Statistical learning theory for scientific applications: an overview

J. Vega

Laboratorio Nacional de Fusión (CIEMAT)

jesus.vega@ciemat.es

Outline

- Machine learning concepts
- Classification
- New paradigm for supervised classification
- Regression
- Applications in nuclear fusion

Machine learning

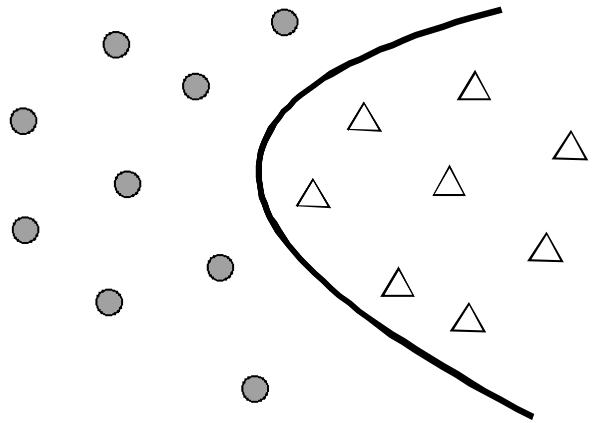
- Data-driven models find relationships among quantities whose formulation cannot be deduced from first principles
  - Nuclear fusion plasmas: disruption prediction and L/H transitions

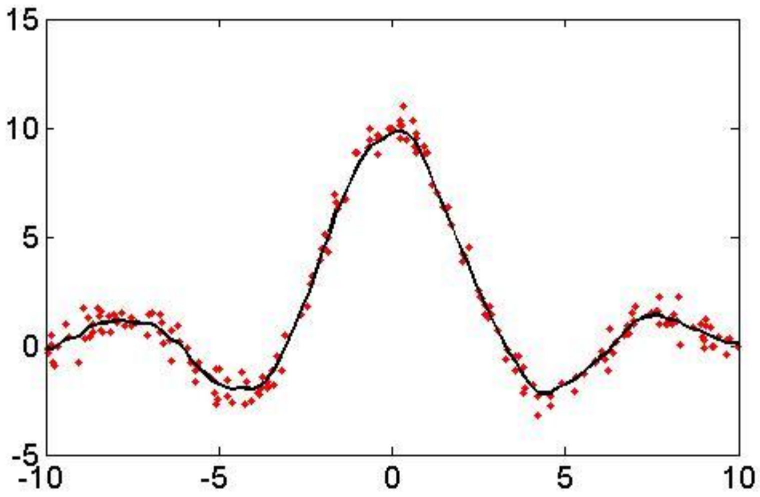
System to be modelled → Samples are observed with the objective of finding **equations** (perhaps in very complex forms) to make predictions

- Main hypothesis: samples are independent and identically distributed
  - iid hypothesis
  - Independent: each sample neither is consequence of a previous sample nor has influence in a future one
  - Identically distributed: all samples belong to the same distribution (typically unknown)

Machine learning

- Equations can be used for two purposes
  - To determine separation frontiers between different behaviours (classification problem)
  - To find the association of one or more independent variables with a dependent variable (regression problem)

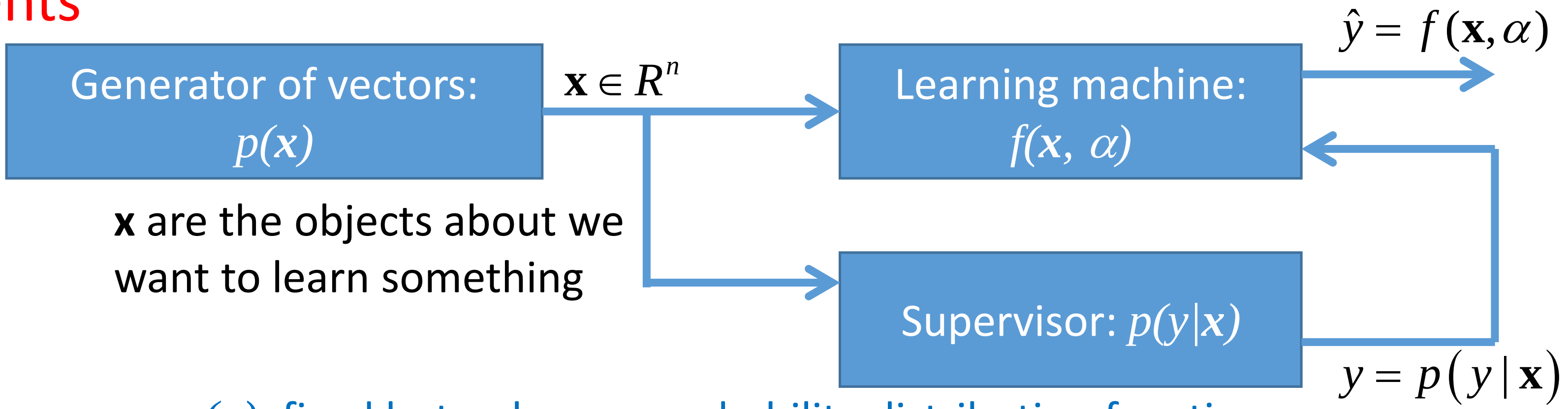




- Data-driven: there is no theory to describe the system to be modelled

Mathematical description

- The general model of *learning from examples* is described through three components



$p(\mathbf{x})$: fixed but unknown probability distribution function
$y = p(y|\mathbf{x})$ (fixed and unknown)
$f(\mathbf{x}, \alpha)$: α indicates an index in the class of functions considered
$(\mathbf{x}_i, y_i), i = 1, \ldots, N$: training samples

- The problem of learning is that of choosing from the given set of functions $f(\mathbf{x}, \alpha)$, the one that best approximates the supervisor's response

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

$$\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\} \text{ "close" to } \{y_1, y_2, \ldots, y_N\}$$

Mathematical description

- Main hypothesis
  - The training set, $(\mathbf{x}_i, y_i), i = 1, \ldots, N$, is made up of independent and identically distributed (*iid*) observations drawn according to $p(\mathbf{x}, y) = p(y|\mathbf{x})p(\mathbf{x})$
- Loss function: $L(y, f(\mathbf{x}, \alpha))$
  - It measures the quality of the approach performed by the learning algorithm, *i.e.* the discrepancy between the response y of the supervisor and the response $f(\mathbf{x}, \alpha)$ of the learning machine. Its values are ≥ 0
- Risk functional: $R(\alpha) = \int L(y, f(\mathbf{x}, \alpha))\, p(\mathbf{x}, y)\, d\mathbf{x}\, dy$

The goal of a learning process is to find the function $f(\mathbf{x}, \alpha_0)$ that minimizes R(α) (over the class of functions $f(\mathbf{x}, \alpha)$) in the situation where $p(\mathbf{x}, y)$ is unknown and the only available information is contained in the training set

Mathematical description

$$R(\alpha) = \int L\left(y, f\left(\mathbf{x}, \alpha\right)\right) p\left(\mathbf{x}, y\right) d\mathbf{x}\, dy$$

- Two main learning algorithms have been considered
  - Pattern recognition (or classification)

$$L\left(y, f\left(\mathbf{x}, \alpha\right)\right) = \begin{cases} 0 \text{ if } y = f\left(x, \alpha\right) \\ 1 \text{ if } y \neq f\left(x, \alpha\right) \end{cases}$$

  - Regression estimation

$$L\left(y, f\left(\mathbf{x}, \alpha\right)\right) = \left(y - f\left(\mathbf{x}, \alpha\right)\right)^2$$

Classification

Description of objects

Dataset: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_N, y_N)$

$\mathbf{x}_i \in R^m$: features that are of distinctive nature (object description with attributes managed by computers)

$y_i \in \{L_1, L_2, \ldots L_K\}$: label of the sample $\mathbf{x}_i$

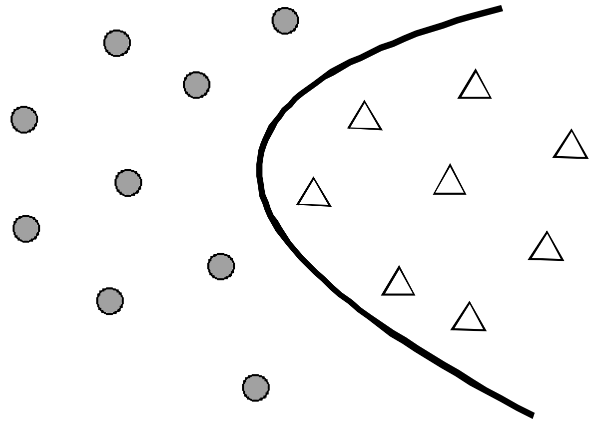
Feature types

- Quantitative (numerical)
  - Continuous-valued (length, pressure)
  - Discrete (total basketball score, number of citizens in a town)
- Qualitative (categorical)
  - Ordinal (education degree)
  - Nominal (profession, brand of a car)

Supervised classifiers

Training dataset: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_N, y_N),\ \mathbf{x}_i \in R^m, y_i \in \{L_1, L_2, \ldots L_K\}$

Test sample: $(\mathbf{x}, y)$, $\mathbf{x}$ is known, y is unknown



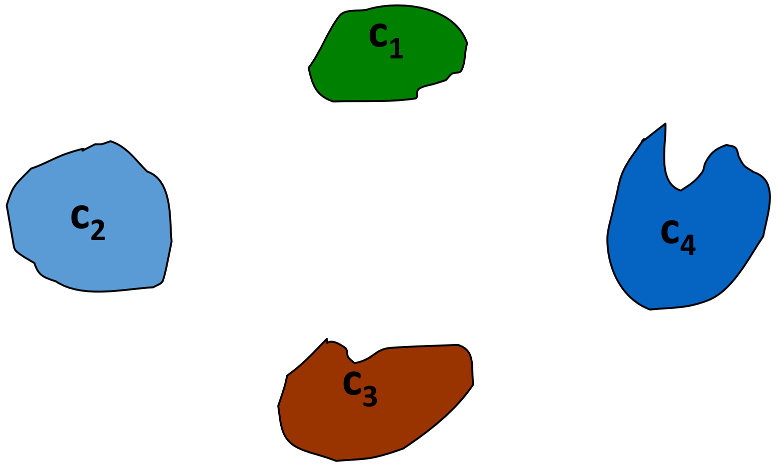
$K = 2$

Types of supervised classifiers

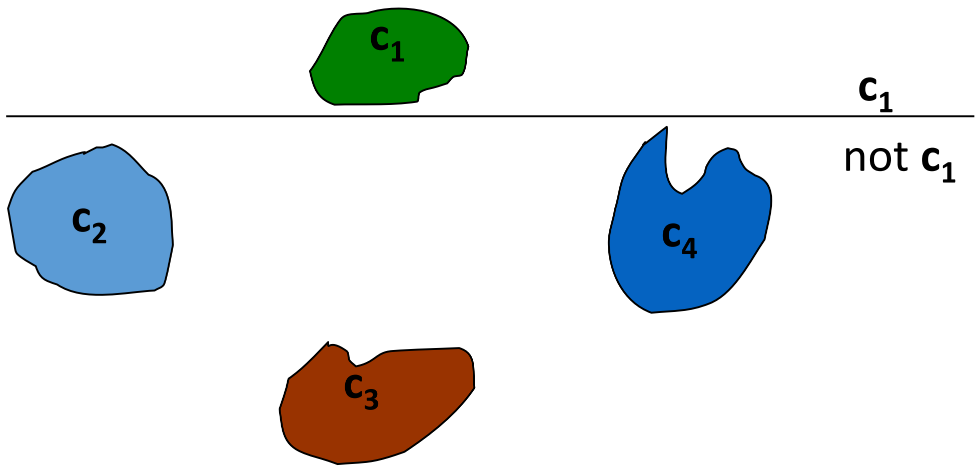
- Support Vector Machines (SVM)
- Neural networks
- Bayes decision theory
  - Parametric method
  - Non-parametric method
- Classification trees

Supervised classifiers ($K > 2$)

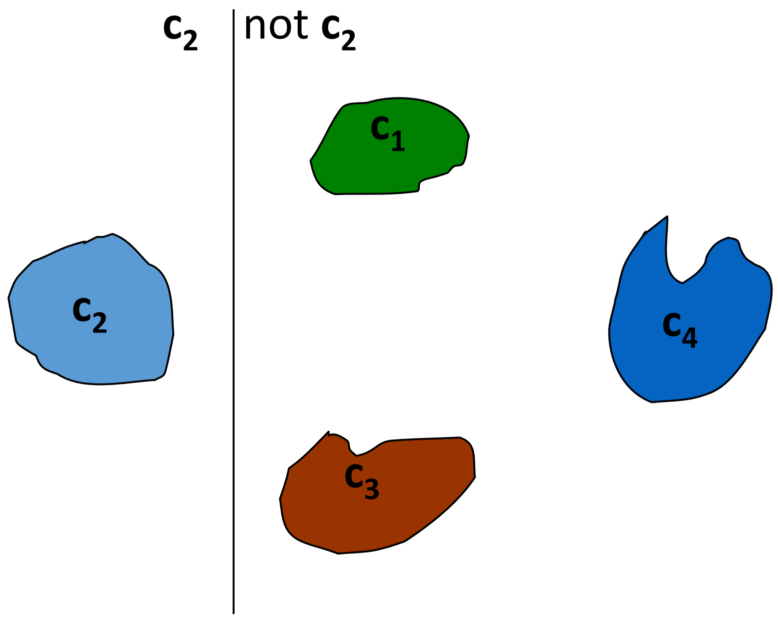
- This case can be tackled as K binary problems. In the training process, each class is compared with the rest (one-versus-the-rest approach)



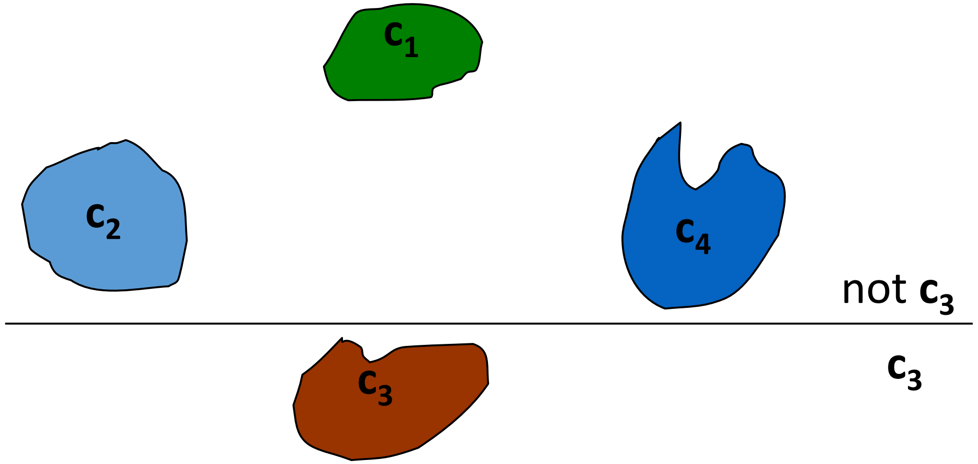
Supervised classifiers ($K > 2$)



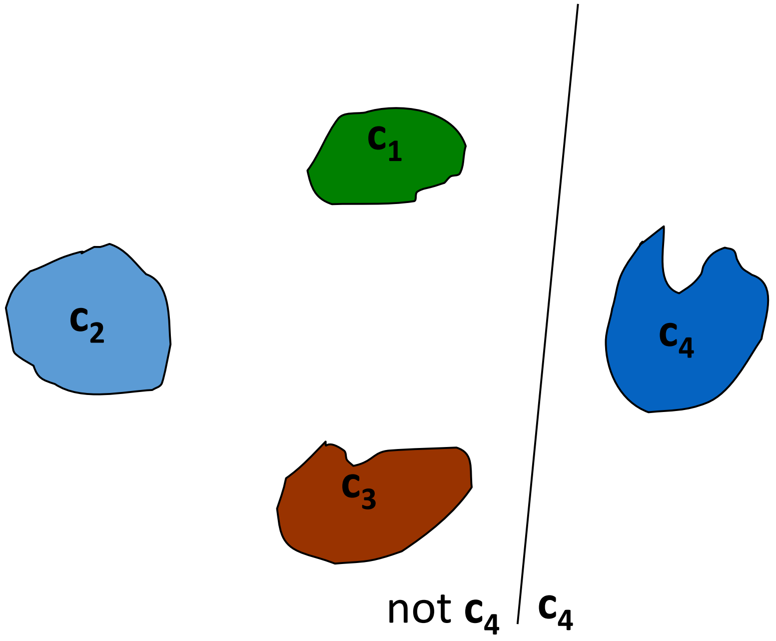
Supervised classifiers ($K > 2$)



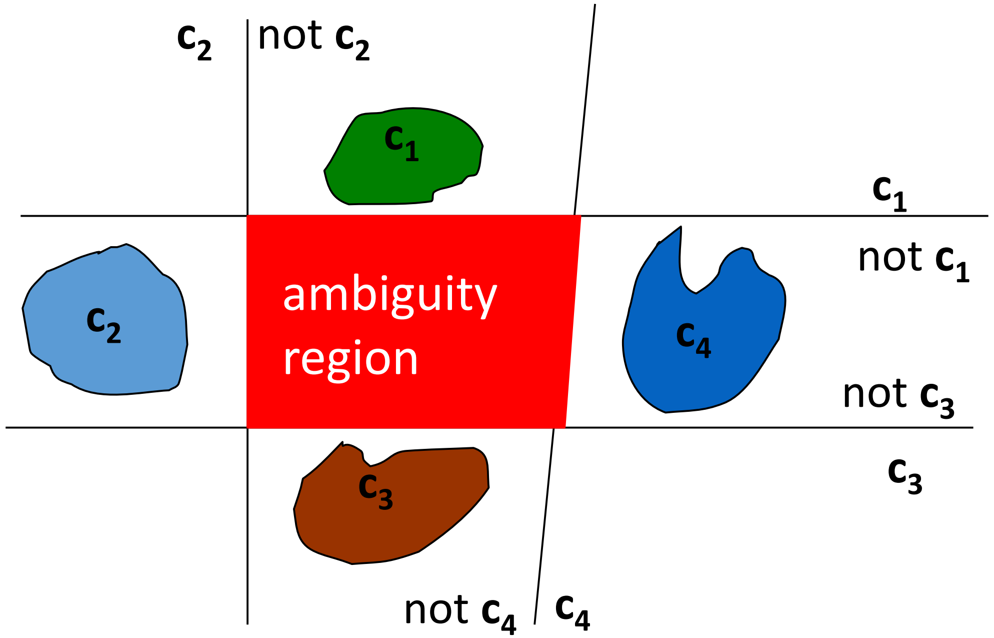
Supervised classifiers ($K > 2$)



Supervised classifiers ($K > 2$)



Supervised classifiers ($K > 2$)



Supervised classifiers

- How good is a classifier?

Dataset: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$

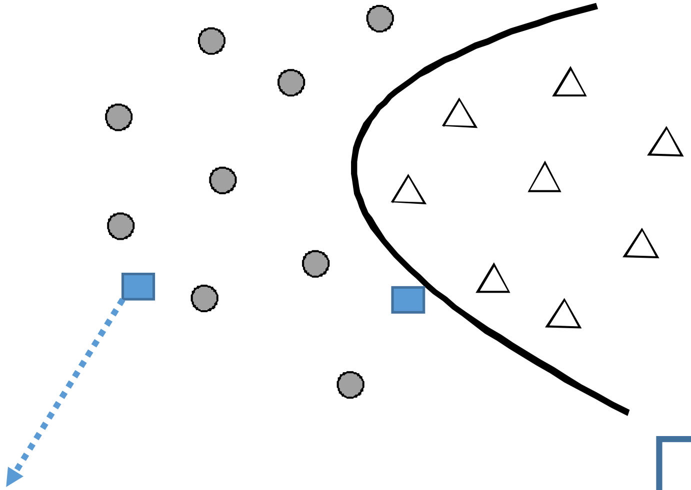
$(\mathbf{x}_i, y_i)$, i = 1, ..., J: training set

$(\mathbf{x}_i, y_i)$, i = J+1, ..., N: test set

- Training set: a model is created to make predictions
  - Given $\boldsymbol{x}$, the model predicts y
- Test set: model validation
  - The success rate is taken as the level of confidence and *it is assumed to be the same* for all future samples
- Predictions corresponding to different samples can have different levels of confidence

Supervised classifiers

Different samples can have different levels of confidence



Higher level of confidence
in the classification

Objective: to qualify each particular prediction with a measure of its reliability

Accuracy and reliability of classifiers

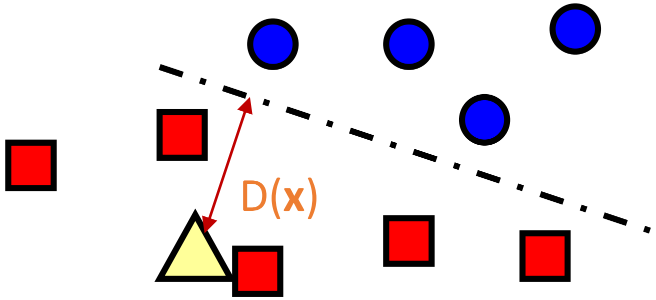
- Bayes classifiers

likelihood

must be known or assumed

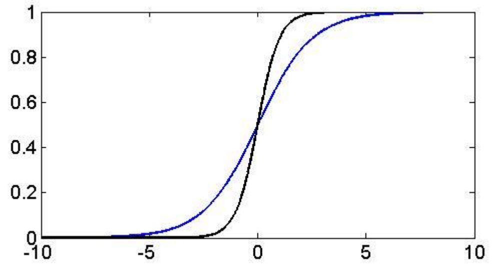
$$P\left(C_j \mid \mathbf{x}\right)=\frac{p\left(\mathbf{x} \mid C_j\right)P\left(C_j\right)}{p\left(\mathbf{x}\right)},\ j\in\left\{L_1,L_2,\ldots L_C\right\}$$

- Logistic regression



$$P_{class\{+1\}}=\frac{1}{1+\exp\left[-kD\left(\mathbf{x}\right)\right]}$$

$$P_{class\{-1\}}=1-P_{class\{+1\}}$$



The greater the distance D(x) the deeper is the point in its corresponding class. This has a translation in terms of probability

Learning using privileged information (LUPI)

- Classical machine learning paradigm: given a dataset of pairs for training purposes, the separation frontier between classes is found

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n), \quad \mathbf{x}_i \in \mathbf{X}, y_i \in \{-1, 1\}$$

Given a new feature vector $\mathbf{x} \in \mathbf{X}$, label y is predicted

- LUPI paradigm: given a dataset of triplets for training purposes, the separation frontier between classes is found

$$(\mathbf{x}_1, \mathbf{x}_1^*, y_1), (\mathbf{x}_2, \mathbf{x}_2^*, y_2), \ldots, (\mathbf{x}_n, \mathbf{x}_n^*, y_n), \quad \mathbf{x}_i \in \mathbf{X}, \mathbf{x}_i^* \in \mathbf{X}^*, y_i \in \{-1, 1\}$$

Given a new feature vector $\mathbf{x} \in \mathbf{X}$, label y is predicted

- The additional information $\mathbf{x}^* \in \mathbf{X}^*$ is only available at the training stage
- The additional information $\mathbf{x}^* \in \mathbf{X}^*$ belongs (generally speaking) to the space $\mathbf{X}^*$ which is different from the space $\mathbf{X}$

GOBIERNO DE ESPAÑA | MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES | Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

V. Vapnik , A. Vashist. Neural Networks 22 (2009) 544-557

Unsupervised classifiers

Dataset: $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_N, y_N)$

$y_i \in \{L_1, L_2, \ldots, L_K\}$: class labels of the training samples $\boldsymbol{x}_\mathrm{i}$ are not available

Objective: to "*reveal*" the organization of the samples into a number of "*sensible*" clusters which will allow us to discover similarities and differences among samples and to derive useful conclusions about them

- The labels are known just after the training
- A clustering criterion is needed: **proximity measure (distance or similarity)**

$\{\underbrace{\text{sheep, dog, cat,}}_{\text{mammals}}\ \underbrace{\text{sparrow, seagull,}}_{\text{birds}}\ \underbrace{\text{viper, lizard,}}_{\text{reptiles}}\ \underbrace{\text{goldfish, red mullet, blue shark,}}_{\text{fish}}\ \underbrace{\text{frog}}_{\text{amphibians}}\}$

Class of animals

Existence of lungs

sheep frog cat dog lizard sparrow viper seagull

goldfish red mullet shark

frog

sheep dog cat lizard sparrow viper seagull

goldfish red mullet shark

Environment where the animals live

Unsupervised clustering categories of clustering algorithms

- Sequential algorithms
  - k-means
- Hierarchical clustering algorithms
  - Agglomerative algorithms
  - Divisive algorithms
- Clustering algorithms based on cost function optimization
  - Hard or crisp clustering algorithms
  - Probabilistic clustering algorithms
  - Fuzzy clustering algorithms
  - Boundary detection algorithms

Regression

Regression estimations (high dimensional cases)

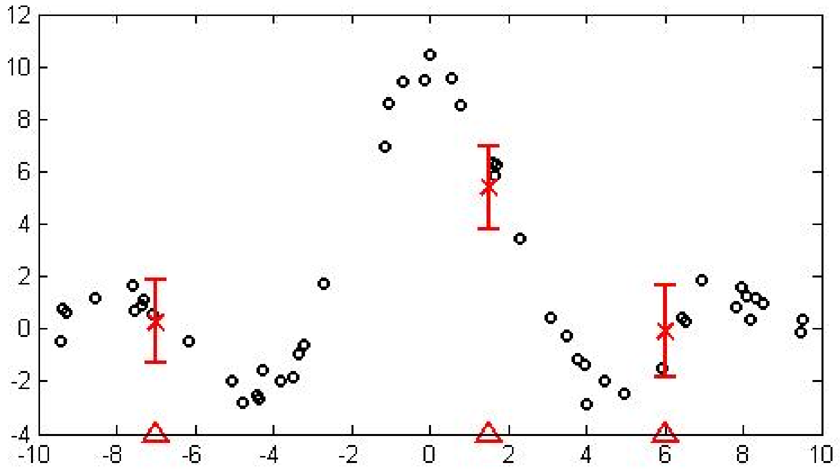
Dataset: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_N, y_N)$

$\mathbf{x}_i \in R^m$: features that are of distinctive nature (object description with attributes managed by computers)
$y_i \in R$: label of the sample $\mathbf{x}_i$

What is the prediction region of the estimations?

$$y = f\left(x_1, x_2, \ldots x_m\right), m >>$$

- Support vector machines
  - No estimation of error bars
- Bayesian estimators
  - Error bars estimation
  - Expensive from a computational point of view



Applications in nuclear fusion

- **Disruption prediction in JET**
  - Linear model
    - Success rate: 99%
    - False alarm rate: 1%
    - Average warning times: 400 ms
  - It is possible to estimate the time to the disruption
- Intelligent system for feature extraction to characterize L/H transitions in JET and DIII-D
- Automatic determination of L/H transition times in JET
- Automatic determination of L/H transition times in DIII-D

- Intelligent data retrieval of waveforms and images based on patterns from massive databases (JET and TJ-II)
- Automatic detection of plasma events in waveforms and video-movies (JET)
- Automatic ELM location in JET
- Automatic analysis system in the TJ-II Thomson Scattering based on pattern recognition
- Noise reduction in images (TJ-II Thomson scattering)
- Application of event-based sampling strategies