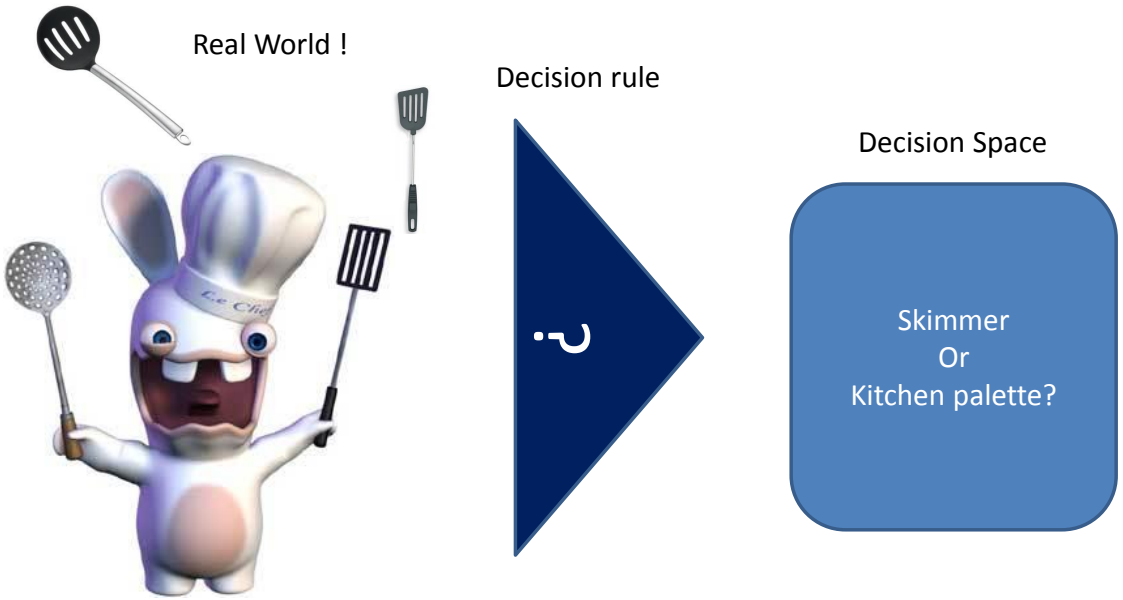
Pattern recognition and applications to monitoring

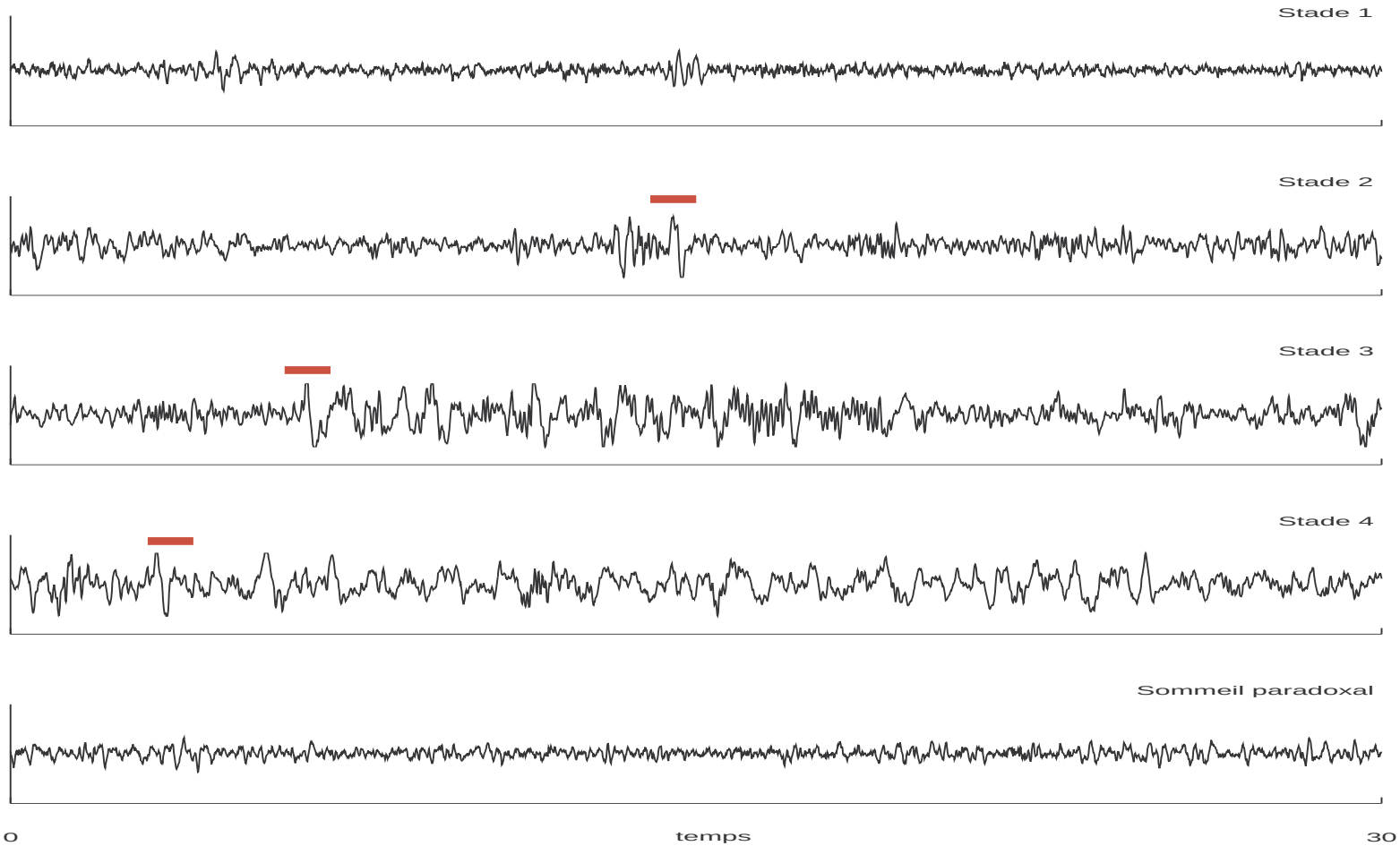
Pierre Beauseroy - Edith Grall - Paul Honeine - Régis Lengellé

– introduction –

Type of Problems



- Real world
- Data gathering
- Decision rule
- Decision space



Example : Complex K detection in EEG sleep signal.

The problem can be written as follow :

$$\begin{cases} \omega_0 : \boldsymbol{x} = \boldsymbol{b} & \text{hypothesis "noise alone"} \\ \omega_1 : \boldsymbol{x} = \boldsymbol{b} + \boldsymbol{s} & \text{hypothesis "signal plus noise"} \end{cases}$$

The aim is to build a detector d, or a decision rule d which minimizes a criteria - the error probability for example

$$P_e(d) = p(d(\boldsymbol{X}) \neq Y),$$

where $\boldsymbol{X}$ is an observation and Y its associate hypothesis.

The strategy to design a solution to this problem depends on the nature of the available information on the problem

Rule based approach

If X is grey and $weight(X) > 1000$ then
X is an elephant
Else
X is a mouse ...
$\Rightarrow$ Need an expert
$\Rightarrow$ Need to translate expert thought to rules - complex, long, most of the time not reliable.

Hypothesis testing approach

Express the problem as an hypothesis testing problem :

$$\begin{cases} H_0 : X \in \omega_0; & X \sim p\left(X \,|\omega_0\right) \\ H_1 : X \in \omega_1; & X \sim p\left(X \,|\omega_1\right) \end{cases}$$

Deduce the test.

Data driven approach

Use a set of data $\mathcal{A}$

2 classes

$$\begin{cases} H_0 : X \in \omega_0; & X \sim p\left(X \,|\omega_0\right) \\ H_1 : X \in \omega_1; & X \sim p\left(X \,|\omega_1\right) \end{cases}$$

1 class

$$\begin{cases} H_0 : X \in \omega_0 \\ H_1 : X \notin \omega_0 \end{cases}$$

Multiclass

$$\begin{cases} H_0 : X \in \omega_0 \\ H_1 : X \in \omega_1 \\ H_2 : X \in \omega_2 \\ \ldots \\ H_K : X \in \omega_K \end{cases}$$

$$\begin{cases} H_0 : X \in \omega_0; & X \sim p\left(X \,|\omega_0\right) \\ H_1 : X \in \omega_1; & X \sim p\left(X \,|\omega_1\right) \end{cases}$$

We want to determine $d(\boldsymbol{x}) = \omega_0$ if $q(\boldsymbol{x}) < s$ so that $\min_{q,s} P_e$

with :

$$P_e = P\left(D_0 \,|\omega_1\right) P\left(\omega_1\right) + P\left(D_1 \,|\omega_0\right) P\left(\omega_0\right)$$

$$P\left(D_0 \,|\omega_1\right) = \int_{\mathcal{X}} \mathbb{1}_{(q(x)<s)} p\left(x \,|\omega_1\right) dx$$

$$P\left(D_1 \,|\omega_0\right) = \int_{\mathcal{X}} \mathbb{1}_{(q(x)\geq s)} p\left(x \,|\omega_0\right) dx$$

For each x we wish to minimize :

$$r(x) = \mathbb{1}_{(q(x)<s)} p\left(x \,|\omega_1\right) P\left(\omega_1\right) + \mathbb{1}_{(q(x)\geq s)} p\left(x \,|\omega_0\right) P\left(\omega_0\right)$$

$$P_e = r(\mathcal{D}) = \int_{\mathcal{X}} r(x) dx$$

We choose the class ω_i such that :

$$p\left(x\left|\omega_i\right.\right) P\left(\omega_i\right) \overset{D_i}{>} p\left(x\left|\omega_j\right.\right) P\left(\omega_j\right)$$

Thus the decision rule is :

$$p\left(x\left|\omega_1\right.\right) P\left(\omega_1\right) \underset{D_1}{\overset{D_0}{\lessgtr}} p\left(x\left|\omega_0\right.\right) P\left(\omega_0\right)$$

or

$$\frac{p\left(x\left|\omega_1\right.\right)}{p\left(x\left|\omega_0\right.\right)} \underset{D_1}{\overset{D_0}{\lessgtr}} \frac{P\left(\omega_0\right)}{P\left(\omega_1\right)}$$

where
$q(x) = \frac{p(x|\omega_1)}{p(x|\omega_0)}$ is the likelihood ratio
and
$s = \frac{P(\omega_0)}{P(\omega_1)}$

Multiclass case

$$d(\boldsymbol{x}) = \arg\min_i \left(\sum_{j \neq i} P\left(D_i \,|\omega_j\,\right) P\left(\omega_j\right) \right)$$

Criteria is a cost function

We want to determine $d(\boldsymbol{x}) = \omega_0$ if $q(\boldsymbol{x}) < s$ so that $\min_{q,s} r(\mathcal{D})$

where

$$r(\mathcal{D}) = \sum_{i,j=1}^{K} c_{ij} P\left(D_i \,|\omega_j\,\right) P\left(\omega_j\right)$$

$$d(x) = \arg\min_i \left(\sum_j c_{ij} P\left(D_i \,|\omega_j\,\right) P\left(\omega_j\right) \right)$$

In the 2 class case :

$$\frac{p\left(x \,|\omega_1\,\right)}{p\left(x \,|\omega_0\,\right)} \underset{D_1}{\overset{D_0}{\lessgtr}} \frac{\left(c_{10} - c_{00}\right) P\left(\omega_0\right)}{\left(c_{01} - c_{11}\right) P\left(\omega_1\right)}$$

We want to define α, the first order error $P(D_1|\omega_0)$. So the problem is :
$\min_{q,s} P\left(D_0\left|\omega_1\right.\right)$
with :
$P\left(D_1\left|\omega_0\right.\right) = \alpha$
Let assume that solving

$$\min_{q,s} \max_{\mu} P\left(D_0\left|\omega_1\right.\right) + \mu\left(P\left(D_1\left|\omega_0\right.\right) - \alpha\right)$$

solves the initial problem.
And :

$$\begin{aligned} & P\left(D_0\left|\omega_1\right.\right) + \mu\left(P\left(D_1\left|\omega_0\right.\right) - \alpha\right) \\ & = \int_{\mathcal{X}} \mathbb{1}_{(q(x)<s)} p\left(x\left|\omega_1\right.\right) + \mu\left(\mathbb{1}_{(q(x)\geq s)} p\left(x\left|\omega_0\right.\right)\right) dx - \mu\alpha \end{aligned}$$

Thus, for a given μ the decision rule is :

$$p\left(x\left|\omega_1\right.\right) \underset{D_1}{\overset{D_0}{\lessgtr}} \mu p\left(x\left|\omega_0\right.\right)$$

or

$$\frac{p\left(x\left|\omega_1\right.\right)}{p\left(x\left|\omega_0\right.\right)} \underset{D_1}{\overset{D_0}{\lessgtr}} \mu$$

which is a likelihood ratio test (as in Bayes case).

Free structure detection

Based on simple hypothesis, applying decision criterion such as Bayes' cost leads to Bayes' rule :

$$d^*(\boldsymbol{x}) = \begin{cases} \omega_1 & \text{si} \quad p(\boldsymbol{x}|\omega_1)/p(\boldsymbol{x}|\omega_0) \geq \lambda_0 \\ \omega_0 & \text{sinon,} \end{cases}$$

$p(\boldsymbol{x}|\omega_0)$ and $p(\boldsymbol{x}|\omega_1)$ have to be known. The threshold λ_0 is the only parameter of the rule. It depends on the criterion one wishes to optimize.

Thus the decision rule or the detector is not subject to any structural constraints but depends only on the choice of a criteria.

$$\begin{cases} P(X|\omega_0) \sim N(M_0, \Sigma) \\ P(X|\omega_1) \sim N(M_1, \Sigma) \end{cases}$$

and $P(\omega_0) = P(\omega_1) = \frac{1}{2}$
Thus :

$$P(X|\omega_i) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(X-M_i)^T \Sigma^{-1}(X-M_i)}$$

So :

$$\log(q(X)) = \frac{1}{2}(X-M_0)^T \Sigma^{-1}(X-M_0) - \frac{1}{2}(X-M_1)^T \Sigma^{-1}(X-M_1)$$

And the decision rule is :

$$(M_1-M_0)^T \Sigma^{-1} X + \frac{1}{2} M_0^T \Sigma^{-1} M_0 - \frac{1}{2} M_1^T \Sigma^{-1} M_1 \underset{D_1}{\overset{D_0}{\lessgtr}} 0$$

Which is equivalent to

$V^T X + U \underset{D_1}{\overset{D_0}{\lessgtr}} 0$ a linear function !

$$\begin{cases} P(X|\omega_0) \sim N(M_0, \Sigma_0) \\ P(X|\omega_1) \sim N(M_1, \Sigma_1) \end{cases}$$

and let $P(\omega_0) = P(\omega_1) = \frac{1}{2}$
Thus :

$$\log(q(X)) = \frac{1}{2}\log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) + \frac{1}{2}(X - M_0)^T \Sigma_0^{-1}(X - M_0) - \frac{1}{2}(X - M_1)^T \Sigma_1^{-1}(X - M_1)$$

So

$$\frac{1}{2}X^T\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right)X - M_0^T\Sigma_0^{-1}X + M_1^T\Sigma_1^{-1}X + \frac{1}{2}M_0^T\Sigma_0^{-1}M_0$$
$$-\frac{1}{2}M_1^T\Sigma_1^{-1}M_1 + \frac{1}{2}\log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \underset{D_1}{\overset{D_0}{\lessgtr}} 0$$

Which is equivalent to :

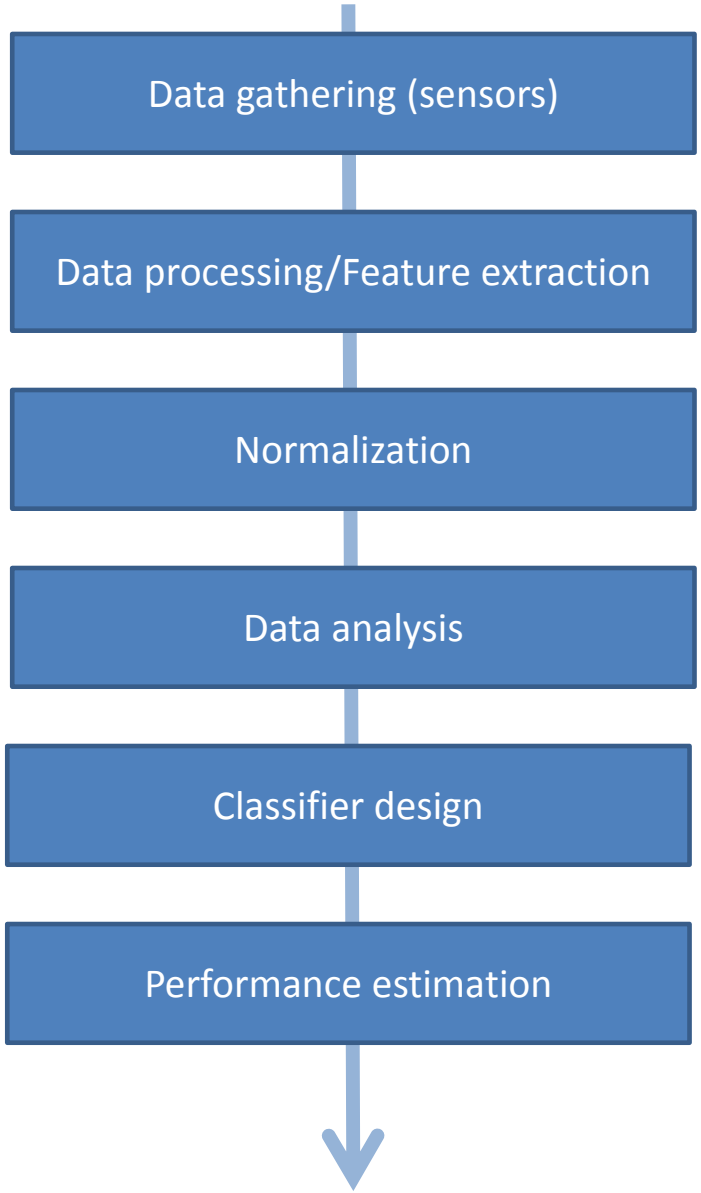
$X^T T X + V^T X + U \underset{D_1}{\overset{D_0}{\lessgtr}} 0$ a quadratic function !

To use these approaches $p(\boldsymbol{x}|\omega_0)$ and $p(\boldsymbol{x}|\omega_1)$ have to be known...

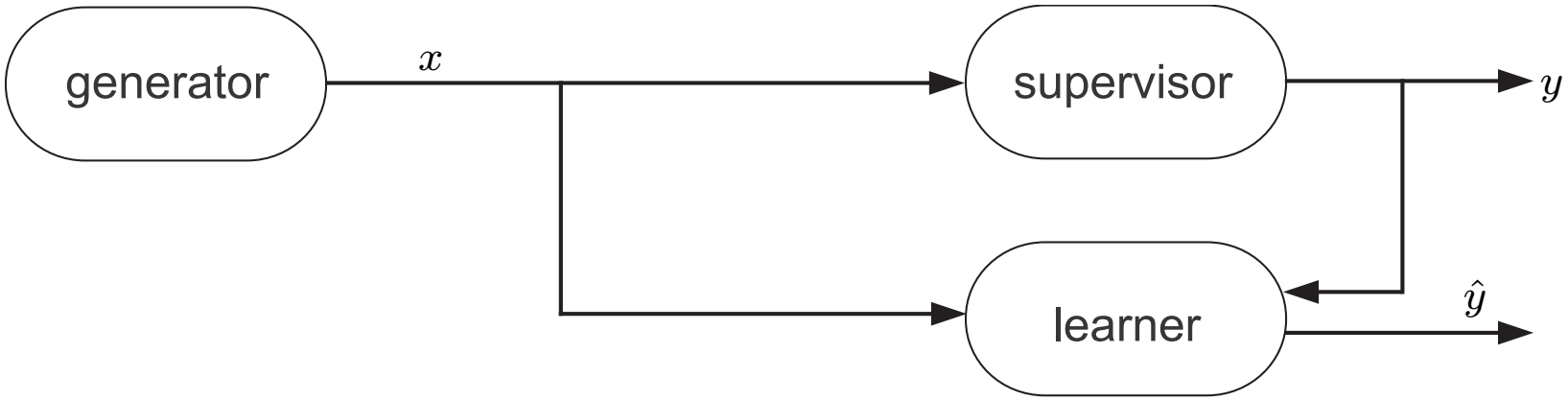
If not the case ?

can assume that
$p(\boldsymbol{x}|\omega_0)$ and $p(\boldsymbol{x}|\omega_1) \subset \mathcal{F}_\theta$

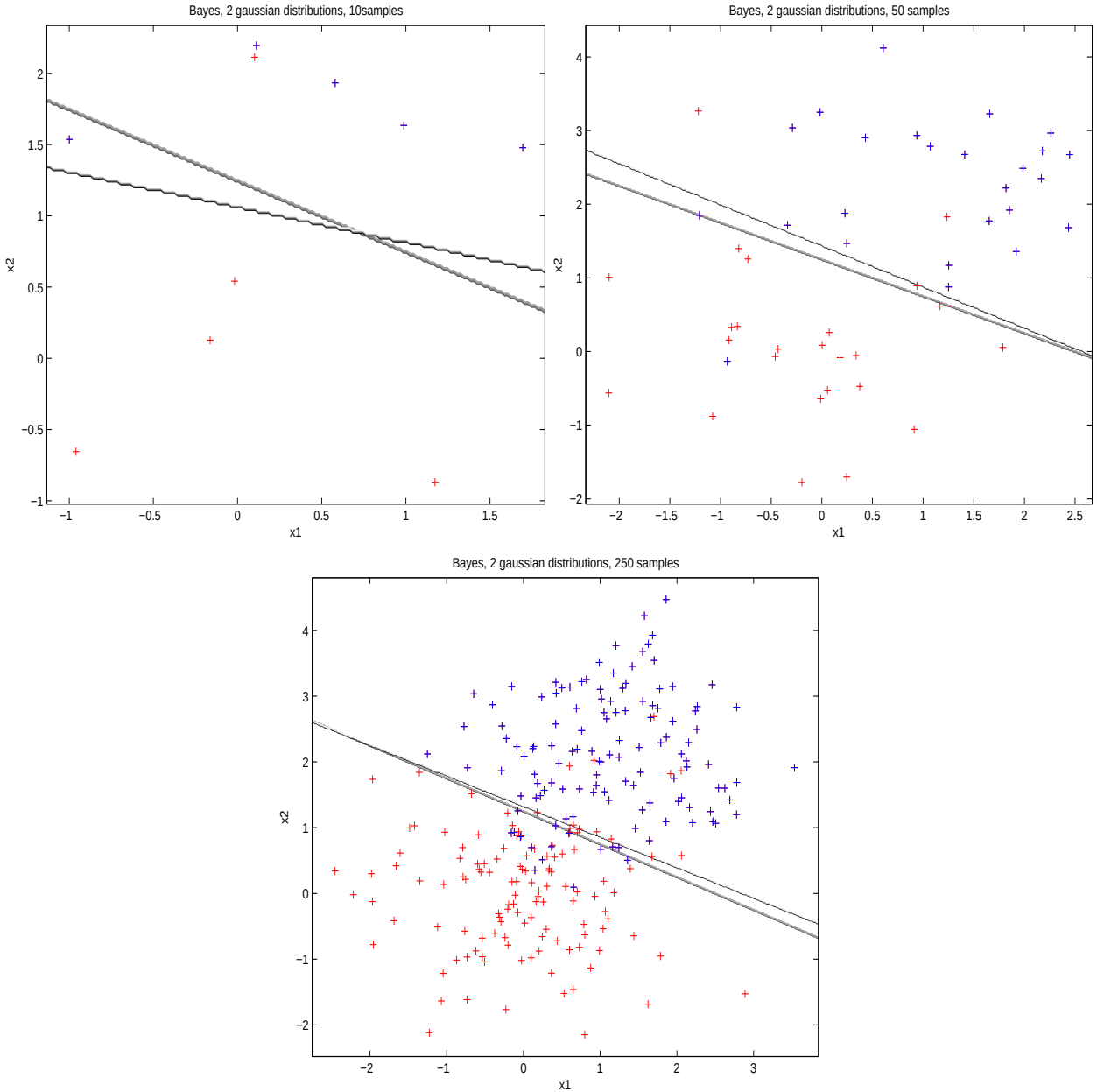
then select or estimate θ based on data
and plugging the estimator in the decision rule.

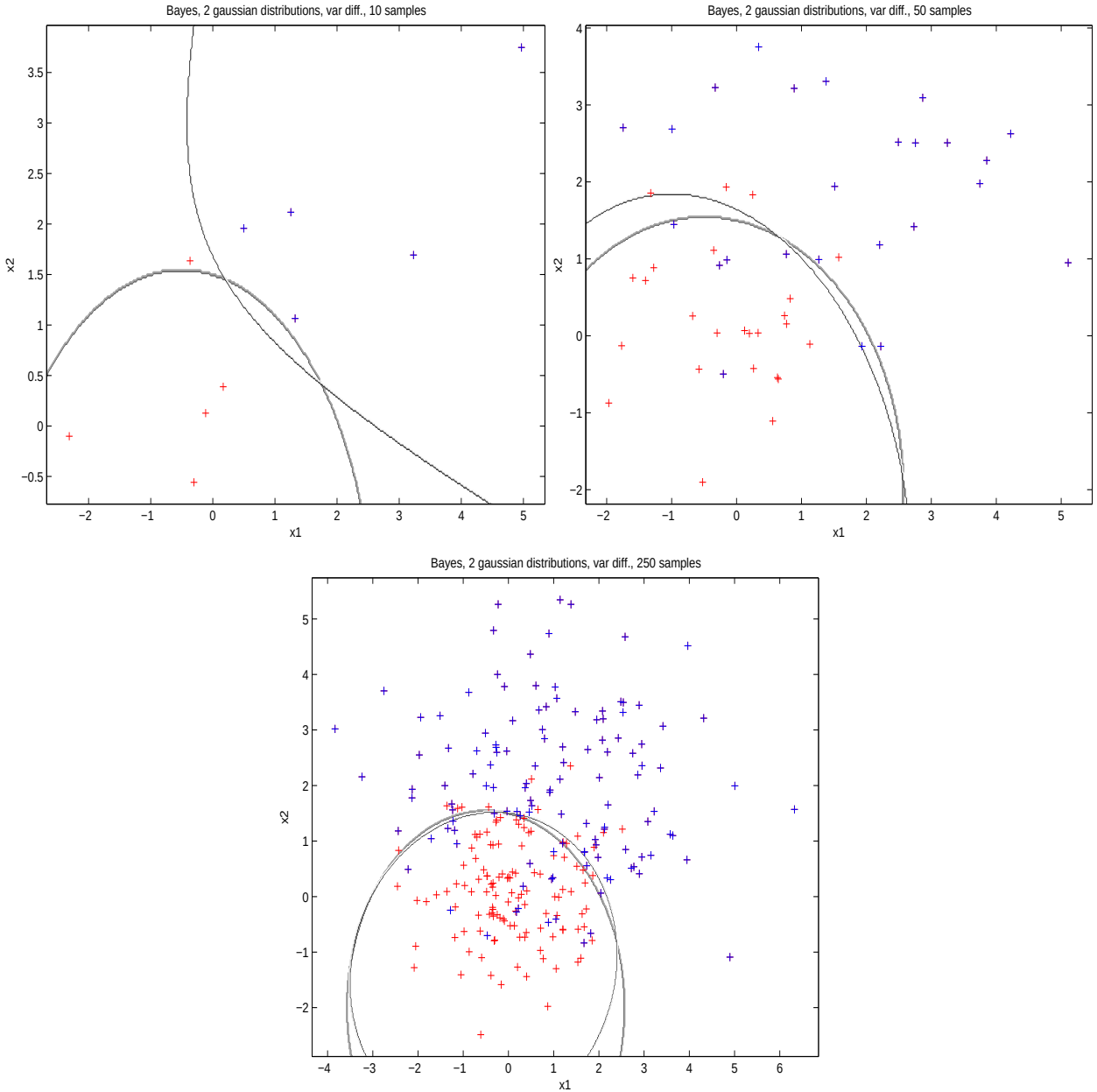

Data gathering (sensors)
Data processing/Feature extraction
Normalization
Data analysis
Classifier design
Performance estimation

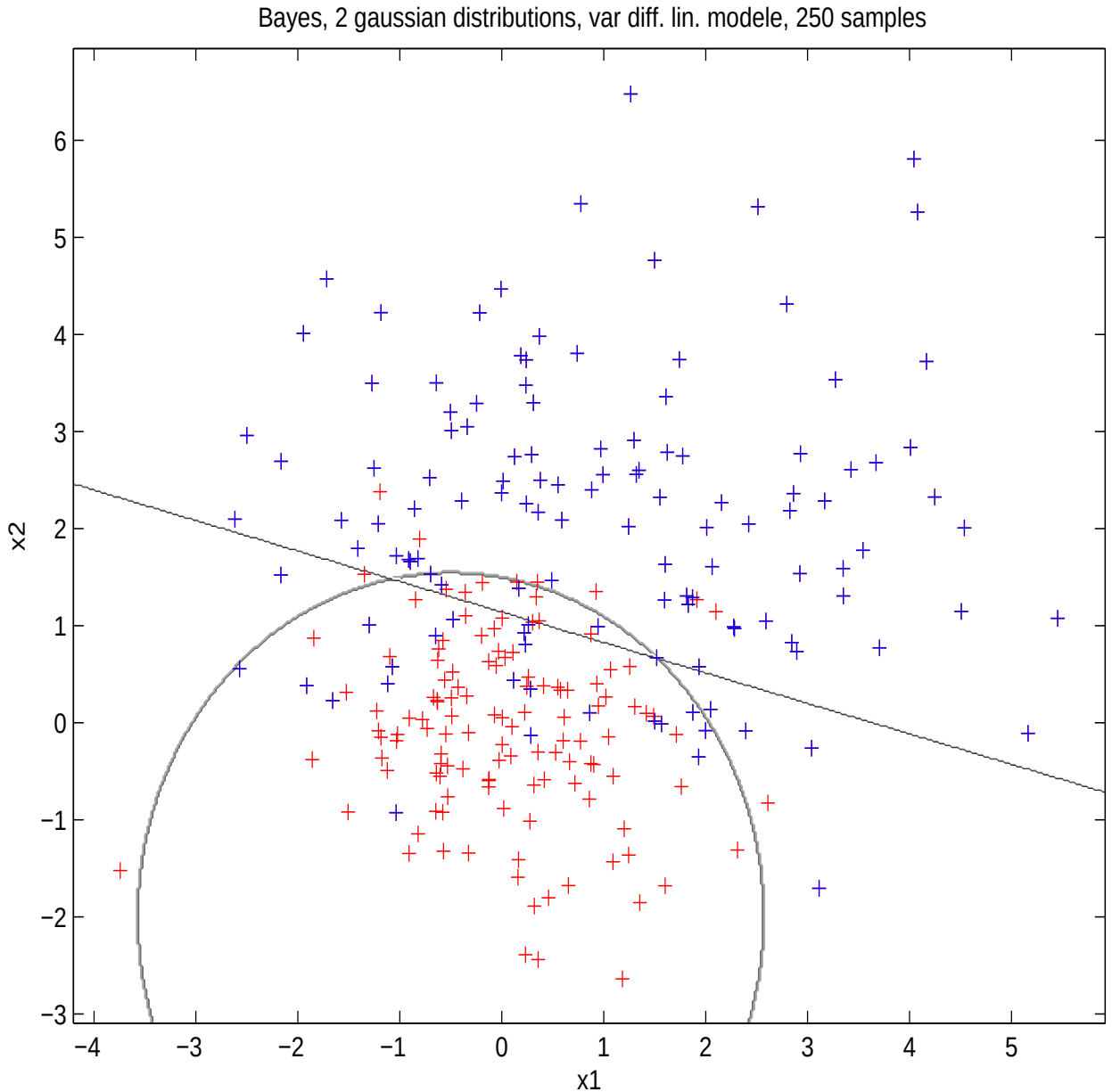
The learning model is composed of 3 elements :



1. Generator : $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^l$, random variables i.i.d.
2. Supervisor : $Y \in \mathcal{Y} \subset \mathbb{R}$, random variables
3. Learner : represented by $d(\boldsymbol{x}; \theta) \in \mathcal{D}$


Bayes, 2 gaussian distributions, 10samples
x1
x2
Bayes, 2 gaussian distributions, 50 samples
x1
x2
Bayes, 2 gaussian distributions, 250 samples
x1
x2


Bayes, 2 gaussian distributions, var diff., 10 samples
Bayes, 2 gaussian distributions, var diff., 50 samples
Bayes, 2 gaussian distributions, var diff., 250 samples
x1
x2


Bayes, 2 gaussian distributions, var diff. lin. modele, 250 samples
x1
x2
−4
−3
−2
−1
0
1
2
3
4
5
6
−3

Avoid to use strong assumption about data distribution.

Main methods :

1. Parzen density estimator,
2. k-nearest neighbor estimator.

Principle

Draw N samples from $p(X)$, count the proportion of them $\frac{n(X)}{N}$ that fall in a given domain of volume ν.

The volume ν is defined a priori.

Estimator

$$\hat{p}(X) = \frac{n(X)}{\nu N}$$

or

$$\hat{p}(X) = \frac{1}{N}\sum_{i=1}^{N} K(X - X_i) = \frac{1}{N}\sum_{i=1}^{N} \delta(X - X_i) * K(X)$$

with :

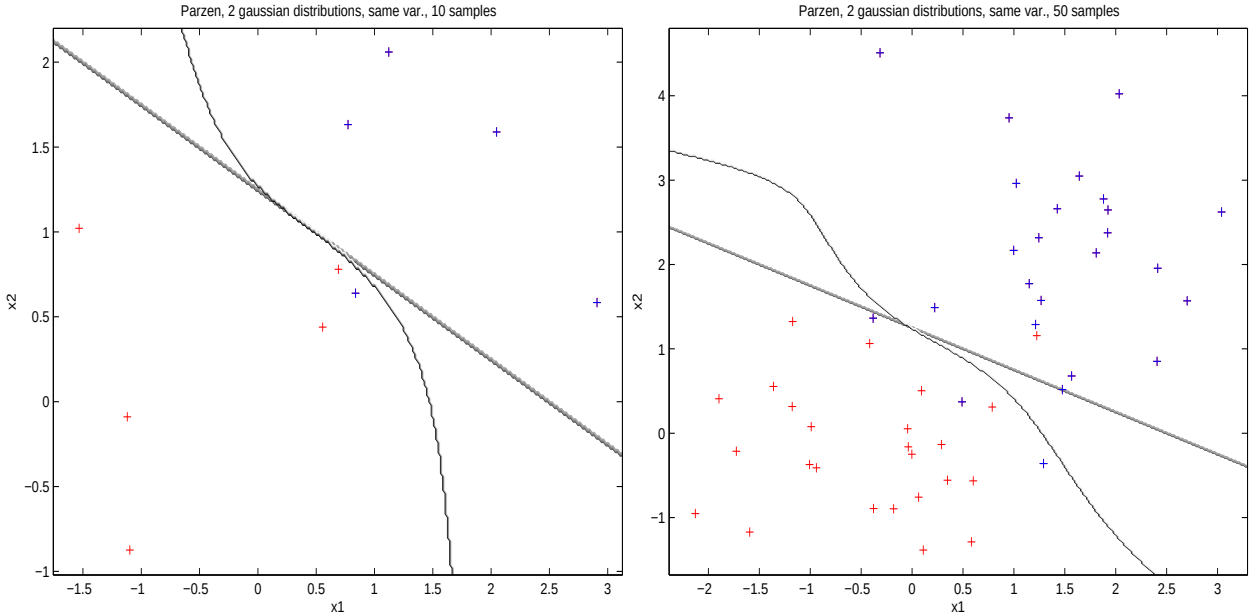
$$\int K(x)dx = 1$$

Usual kernels :

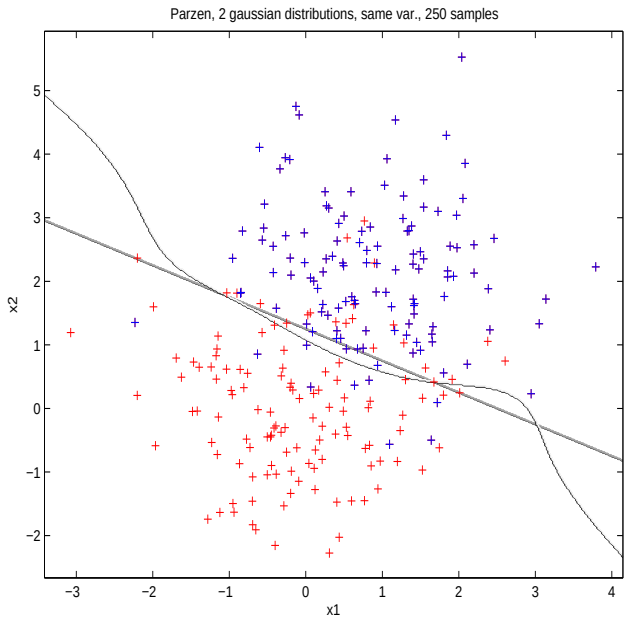
- Uniform kernel
- Gaussian kernel

$\hat{p}(X)$ depends on kernel parameters.

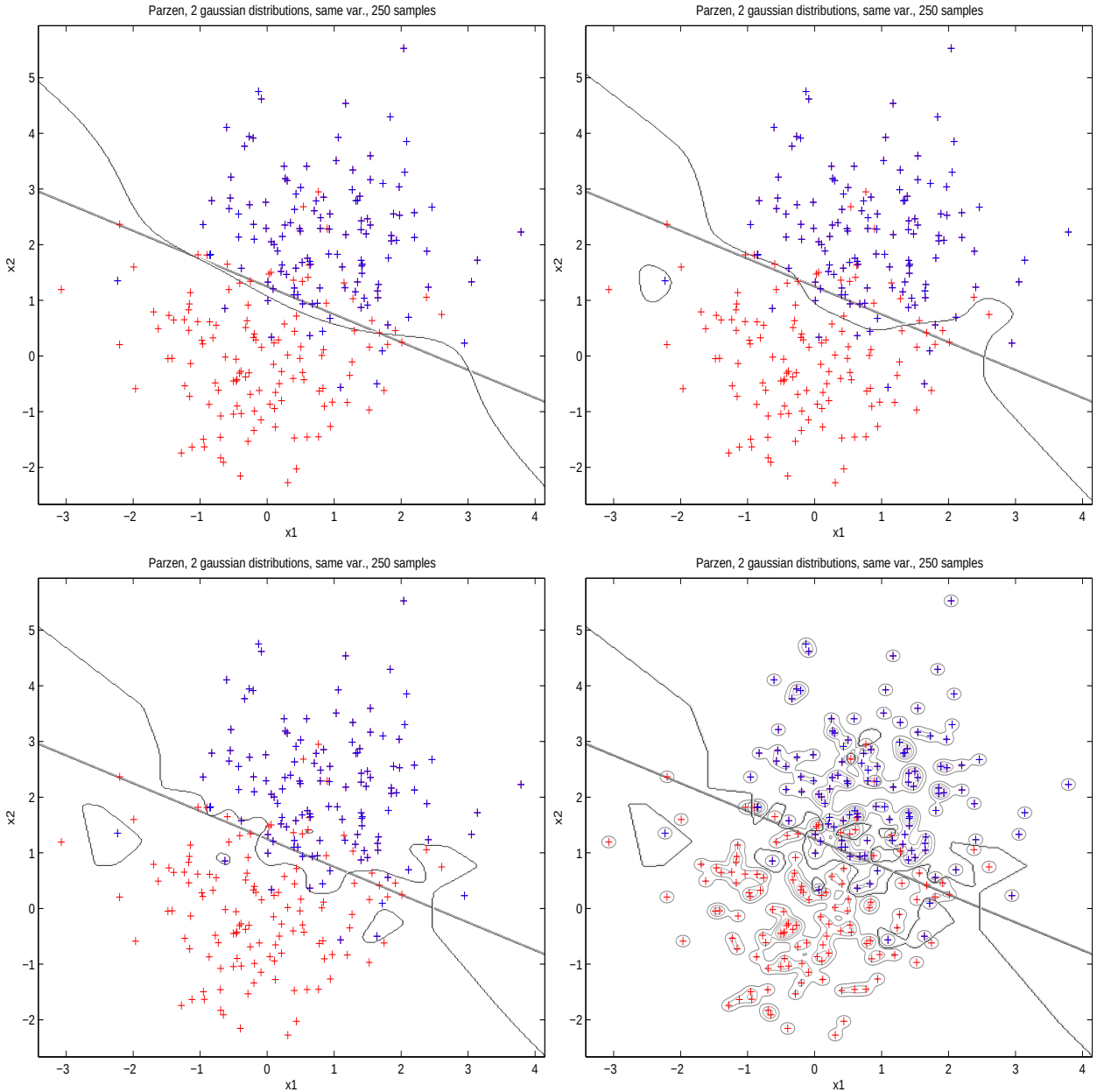
Examples (1)


Parzen, 2 gaussian distributions, same var., 10 samples
x1
x2

Parzen, 2 gaussian distributions, same var., 50 samples
x1
x2


Parzen, 2 gaussian distributions, same var., 250 samples
x1
x2

Examples (2)


Parzen, 2 gaussian distributions, same var., 250 samples
x1
x2
Parzen, 2 gaussian distributions, same var., 250 samples
x1
x2
Parzen, 2 gaussian distributions, same var., 250 samples
x1
x2
Parzen, 2 gaussian distributions, same var., 250 samples
x1
x2

Estimator

$$\hat{p}(X) = \frac{k-1}{V(X)N}$$

Decision

Two options :

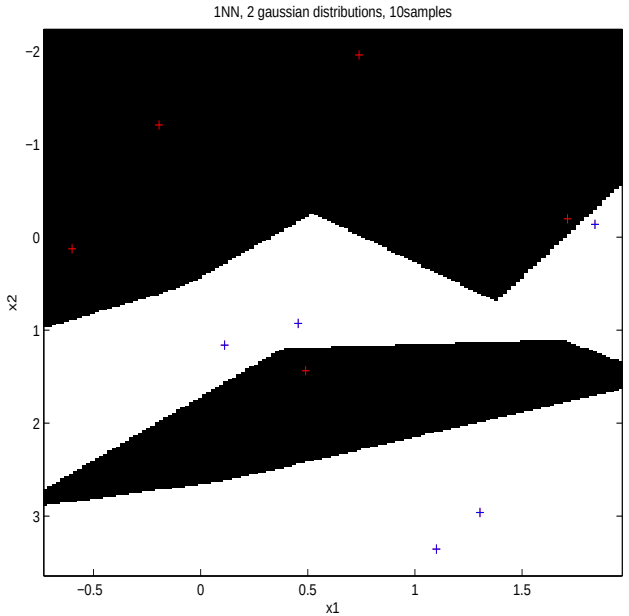
1. Compare the volume of kNN of the différent classes.
2. Compare the number of neighbors of each class among the kNN (most popular form)

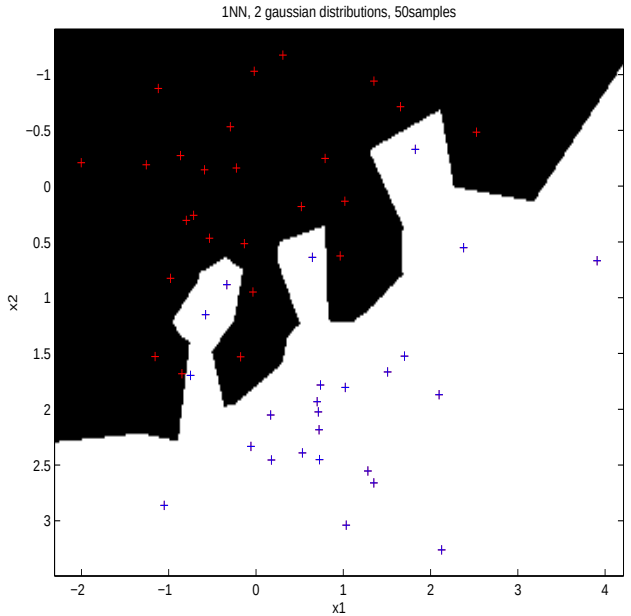
Asymptotic property

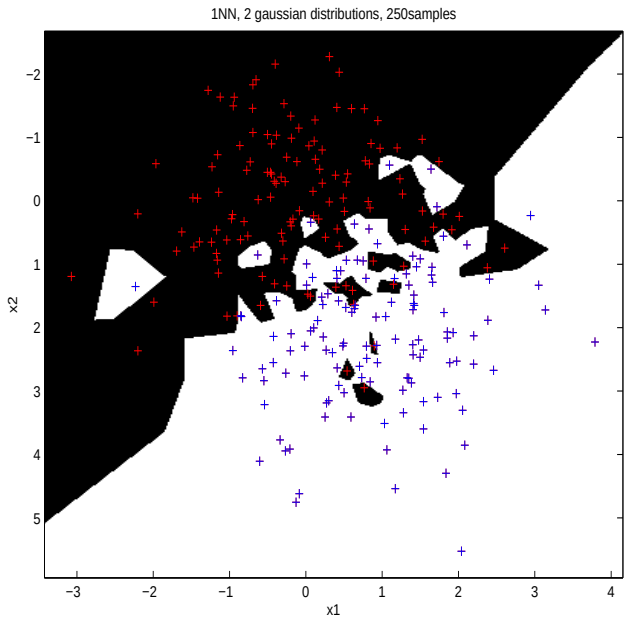
when n goes to infinity

$$r^*(X) \leq r_{kNN}(X) \leq 2r^*(X)$$

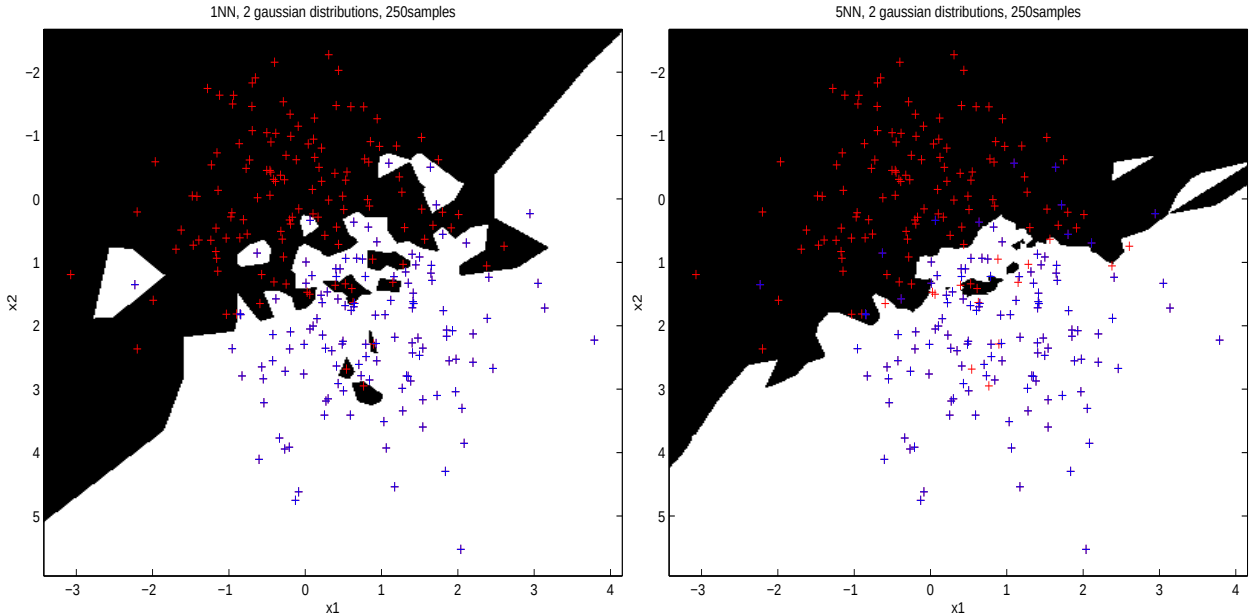
Examples

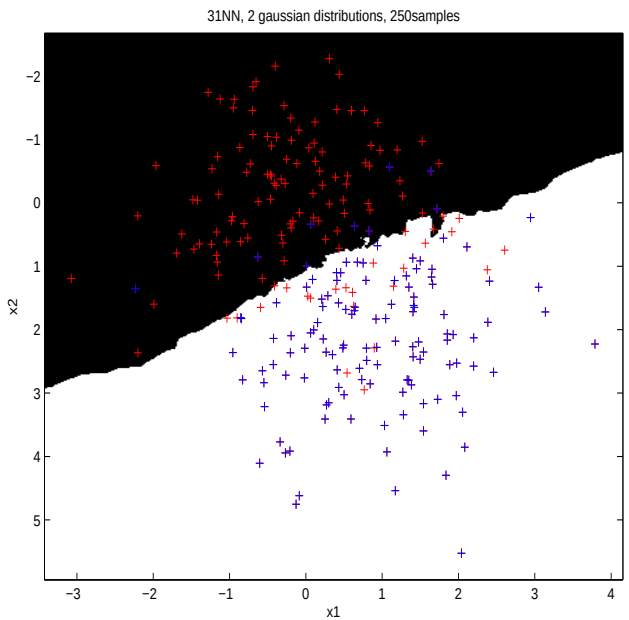

1NN, 2 gaussian distributions, 10samples
x1
x2


1NN, 2 gaussian distributions, 50samples
x1
x2


1NN, 2 gaussian distributions, 250samples
x1
x2

Examples


1NN, 2 gaussian distributions, 250samples
5NN, 2 gaussian distributions, 250samples
x1
x2


31NN, 2 gaussian distributions, 250samples
x1
x2

Imposed structure detection

Without knowledge on conditional distributions of the samples one has to choose another approach which can be :

1. define a detector class $\mathcal{D} = \{d(\boldsymbol{x}, \theta) \ : \ \theta \in \Theta\}$
2. select the detector in $\mathcal{D}$ which is the best according to a given criteria

Simple at first sight, the implementation of this approach is based on our ability to answer properly to the following questions :

1. How to choose and to define the detector class $\mathcal{D}$?
2. What criterions are pertinent to select a detector in $\mathcal{D}$?
3. How to explore $\mathcal{D}$ in order to find the best detector ?

Chapitre 1 *: Statistical learning theory - Basic concepts*
→ *fonctionnel training, consistency, generalization capability, . . .*

Chapitre 2 *: Regularization*
→ *Well posed and ill posed problems, Tiknonov regularization, . . .*

Chapitre 3 *: Kernel methods*
→ *RKHS, Mercer's condition, kernelization example, . . .*

Chapitre 4 *: Support Vector Classifiers*
→ *Primal problem, Lagrangian, . . .*

Chapitre 5 *: Support Vector Regression*
→ *Primal problem, Lagrangian, . . .*

Chapitre 6 *: Support Vector one class*
→ *Primal problem, Lagrangian, . . .*

Chapitre 7 *: Support Vector multi-class*
→ *Primal problem, Lagrangian, . . .*

Chapitre 8 *: Semi-supervised classification*
→ *clustering methods, Generative methods, . . .*