lambda_R}{\lambda_E - \lambda_C} = \frac{1 - 0.3}{1} = 0.7$$

Example of classification with reject option (3)



We know that the inequality $\max[P(\omega_i \mid x)] < T$ defines the region of *rejection* (reject region).

Reject region R0: $x \in [x_{S1}, x_{S2}]$

Let's find the reject region R0.

In the following, we show the calculations to compute x_{S1} *and* x_{S2}.

Example of classification with reject option (4)

$$x_{S1} \rightarrow P(\omega_1/x) < T \text{ and } P(\omega_1|x) > P(\omega_2|x)$$

$$x_{S2} \rightarrow P(\omega_2/x) < T \text{ and } P(\omega_2|x) > P(\omega_1|x)$$

$$x_{S1} \rightarrow P(\omega_1/x) < T \Rightarrow \frac{p(x/\omega_1)P(\omega_1)}{p(x)} < T$$

$$\Rightarrow \frac{p(x/\omega_1)P(\omega_1)}{p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2)} < T$$

Example of classification with reject option (5)

$$\frac{p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2)}{p(x/\omega_1)P(\omega_1)} > \frac{1}{T} \Rightarrow$$

$$\Rightarrow 1 + \frac{p(x/\omega_2)}{p(x/\omega_1)}\frac{P(\omega_2)}{P(\omega_1)} > \frac{1}{T}$$

$$\Rightarrow \frac{p(x/\omega_2)}{p(x/\omega_1)} > \left[\frac{1}{T} - 1\right]\frac{P(\omega_1)}{P(\omega_2)}$$

Given that:

$$\frac{p(x/\omega_2)}{p(x/\omega_1)} =$$

$$= \exp\left[\frac{1}{2}(x^2 - 8x + 16) - \frac{1}{2}(x^2 - 16x + 64)\right] = \exp[4x - 24]$$

Example of classification with reject option (6)

$$\exp[4x - 24] > \left[\frac{1}{T} - 1\right]\frac{P(\omega_1)}{P(\omega_2)} = \left[\frac{1}{0.7} - 1\right]\frac{85}{15} = \frac{3}{7}\frac{85}{15} = \frac{17}{7}$$

$$4x - 24 > \ln\left(\frac{17}{7}\right); x > 6 + \frac{1}{4}\ln\left(\frac{17}{7}\right);$$

$$x_{s1} = 6 + \frac{1}{4}\ln\left(\frac{17}{7}\right) \cong 6.2218$$

Example of classification with reject option (7)

$$x_{S2} \rightarrow P(\omega_2/x) < T \Rightarrow \frac{p(x/\omega_2)P(\omega_2)}{p(x)} < T \Rightarrow$$

$$\Rightarrow \frac{p(x/\omega_2)P(\omega_2)}{p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2)} < T$$

Example of classification with reject option (8)

$$\frac{p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2)}{p(x/\omega_2)P(\omega_2)} > \frac{1}{T} \Rightarrow 1 + \frac{p(x/\omega_1)}{p(x/\omega_2)}\frac{P(\omega_1)}{P(\omega_2)} > \frac{1}{T}$$

$\Rightarrow$

$$\Rightarrow \frac{p(x/\omega_1)}{p(x/\omega_2)} > \left[\frac{1}{T} - 1\right]\frac{P(\omega_2)}{P(\omega_1)}$$

$$\Rightarrow \frac{N(\mu_1, \sigma^2)}{N(\mu_2, \sigma^2)} > \left[\frac{1}{T} - 1\right]\frac{P(\omega_2)}{P(\omega_1)}$$

Example of classification with reject option (9)

$$\exp[24 - 4x] > \left[\frac{1}{T} - 1\right]\frac{P(\omega_2)}{P(\omega_1)} = \left[\frac{1}{0.7} - 1\right]\frac{15}{85} = \frac{3}{7}\frac{15}{85} = \frac{9}{119}$$

$$24 - 4x > \ln\left(\frac{9}{119}\right) \; ; x < 6 - \frac{1}{4}\ln\left(\frac{9}{119}\right) \; ;$$

$$x_{s2} = 6 - \frac{1}{4}\ln\left(\frac{9}{119}\right) \cong 6.6455$$

Example of classification with reject option (10)

Now we can compute the error probability and the reject probability.

False positive rate:

$$\int_{x_{S2}}^{\infty} p(x/healthy)\, P\left(\omega_{healthy}\right) dx \cong 3.45 \times 10^{-3}$$

False negative rate:

$$\int_{-\infty}^{x_{S1}} p(x/\omega_{sick})\, P(\omega_{sick})\, dx \cong 5.65 \times 10^{-3}$$

Total error probability = 9.12 10^{-3}

Example of classification with reject option (11)

The error is now lower than the Bayesian one without reject option. However, we do NOT classify (we reject) 15 patients every 1000.

$$\int_{x_{S1}}^{x_{S2}} p(x/\omega_{sick})\, P(\omega_{sick})\, dx$$

$$+ \int_{x_{S1}}^{x_{S2}} p\left(x/\omega_{healthy}\right) P\left(\omega_{healthy}\right) dx \cong 15.22 \times 10^{-3}$$

Homework 4 - send me your answer to roli@unica.it

With reference to the previous example on reject option:

1) Using the MAP decision rule, compute separately the two error probabilities for the two classes $\omega_{healthy}$ and ω_{sick} , and then the total (minimum) error probability.

2)Explain how and why the above error probabilities are different from the ones that we have computed with the reject option.

Homework 5 (Bonus 1/30) - send me your answer to roli@unica.it

Let's consider a 2-class problem in a one-dimensional feature space.
The class-conditional probability densities of the 2 classes are:

$$p(x|\omega_1) = N(x;\ \mu_1 = -1, \sigma_1 = 1)$$
$$p(x|\omega_2) = N(x;\ \mu_2 = +1, \sigma_2 = 1)$$

Recall that $N(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, and that $\int_{-\infty}^{x} N(x;\ \mu, \sigma)\, dx = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)$. Assume that the prior probabilities of the two data classes are $P(\omega_1) = 0.5P(\omega_2)$, and that the cost of errors of class ω_1 is double than those of class ω_2, that is: $\lambda_{21} = 2\,\lambda_{12}$ $\lambda_{11} = \lambda_{22} = 0$

1)Compute the minimum-risk decision regions and, separately, the two errors (one per class). Compare these errors with the corresponding errors achieved using the Bayesian optimal threshold. Justify the results. *Is the Bayesian error higher/lower than the error obtained with the minimum-risk approach? Why?*

Homework 5 - send me your answer to roli@unica.it

2) Let us now assume that $P(\omega_1) = P(\omega_2)$, while $p(x|\omega_i)$ is defined as in previous question 1. We want to minimize the classification error (not the risk!) using the rejection option with the Chow's rule.

Let us assume that the rejection region is defined as [-0.5, 0.5].

- Compute the rejection threshold T on the posterior probabilities.
- Compute the fraction of rejected samples of class 2.
- Compute the two errors, separately for each class.

Generative machine learning models: the Gaussian Pattern Classifiers

Generative vs. discriminative models

In machine learning, **generative** models assume to know the parametric form of the distribution $p(\boldsymbol{x}/\omega_i)$ and model the joint probability distribution $p(\boldsymbol{x},\omega_i)$ to design the pattern classifier.

Now, we see an example of pattern classifier based on a generative model: the **Gaussian Classifier**

In part 4, we see **discriminative** models…

Homework 6 (Bonus 0.2/30) - send me your answer to roli@unica.it

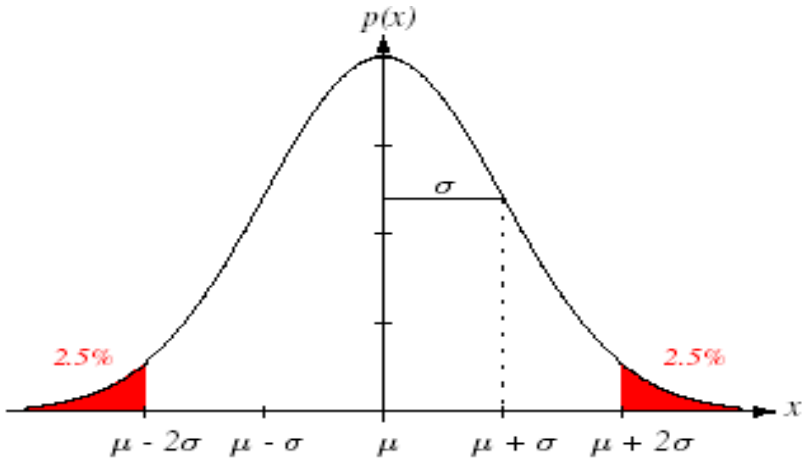
Why generative machine learning models are called ***«generative»*** ?

Now let's summarize the Gaussian distribution…

Carl Friedrich Gauss

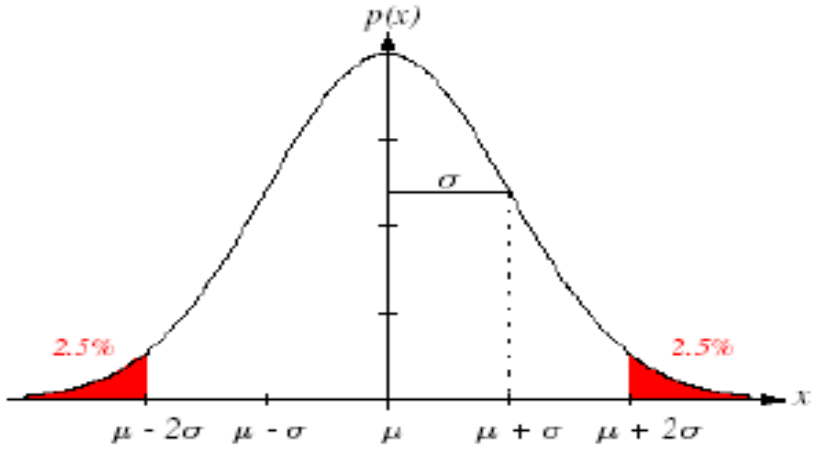
Gaussian probabilistic model

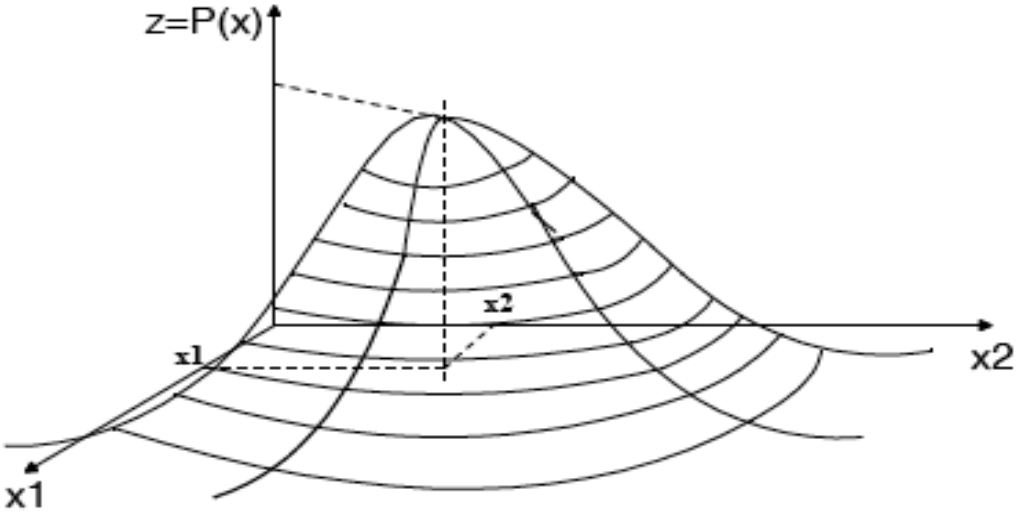
- Mono-dimensional Gaussian model: $p(x/\omega_i) = N(\mu, \sigma)$



If the model of $p(x/\omega_i)$ is Gaussian (see on the left), the problem is only to estimate the two parameters μ and σ.

The Gaussian model





Why is the Gaussian model so widely used?

✓Several natural and/or artificial phenomena fit the Gaussian model (even scores of university exam are roughly "normally-distributed"…)

✓Central Limit Theorem: the sum of N independent random variables will lead to a Gaussian distribution for $N \rightarrow +\infty$

✓In several machine learning tasks, a pattern can be regarded as an ideal 'prototype' corrupted by a sum of random and independent noisy sources

✓In some cases, the distribution is not Gaussian, but it can be approximated by a Gaussian distribution.

Quick recap: parametric form of the Gaussian model

- One-dimensional case:

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

$$\mu \equiv \varepsilon[x] = \int_{-\infty}^{+\infty} x p(x) dx$$

$$\sigma^2 \equiv \varepsilon\left[(x-\mu)^2\right] = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx$$

- Multidimensional case:

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

$\mathbf{x}$ and μ are column vectors of d components, Σ is the **covariance matrix** d x d, $|\Sigma|$ and Σ^{-1} are its determinant and its inverse.

$$\boldsymbol{\mu} \equiv \varepsilon[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$$\Sigma \equiv \varepsilon\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}$$

$$\sigma_{ij} \equiv \varepsilon\left[(x_i - \mu_i)(x_j - \mu_j)\right]$$

Remarks on the covariance matrix

Properties of Σ:

- Σ is a symmetric matrix: $\Sigma = \Sigma^t$;
- Σ is a semi-positive definite matrix. However, in order to obtain a well-defined Gaussian probability density function, Σ must be positive definite (in fact, the expression of $p(\mathbf{x})$ involves the inverse of Σ and the division by the determinant $|\Sigma|$).

Independent random variables::

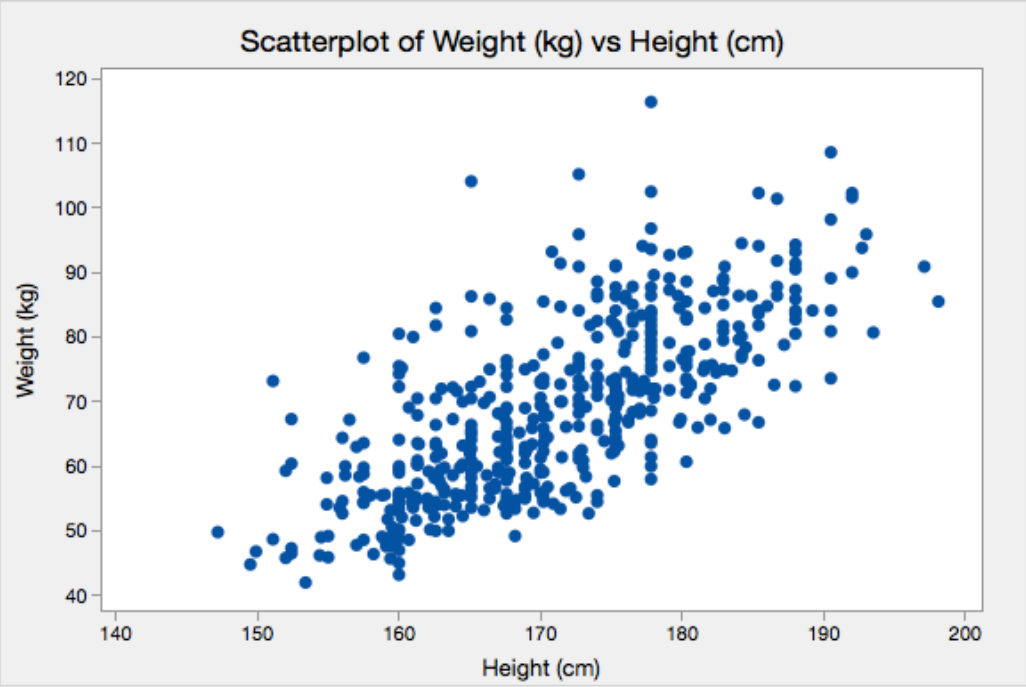
- Given the covariance matrix Σ

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \cdots & \cdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}$$

- If $\sigma_{ij} = 0$, then the random variables x_i and x_j are uncorrelated and, being Gaussian distributed, they are also independent;
- If $\sigma_{ij} = 0$ for all $i \neq j$ (that is, if Σ is a diagonal matrix), we have:

$$p(\mathbf{x}) = p(x_1)\, p(x_2) \ldots p(x_d)$$

Covariance matrix



Correlation (x, y)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

What about the covariance between the feature «height» and «weight»? is positive, negative, zero?

Estimation of Gaussian parameters

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}\right)\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}\right)^t$$

One-dimensional feature space:

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

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \hat{\mu}\right)^2 \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \hat{\mu}\right)^2$$ Unbiased estimator

Gaussian Classifiers

- Let us now consider which classifiers, and which types of discriminant functions, may be obtained by assuming a Gaussian model of the data.
- We will also see the different types of classifiers (generative machine learning models) that are obtained with different hypotheses on the covariance matrix

If $p(\mathbf{x}/\omega_i) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

The associate discriminant function is:

$$g_i(\mathbf{x}) = \ln\left(p(\mathbf{x}/\omega_i)\right) + \ln\left(P(\omega_i)\right)$$

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

$$P(\omega_i) = \frac{n_i}{\sum_j n_j}$$

Gaussian model: case $\Sigma_i=\sigma^2\mathbf{I}$

- $\Sigma_i=\sigma^2\mathbf{I}$ means that
  - The "feature" are statistically independent and have the same variance.
  - The data ("pattern") form hyper-spherical "clusters" (groups) of identical size, and centers μ_i
- In this simple case we obtain:

$$\left|\Sigma_i\right| = \sigma^{2d} \qquad \Sigma_i^{-1} = \left(1/\sigma^2\right)\mathbf{I}$$

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

Gaussian model: case $\Sigma_i=\sigma^2\mathbf{I}$

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

- Given that we assign the pattern $\mathbf{x}$ to the class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})$, $j \neq I$
- We can disregard terms do not depend on the class !

Therefore, the above $g_i(\boldsymbol{x})$ can be rewritten as:

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

Nearest mean classifier

- Special case: $P(\omega_i) = P(\omega_j)$ for each class
  - $P(\omega_i)$ becomes irrelevant to the classification
  - The discriminant function becomes

$$g_i(\mathbf{x}) = -\left\|\mathbf{x} - \boldsymbol{\mu}_i\right\|^2$$

  - This classifier is called **nearest mean classifier** and it is used in the classification procedure called template matching (where each class is represented by its "prototype" μ_i)

Gaussian model, with $\Sigma_i=\sigma^2\mathbf{I}$, the linear classifier

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

- Rewriting the $g_i(\boldsymbol{x})$ above and noting that the term $(\mathbf{x}^t\,\mathbf{x})$ is the same for all values of i, we obtain the *linear* discriminant function:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0} \qquad w_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i;\;\; w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$$

w_{i0} is called *threshold* or "*bias*" for the i-th class

Gaussian model, with $\Sigma_i=\sigma^2\mathbf{I}$: Decision boundary

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

w_{i0} is called *threshold* or "*bias*" for the i-th class

- The decision boundary is a portion of hyperplane of dimension d-1 defined by

$g_i(\mathbf{x}) = g_j(\mathbf{x})$ for classes with the highest posterior probability

Gaussian model, with $\Sigma_i=\sigma^2 I$: Decision boundary

$g_i(\mathbf{x}) = g_j(\mathbf{x})$ for classes with the highest posterior probability

In this case the equation of hyperplanes can be written as

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

where

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

Gaussian model, with $\Sigma_i=\sigma^2 I$: Decision boundary

$g_i(\mathbf{x}) = g_j(\mathbf{x})$ for classes with the highest posterior probability

In this case the equation of hyperplanes can be written as

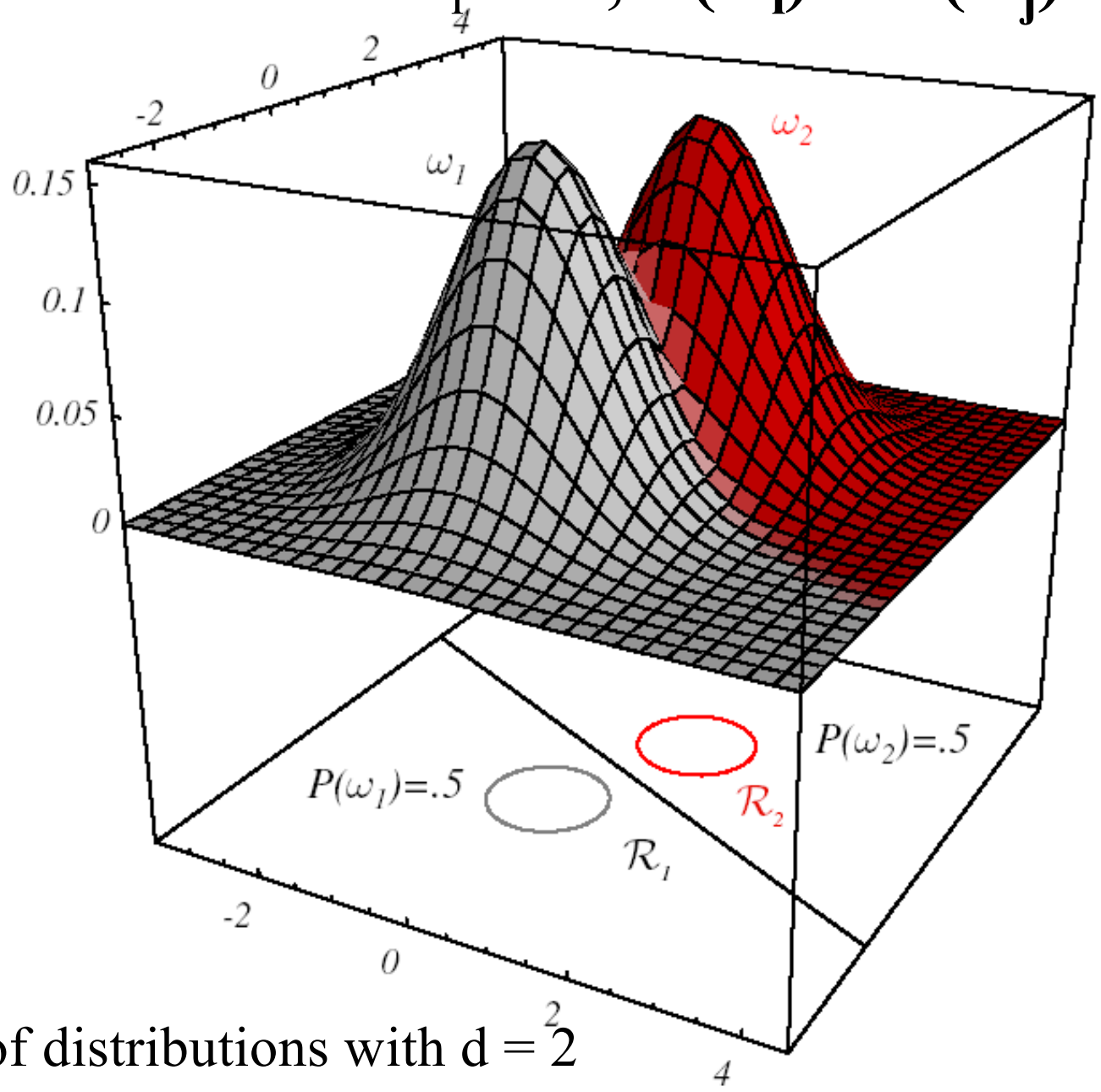
$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$

where

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

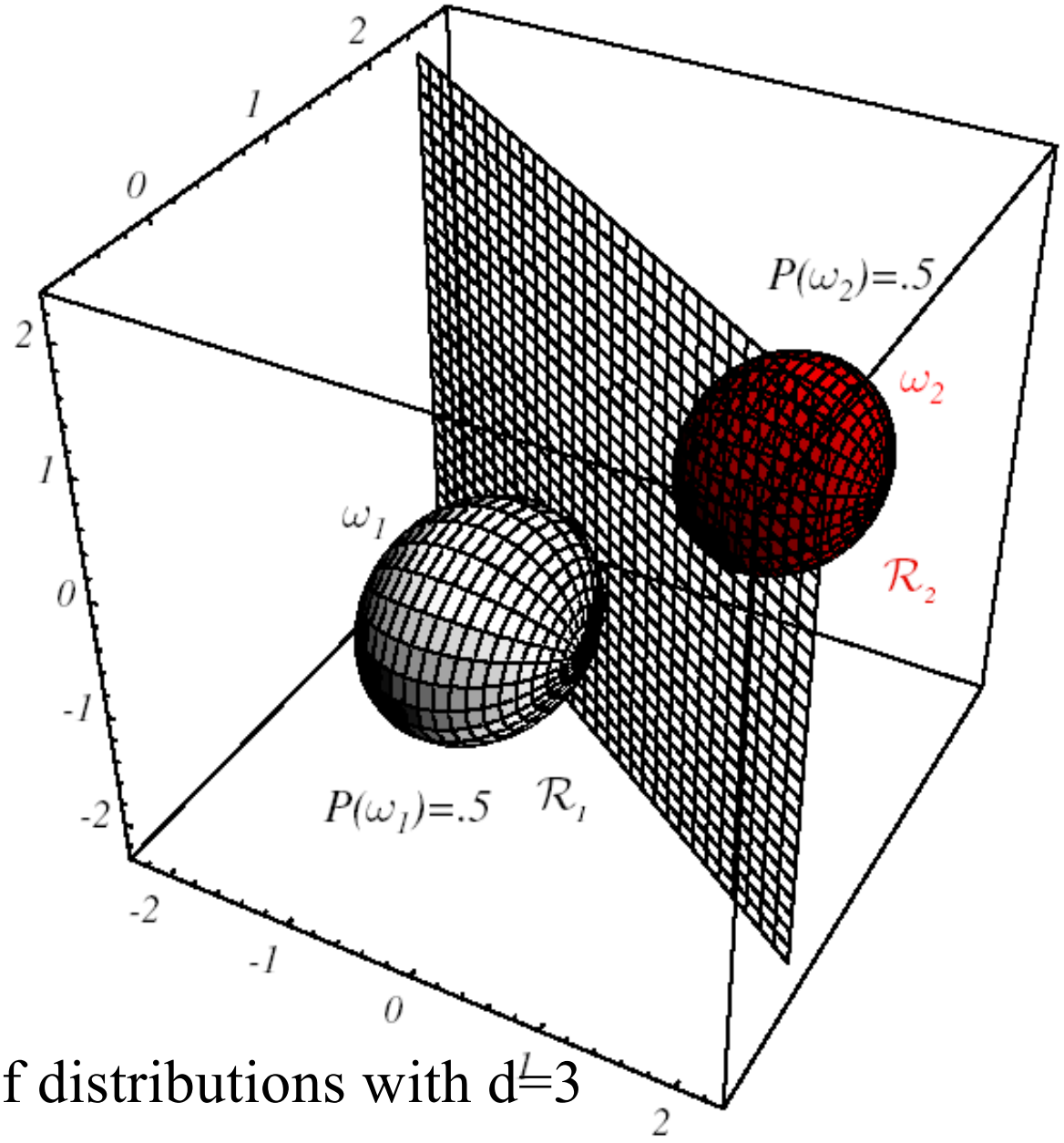
- The hyperplane that separates the regions R_i and R_j is orthogonal to the line joining the means
- The values of $P(\omega_i)$ and $P(\omega_j)$ determine the position of the point $\mathbf{x}_0$ in which the hyperplane passes

Example for the case $\Sigma_i=\sigma^2 I$, $P(\omega_i) = P(\omega_j)$



- Example of distributions with d = 2

Example for the case $\Sigma_i=\sigma^2 I$, $P(\omega_i) = P(\omega_j)$

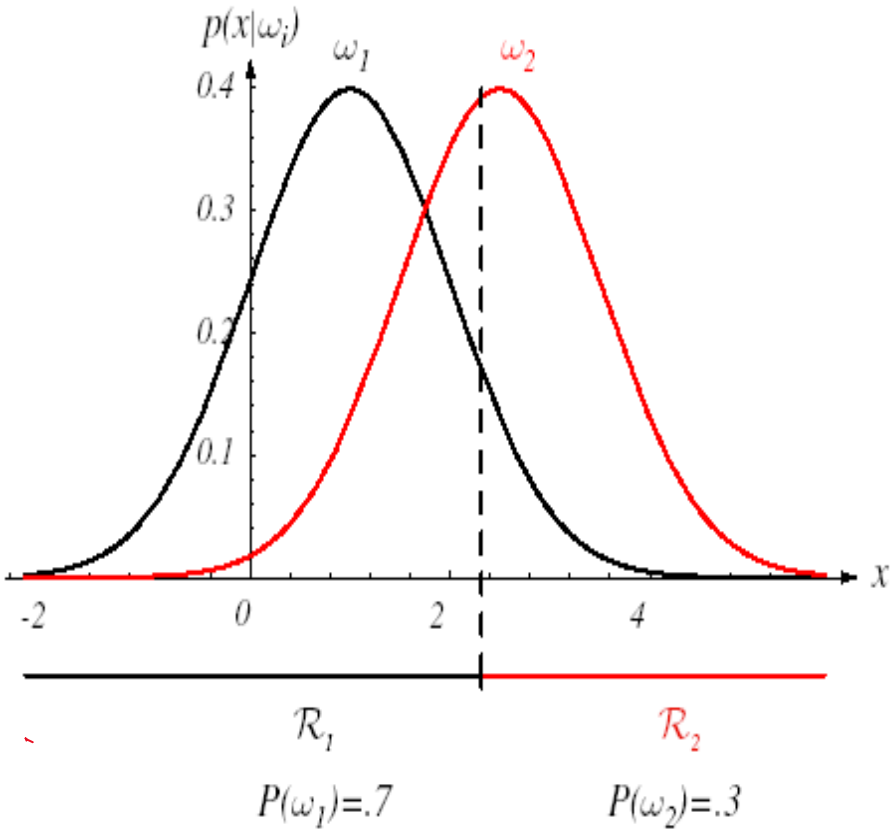


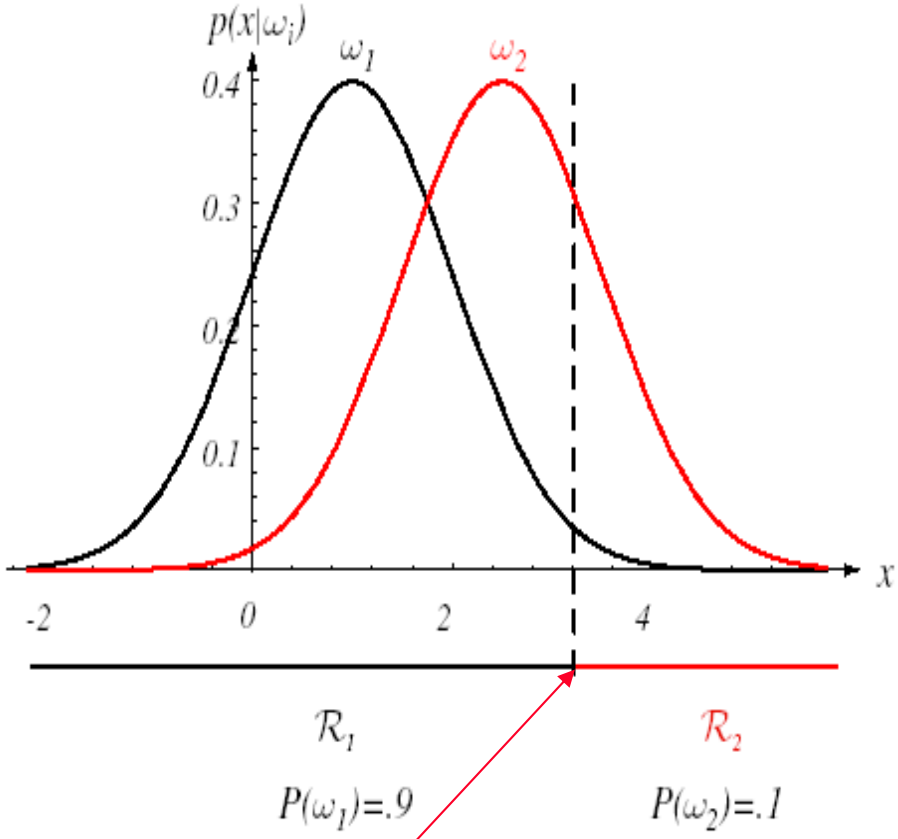
- Example of distributions with d=3

Case $\Sigma_i=\sigma^2 I$ with different a priori probabilities

- As shown in the following slides, changing the prior probabilities changes decision surfaces.
- For prior probabilities quite different, decision boundaries do not lie in between the means of the distributions
- In the next slides, examples for one, two and three dimensions are shown

Example $\Sigma_i=\sigma^2 I$ with different *a priori* probabilities



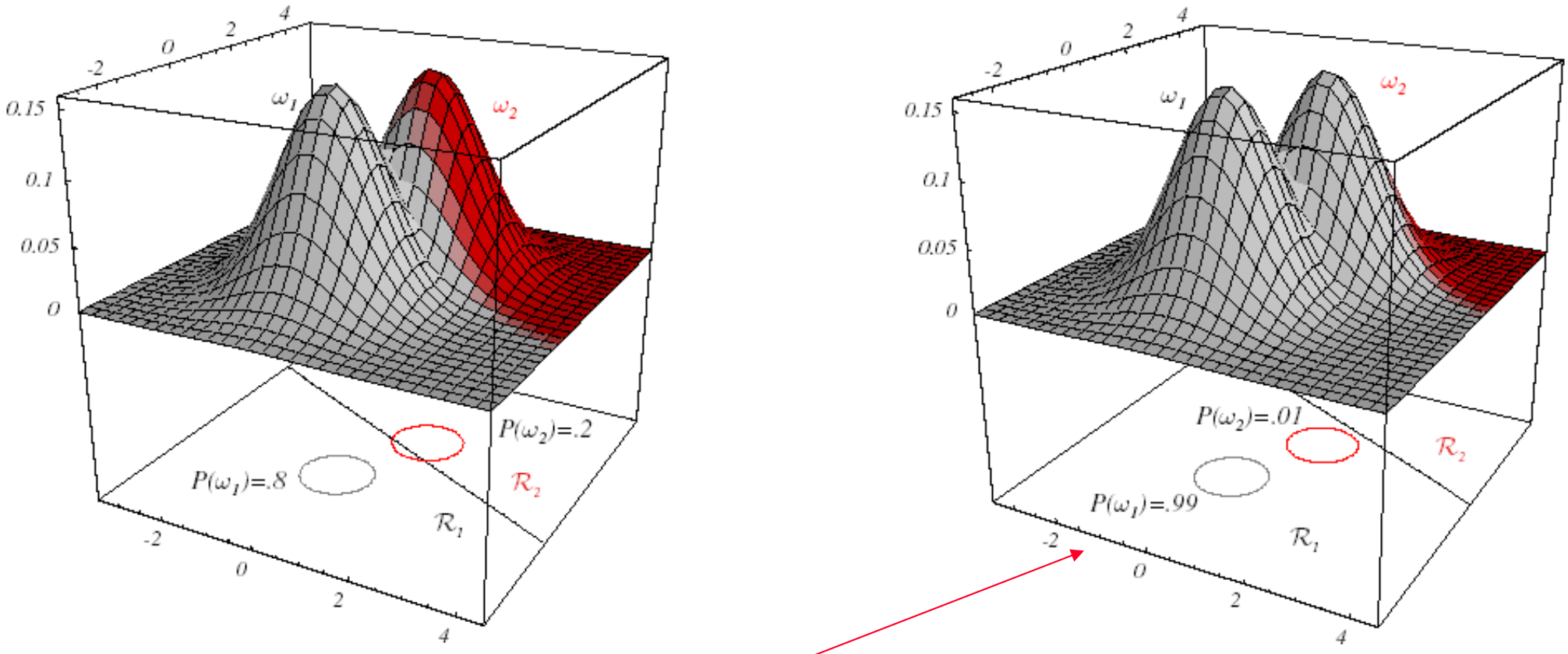


➢If the prior probabilities are different the boundary point x_0 "moves away" as is intuitive, from the mean of the most probable class

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

The smaller the variance (compared to the distance between the averages), the more the influence of "priors" decreases

Example $\Sigma_i=\sigma^2 I$ with different *a priori* probabilities



Extreme case where the big difference between $P(\omega_i)$ brings me to choose almost always for ω_1

It is clear that this is a problem when I have to recognize classes "rare" (a rare disease, intrusive traffic, "spamming")

Gaussian model: case $\Sigma_i = \Sigma$

- In this case, the covariance matrices are equal (but arbitrary) for all classes.
- The pattern form hyper-ellipsoidal "clusters" of identical size and shape, centred in μ_i
- Deleting from the discriminant function

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

all the terms that do not depend on *i*, we can write:

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

Gaussian model with $\Sigma_i = \Sigma$ and $P(\omega_i) = P(\omega_j)$

- Special case: $P(\omega_i) = P(\omega_j)$ for each class.

- In this case, the discriminant function becomes:

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

Mahalanobis distance

- The decision rule will be:
  - Given $\mathbf{x}$, we measure the Mahalanobis distance between $\mathbf{x}$ and any μ_i, and assign $\mathbf{x}$ to the class to minimum distance

As in the previous case with Σ diagonal, expanding and eliminating the terms independent on "*i*" we obtain the *linear* function:

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

with:

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

Case $\Sigma_i = \Sigma$: Decision Surfaces

- The separation surfaces between adjacent regions R_i and R_j are hyperplanes of equation

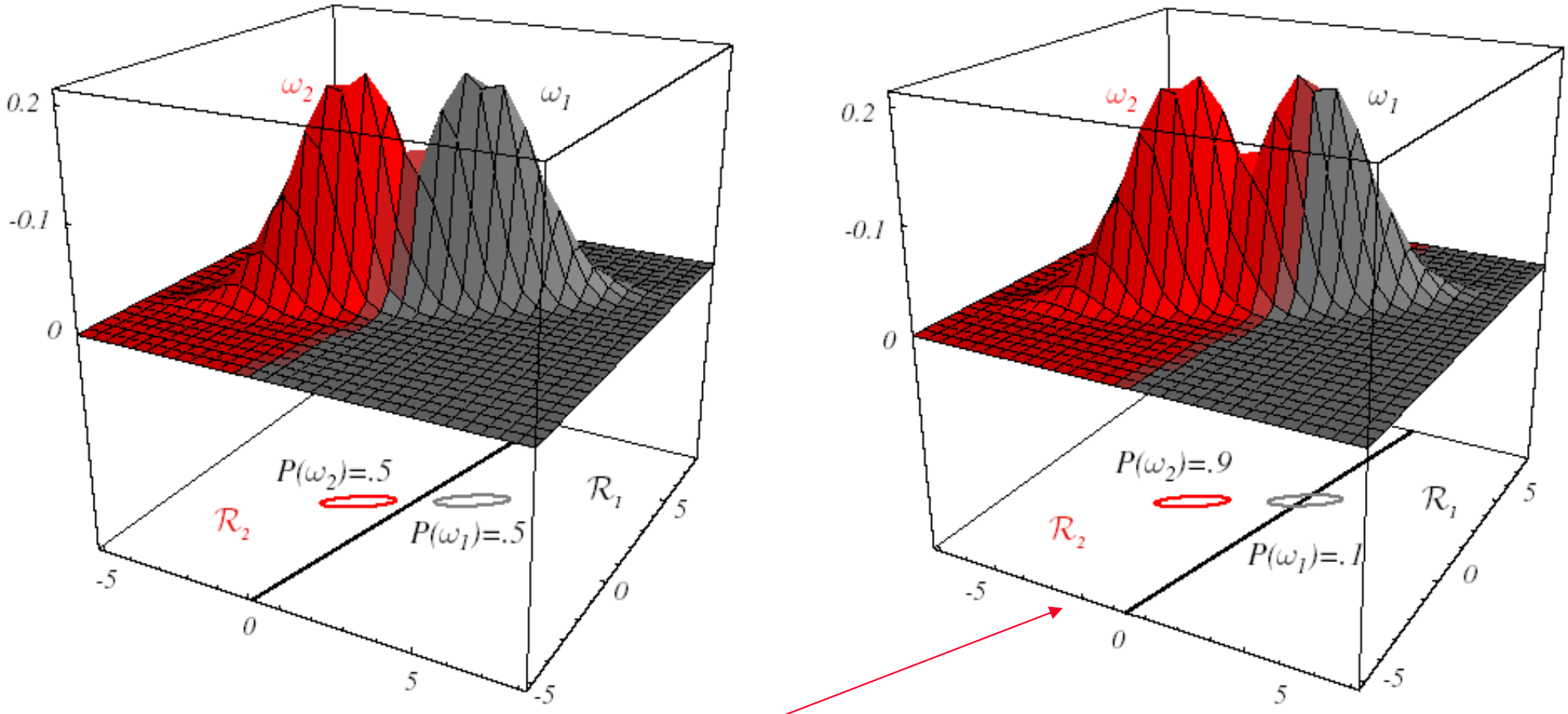
$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0$$

dove

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

- Since $\mathbf{w}$ is not (in general) along the direction of $(\mu_i - \mu_j)$, the hyper-plane is not orthogonal to the line joining the two means.
- However, the hyper-plane intersects this line in $\mathbf{x}_0$; the position of $\mathbf{x}_0$ depends on the *a priori* probability

Example case $\Sigma_i = \Sigma$



- Examples of decision regions for equal normal distributions with very different prior probabilities

Estimation of the covariance matrix if $\Sigma_i = \Sigma$

If the covariance matrix is the same for all classes, it can be shown that its maximum likelihood estimate is:

$$S_w = \sum_{i=1}^{C} \frac{n_i}{n} \hat{\Sigma}_i$$

pooled within-group sample covariance matrix

Where Σ_i is estimated with the samples belonging to class ω_i.

The unbiased estimate of the matrix S_w (for "c" classes) is:

$$\frac{n}{n-c} S_w$$

Gaussian Model: arbitrary Σ_i

- In this case, the only term that we can drop from the discriminant function is (d/2)ln(2π), obtaining:

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

- $g_i(\mathbf{x})$ is a *quadratic function* that can be written as

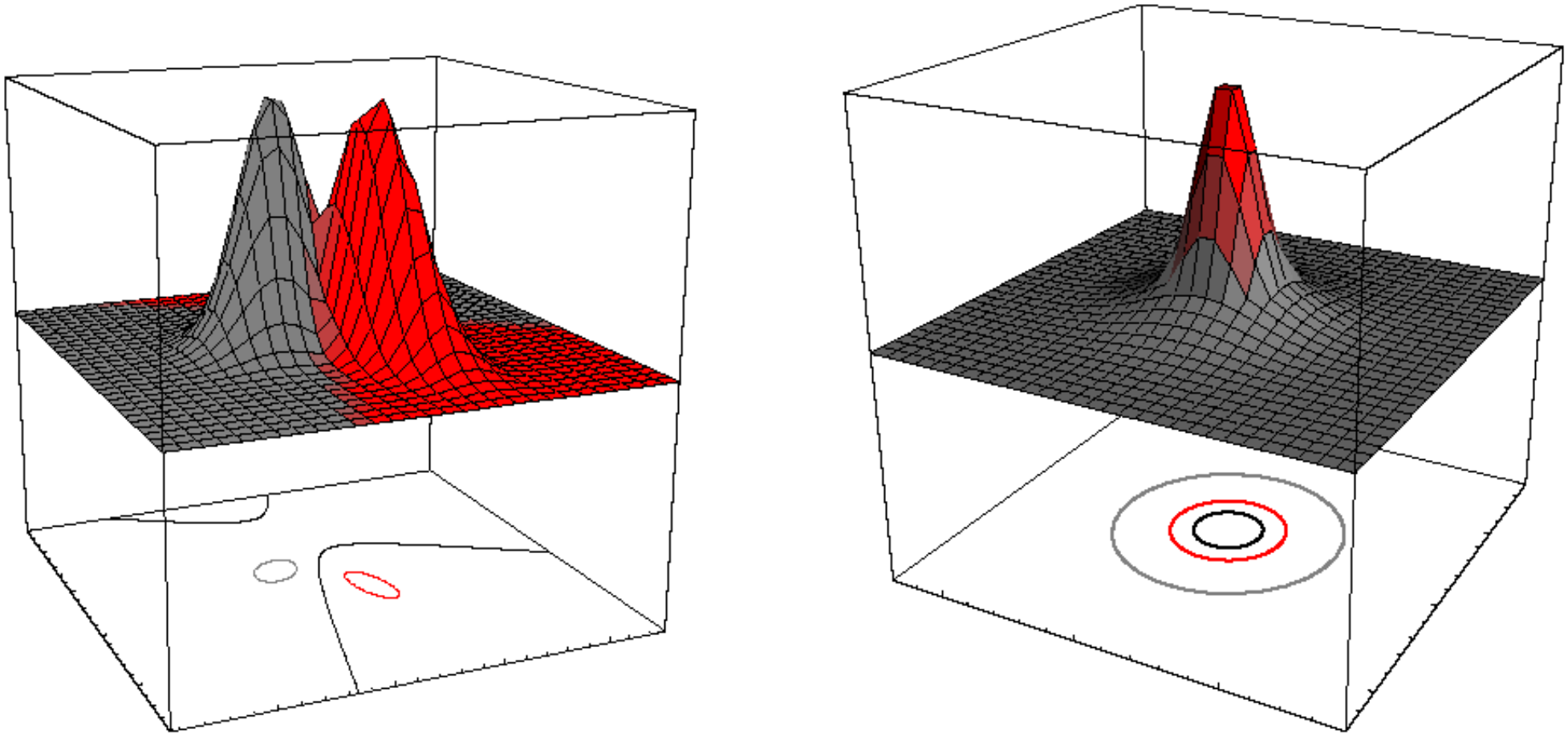
$$g_i(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{W}_i\boldsymbol{x} + \boldsymbol{w}_i'\,\boldsymbol{x} + w_{i0}$$

where

$$\boldsymbol{W}_i = -\frac{1}{2}\,\boldsymbol{\Sigma}_i^{-1};\; w_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

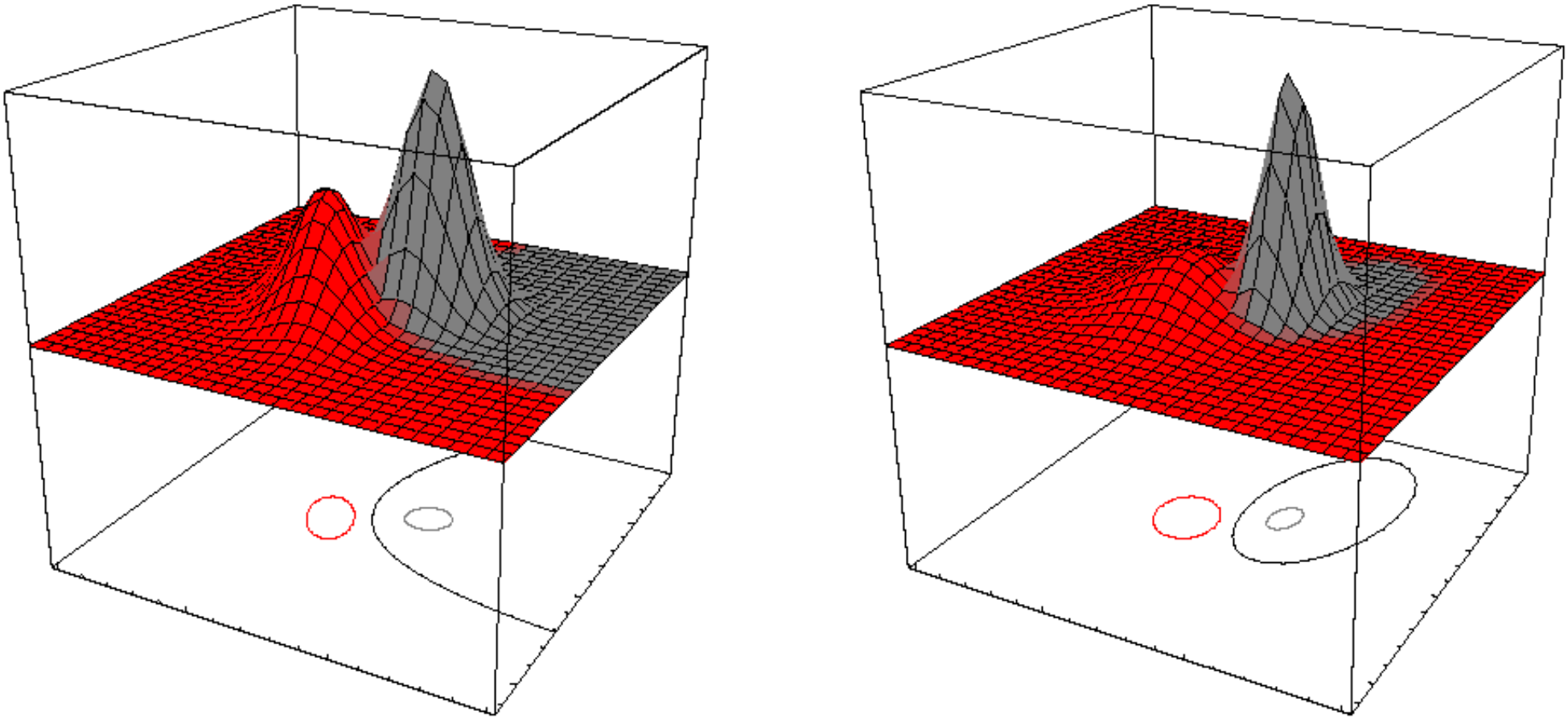
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Arbitrary Σ_i : examples



Arbitrary distributions and hyperquadrics decision surfaces

Arbitrary Σ_i : examples



Arbitrary distributions and hyperquadrics decision surfaces

Homework 7 - send me your answer to roli@unica.it

Consider a two-class problem in $\mathbb{R}^2$ (two-dimensional feature space). Each pattern is characterized by a numerical feature vector in R^2 :

$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Each class has a Gaussian probability density function:

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\,\pi)^{d/2}\,|\Sigma_\mathrm{i}|^{1/2}} \exp\left[-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu_i})^T \Sigma_\mathrm{i}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu_i})\right]$$

with the following values:

$\mathrm{P}(\omega_1) = \mathrm{P}(\omega_2)$; $\Sigma_1 = \Sigma_1 = \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$; $\boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Homework 7 - send me your answer to roli@unica.it

Using the MAP decision rule:

a) Classify the pattern $x_T=(1/2, 1/3)'$ using the MAP rule by computing explicitly the decision regions, and tell to which decision region the pattern $x_T=(1/2, 1/3)'$) belongs to.

b) Classify the same pattern $x_T=(1/2, 1/3)'$ with the likelihood ratio test.

Homework 8 - send me your answer to roli@unica.it

Let us consider a **three-class** problem, with a **bi-dimensional feature space**.

a. Classify the pattern $x_t = (1,1/3)^T$ under the following assumptions:

- the probability density function of each class is Gaussian, with the same covariance matrix $\mathbf{\Sigma}=\sigma^2\mathbf{I}$, the parameter σ^2 (variance) is **unknown**
- all the classes have the same prior probability
- the mean vectors of the three classes are: $\mu_1=(0,0)^T$; $\mu_2=(0,2)^T$; $\mu_3=(2,1)^T$.

b. Now, classify the pattern $x_t =(1,1/3)^T$ with these assumptions about priors (class prior probabilities):

$$P(\omega_1)=P(\omega_2)=1/4; P(\omega_3)=1/2$$

References

- Sections 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, Pattern Classification, R.O. Duda, P. E. Hart, and D. G. Stork, John Wiley & Sons, 2000
- Chapter 1, Statistical Pattern Recognition, Andrew Webb, John Wiley & Sons, 2002
- Sections 2.6, 3.1, 3.2, 3.3, 3.9, Pattern Classification, R. O. Duda, P. E. Hart, D. G. Stork, John Wiley & Sons, 2000.
- Sections 2.1, 2.2.1, 2.2.2, 2.3, Statistical Pattern Recognition, Andrew Webb, John Wiley & Sons, 2002.