Introduction to Modelling and to Statistical Learning (Part 1)

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

March 19, 2019

Plan

Introduction

- Data Mining is the process of discovering patterns and relationships in data, with an emphasis on large observational databases.
- Special interest now because of:
  - Explosive growth of data in a great variety of fields revolution in biology, ecology, genomic, internet, network, images, multimedia.
  - Increasing of the computer power, storage devices with higher capacity
  - Faster communications, better database management systems

Extract information from a data set in such a way it can be understandable and usable.

¿For what?

Descriptive and predictive methods.

Descriptive methods

Objective: detect patterns on data by grouping units, attributes or both.
Data is usually unlabeled so we use non supervised approaches. Some descriptive techniques are:

- Clustering : find existing groups on data
- Segmentation : create groups by partitioning
- Factorial Analysis : find factors, i.e. groups of variables or groups of observations.
- Association rule : look for associations of variables
- Dimensional Reduction: Principal Component Analysis, Multidimensional Scalling, ISOMAP, etc.

Examples :

- Clustering electrical load curves
- Segmentation of clients for oriented marketing
- Look for set of items usually sold together on a supermarket.

Predictive methods

Objective : construct a mapping using available instance that can be used to predict new instances.
Data is labeled so we use supervised approaches. Some predictive techniques are:

- Regression Analysis
- Time Series Analysis
- Classification And Regression Trees (CART)
- Support Vector Machines (SVM)
- k-Nearest Neighbours (kNN)

Examples :

- Credit scoring
- Anticipate the electricity demand for tomorrow
- Estimate the probability of a disease for a patient

The Role of the Statistician

Statistics machine learning plays a central role in data mining.

- provide theoretical foundations for learning algorithms
- give useful tools to analyze an algorithms statistical properties and performance guarantee
- help researchers gain deeper understanding of the approaches, design better algorithms, and select appropriate methods for a given problem.
- help to take a better decision.

Machine Learning

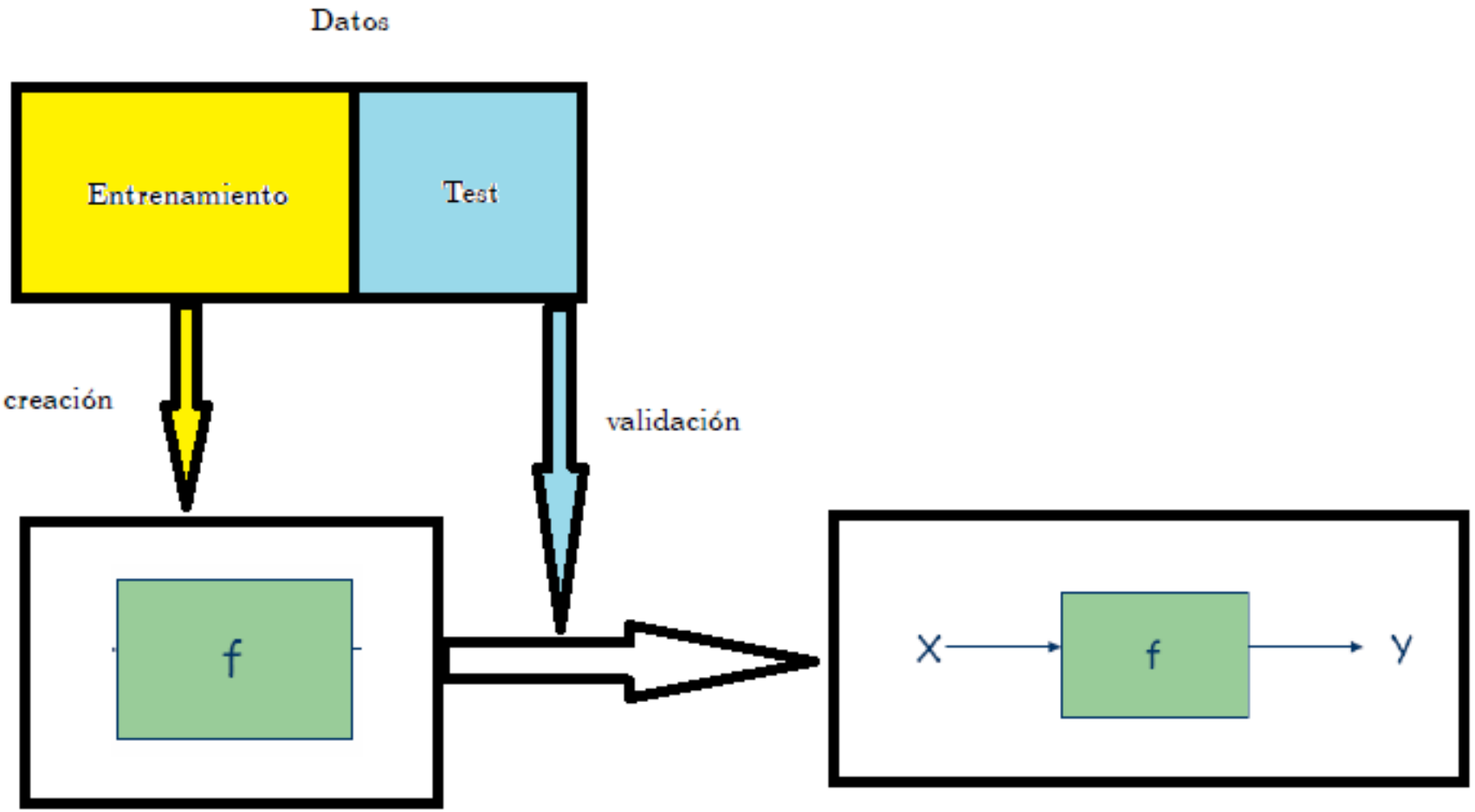
Machine Learning is about predictive methods.

- Another denominations: machine learning, statistical learning, artificial intelligence
- The techniques of Statistical Learning can help solve the problems that frequently arise when modeling an ecological problem, economic phenomenon, medical situation, climatic situation, etc..
- Idea: from a (training) data set, build and train a mathematical model f that will allow, given a new observation, to predict the category to which it belongs or some relevant output value. Predictor f is construct generally without any assumption on distribution or on nature of the dataset.

Examples

- Predict whether an email is spam or not spam.
- Predict whether a patient is prone to heart disease.
- Estimate the ozone rate in a city taking into account climatic variables.
- Predict the absence or presence of a species in a given environment.
- Predicting customer leaks for a financial institution.
- Identify handwritten figures of postcards in envelopes.
- Split a population into several subgroups.

Statistical Learning



Framework of Machine Learning

General framework:
$\mathcal{L}$ a data basis.

Framework of Machine Learning

General framework:
$\mathcal{L}$ a data basis. We search about $f : \mathcal{X} \to \mathcal{Y}$ a good predictor or a good explainer.

- Supervised Learning: $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$

  X: input variable, independent variable, explanatory (real o multidimensional), continuous, categorical, binary, ordinal.

  Y: output variable, dependent variable, real o categorical.

  - Classification: $y \in \{-1, 1\}$ (binary) or $y \in \{1, \ldots, K\}$ (multiclass).
  - Regression: $y \in \mathbb{R}$.

- Unsupervised Learning $\mathcal{L} = \{x_1, \ldots, x_n\} \subset \mathcal{X} \subset \mathbb{R}^d$

  - Clustering
  - Density estimation

  In all cases, the sample $\mathcal{L}$ is a collection of n independents realization of a multivariate random variable (X, Y) or X

Example

1. Environmental pollution
   - Input variables X: vector of environmental variables in the day n (temperature, atmospheric pressure, winds, etc.)
   - Output variable Y: level of environmental pollution in the day $n+1$

The example corresponds to a regression problem, but if the output variable Y divided into categories is considered, the problem is of classification.

Example

2. Selection of habitat of a species.
   - Input variables X: abundance of food, characteristics of the terrain (altitude, slope), distance to water, etc.
   - Output variable Y: presence / absence of the species.

The output is a binary variable, that is, it only takes the values 0 or 1, so the example is of classification.

Example

3. Predicting customer leaks in a banking institution
   - Input variables X: banking behavior (monthly balances, withdrawals, etc.), socio-demographic (personal data), perception of service quality, old customer age.
   - Output variable Y: "leak" or "no leak" of the client.

The output Y is a binary variable, that is, it only takes the values 0 or 1, so the example is of classification.

Objectives

- learn how to formulate a learning problem in a "statistical" framework
- understand existing techniques from a "statistical" perspective; what are the limitations and strengths? Can we do better by relaxing assumptions?

Supervised and Unsupervised

1. *Supervised.* Data bases are of the type

$$X|Y$$

with $X \in \mathcal{M}_{n\times p}$ and $Y \in \mathcal{M}_{n\times 1}$ (categorical label -qualitative, classification problem- or continuous -quantitative, problem of regression-).

We will use this type of database to make inference and construct a predictor f that given a new observation can predict a category or a value having learned from the observations of the data base.

Supervised and Unsupervised

1. *Supervised.* Data bases are of the type

$$X|Y$$

with $X \in \mathcal{M}_{n\times p}$ and $Y \in \mathcal{M}_{n\times 1}$ (categorical label -qualitative, classification problem- or continuous -quantitative, problem of regression-).

We will use this type of database to make inference and construct a predictor f that given a new observation can predict a category or a value having learned from the observations of the data base.

Example: Linear models, Discriminant Analysis, Supervised learning techniques: CART, SVM, kNN, Aggregating Methods.

Supervised and Unsupervised

1. *Supervised.* Data bases are of the type

$$X|Y$$

with $X \in \mathcal{M}_{n\times p}$ and $Y \in \mathcal{M}_{n\times 1}$ (categorical label -qualitative, classification problem- or continuous -quantitative, problem of regression-).

We will use this type of database to make inference and construct a predictor f that given a new observation can predict a category or a value having learned from the observations of the data base.

Example: Linear models, Discriminant Analysis, Supervised learning techniques: CART, SVM, kNN, Aggregating Methods.

2. *Unsupervised.* Data bases are of the type

$$X$$

with $X \in \mathcal{M}_{n\times p}$

We will use this type of database to reduce the number of variables considered, find certain patterns, group, ... We do not have a label that "supervises" our analysis.

Supervised and Unsupervised

1. *Supervised.* Data bases are of the type

$$X|Y$$

with $X \in \mathcal{M}_{n\times p}$ and $Y \in \mathcal{M}_{n\times 1}$ (categorical label -qualitative, classification problem- or continuous -quantitative, problem of regression-).

We will use this type of database to make inference and construct a predictor f that given a new observation can predict a category or a value having learned from the observations of the data base.

Example: Linear models, Discriminant Analysis, Supervised learning techniques: CART, SVM, kNN, Aggregating Methods.

2. *Unsupervised.* Data bases are of the type

$$X$$

with $X \in \mathcal{M}_{n\times p}$

We will use this type of database to reduce the number of variables considered, find certain patterns, group, ... We do not have a label that "supervises" our analysis. Example: Principal Component Analysis, Multidimensional Scalling, Correspondance Analysis, Clustering, Density Estimation.

Supervised and Unsupervised

1. *Supervised.* Data bases are of the type
$$X|Y$$
with $X \in \mathcal{M}_{n\times p}$ and $Y \in \mathcal{M}_{n\times 1}$ (categorical label -qualitative, classification problem- or continuous -quantitative, problem of regression-).
We will use this type of database to make inference and construct a predictor f that given a new observation can predict a category or a value having learned from the observations of the data base.
Example: Linear models, Discriminant Analysis, Supervised learning techniques: CART, SVM, kNN, Aggregating Methods.

2. *Unsupervised.* Data bases are of the type
$$X$$
with $X \in \mathcal{M}_{n\times p}$
We will use this type of database to reduce the number of variables considered, find certain patterns, group, ... We do not have a label that "supervises" our analysis. Example: Principal Component Analysis, Multidimensional Scalling, Correspondance Analysis, Clustering, Density Estimation.

Observation:

- There may be bridges between the two approaches, there is also *semi-supervised learning*,...

Supervised and Unsupervised

1. *Supervised.* Data bases are of the type

$$X|Y$$

with $X \in \mathcal{M}_{n\times p}$ and $Y \in \mathcal{M}_{n\times 1}$ (categorical label -qualitative, classification problem- or continuous -quantitative, problem of regression-).

We will use this type of database to make inference and construct a predictor f that given a new observation can predict a category or a value having learned from the observations of the data base.

Example: Linear models, Discriminant Analysis, Supervised learning techniques: CART, SVM, kNN, Aggregating Methods.

2. *Unsupervised.* Data bases are of the type

$$X$$

with $X \in \mathcal{M}_{n\times p}$

We will use this type of database to reduce the number of variables considered, find certain patterns, group, ... We do not have a label that "supervises" our analysis. Example: Principal Component Analysis, Multidimensional Scalling, Correspondance Analysis, Clustering, Density Estimation.

Observation:

- There may be bridges between the two approaches, there is also *semi-supervised learning*,...
- The method of learning used in the case of supervised learning clearly depends on the nature of the response (whether it is qualitative or quantitative).

Supervised and Unsupervised

1. *Supervised.* Data bases are of the type

$$X|Y$$

with $X \in \mathcal{M}_{n\times p}$ and $Y \in \mathcal{M}_{n\times 1}$ (categorical label -qualitative, classification problem- or continuous -quantitative, problem of regression-).

We will use this type of database to make inference and construct a predictor f that given a new observation can predict a category or a value having learned from the observations of the data base.

Example: Linear models, Discriminant Analysis, Supervised learning techniques: CART, SVM, kNN, Aggregating Methods.

2. *Unsupervised.* Data bases are of the type

$$X$$

with $X \in \mathcal{M}_{n\times p}$

We will use this type of database to reduce the number of variables considered, find certain patterns, group, ... We do not have a label that "supervises" our analysis. Example: Principal Component Analysis, Multidimensional Scalling, Correspondance Analysis, Clustering, Density Estimation.

Observation:

- There may be bridges between the two approaches, there is also *semi-supervised learning*,...
- The method of learning used in the case of supervised learning clearly depends on the nature of the response (whether it is qualitative or quantitative).
- **There is no better method than all the rest on all data sets.**

Example

Dataset Advertising:

```
> datos=read.csv("Advertising.csv",header=T,sep=",")
> datos[,-1]
        TV Radio Newspaper Sales
1    230.1  37.8      69.2  22.1
2     44.5  39.3      45.1  10.4
3     17.2  45.9      69.3   9.3
4    151.5  41.3      58.5  18.5
5    180.8  10.8      58.4  12.9
```

In this case, each row of the dataset is an independent realization of the random multivariate variable (X, Y) where:

- $X = (X_1, X_2, X_3)$ is the *input* vector:
  - X_1 budget allocated to advertising by television (TV)
  - X_2 budget allocated to advertising by radio (Radio)
  - X_3 budget allocated to advertising by newspaper (Newspaper)
- Y (Sales) is the amount of sales made and is the output variable (response), dependent variable.

In general we will want models of the general form:

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

where $X_1, X_2, \ldots, X_p$ are predictor variables e Y is the response variable, ϵ is the error term, independent of X and with mean 0.

Data matrix

Two ways to consider the data matrix
$\mathbf{X} = ((x_{ij}))_{i=1,\dots,n}^{j=1,\dots,p} \in \mathcal{M}_{n\times p}$ (n observations with p variables).
By rows (observations):

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \text{obs 1} \\ \text{obs 2} \\ \vdots \\ \text{obs n} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}$$

By columns (variables):

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} v & v & & v \\ a & a & & a \\ r & r & & r \\ i & i & \dots & i \\ a & a & & a \\ b & b & & b \\ l & l & & l \\ e & e & & e \\ 1 & 2 & & p \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_p \end{pmatrix}$$

Let y_i the response of observation i. Our data set is :

$$\mathcal{L} = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \dots, (\mathbf{x_n}, y_n)\}$$

where $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \dots, (\mathbf{x_n}, y_n)$ are independent realizations of variable (X, Y) where Y is dependent of X.

Challenges to the Statisticians

1. Data Complexity: involves many variables which are often related in complex (nonlinear) ways.
2. Big Data (datasets with large number of observations, large number of variables, large number of observations and variables).
3. Feature Selection: many features are available but some are redundant, leading to the feature selection or dimension reduction problem.
4. Optimization: many methods involve finding the "best" parameters values by solving complex and large (containing many parameters) optimization problems. Therefore, efficient optimization techniques are required.
5. Visualization: much harder in a high dimensional space.
6. Curse of dimensionality.

   In high dimension, the points are very far one of the other. Suppose we send out a hypercubical neighborhood about a target point to capture a fraction r of the observations. Since this corresponds to a fraction r of the unit volume, the expected edge length will be $e_p(r) = r^{1/p}$. In ten dimensions $e_{10}(0.01) = 0.63$ and $e_{10}(0.1) = 0.80$, while the entire range for each input is only 1.0. So to capture 1% or 10% of the data to form a local average, we must cover 63% or 80% of the range of each input variable. Such neighborhoods are no longer "local". Reducing r dramatically does not help much either, since the fewer observations we average, the higher is the variance of our fit.

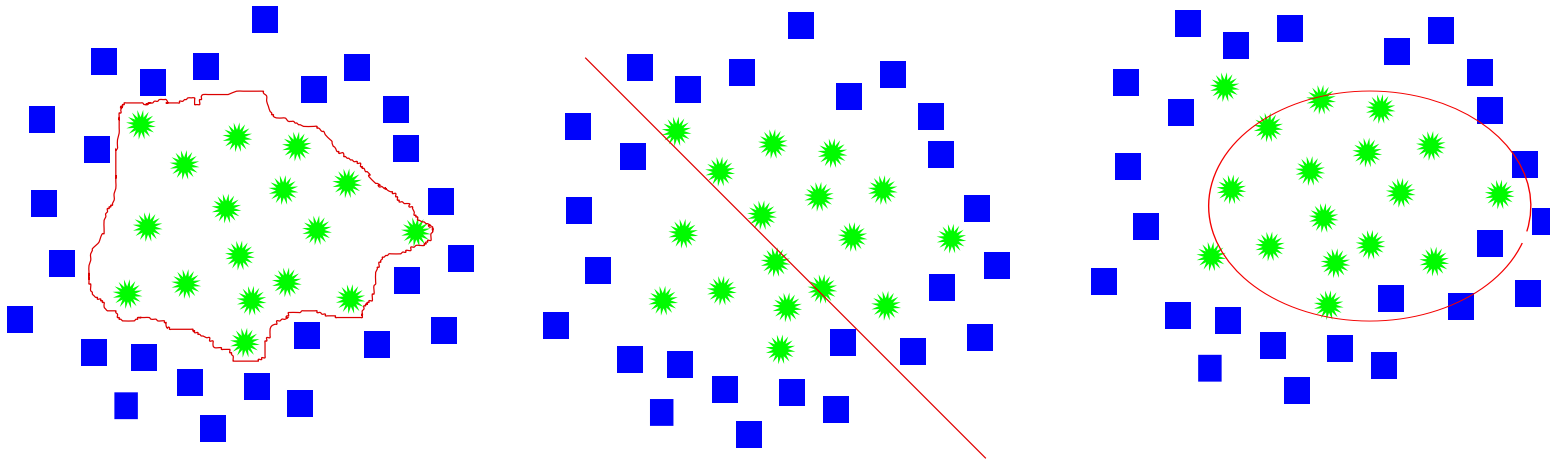
Plan

Overfitting

Notice that a very simple model will probably have a high modelling error and we will not learn too much from the data (underfitting) whereas a model with many parameters will have a high statistical error (overfitting).
We must achieve a compromise between both errors, in such a way that the "generalization error" is the least as possible.

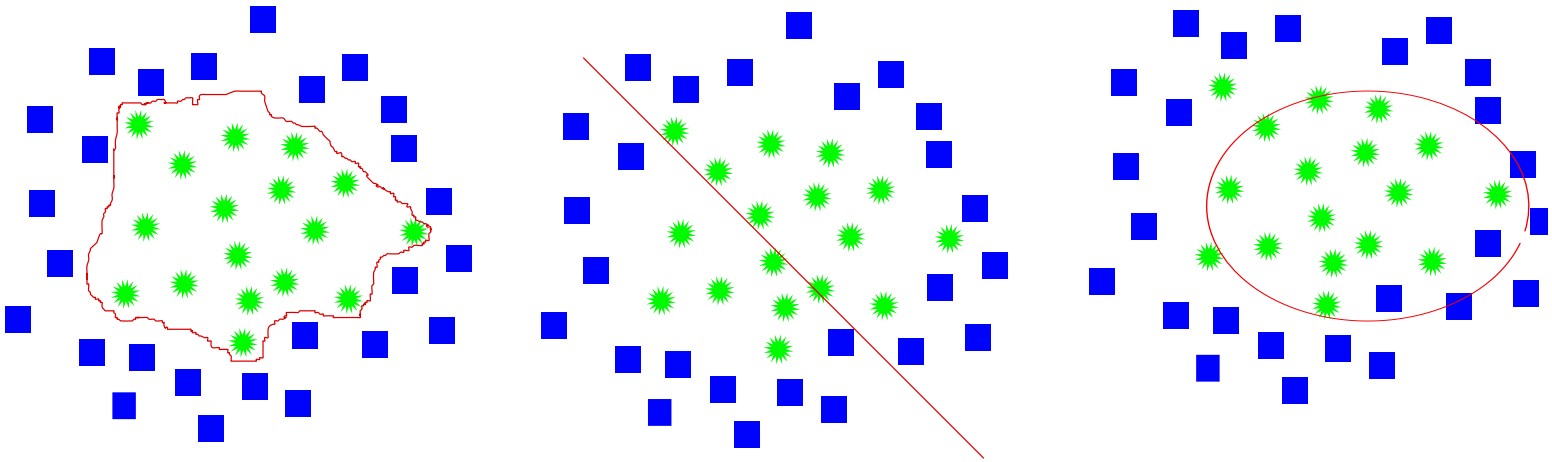
Overfitting

Notice that a very simple model will probably have a high modelling error and we will not learn too much from the data (underfitting) whereas a model with many parameters will have a high statistical error (overfitting).
We must achieve a compromise between both errors, in such a way that the "generalization error" is the least as possible.



Overfitting

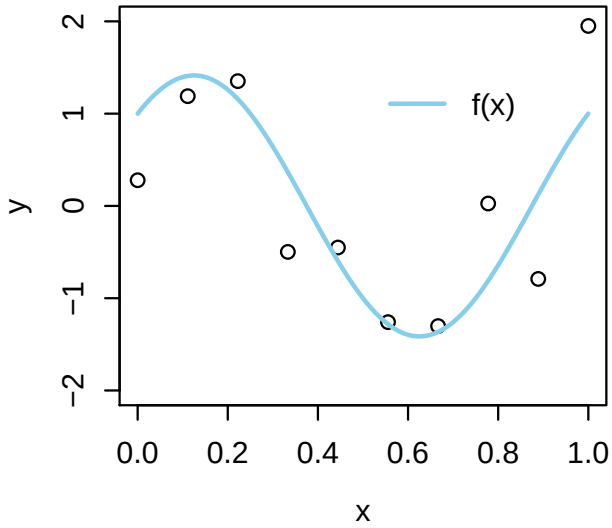
Notice that a very simple model will probably have a high modelling error and we will not learn too much from the data (underfitting) whereas a model with many parameters will have a high statistical error (overfitting).
We must achieve a compromise between both errors, in such a way that the "generalization error" is the least as possible.

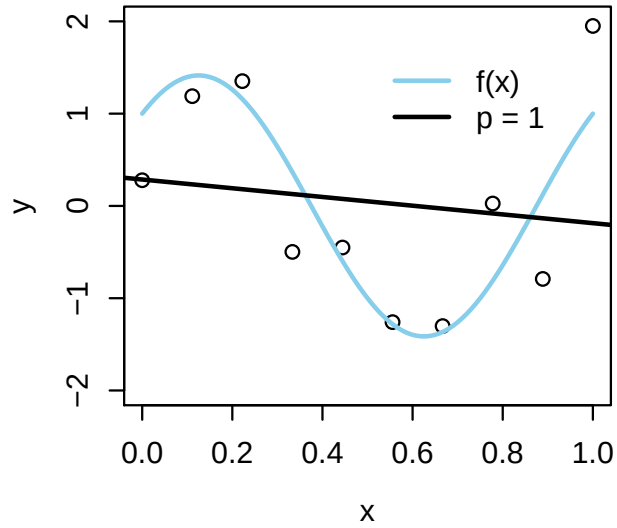


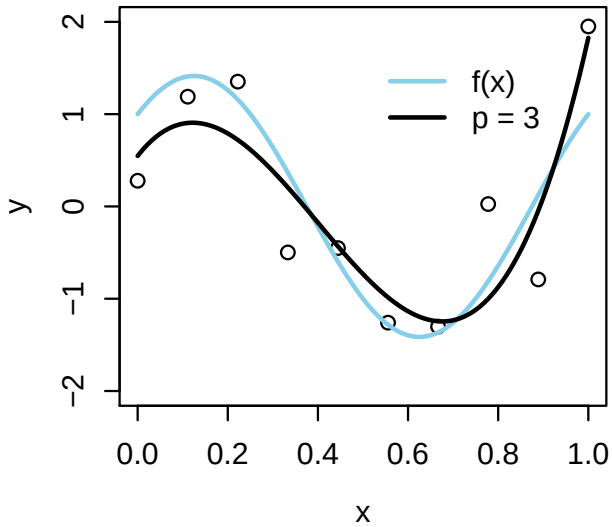
To avoid overfitting, the predictor performance (classification error, mean quadratic error) is evaluated with a new sample called the evaluation sample, independent of the training sample.

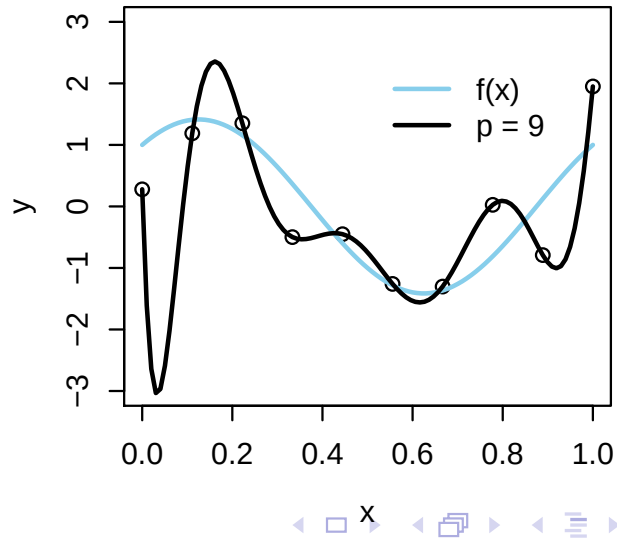
Other ways to evaluate the predictor: cross validation, bootstrap.

Overfitting









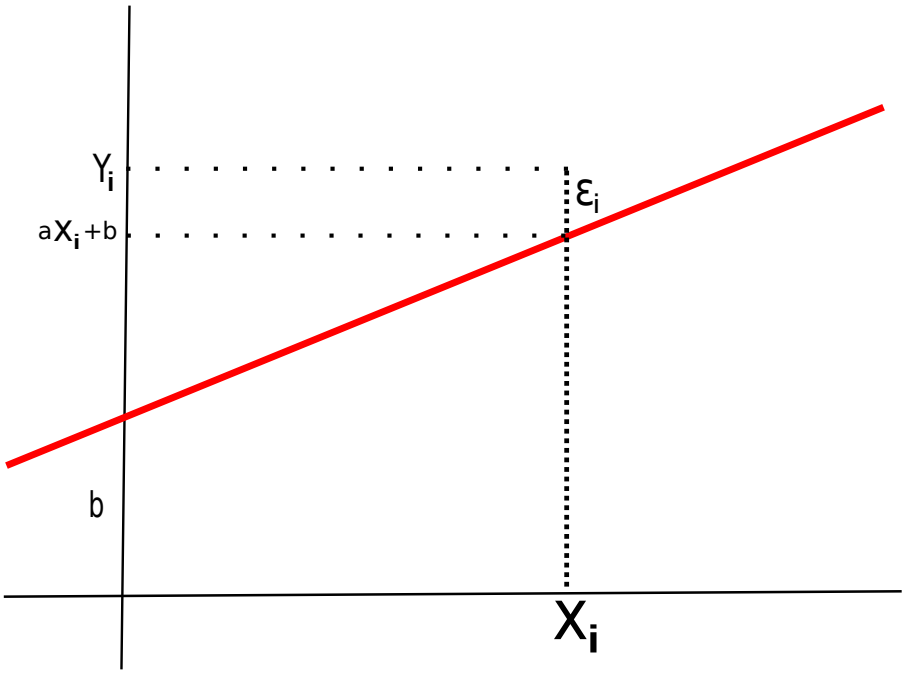
Plan

Simple Linear Model: method of least squares

Data: $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

Simple Linear Model: method of least squares

Data: $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
We look for the line $y = ax + b$ that passes as close as possible to the data.



We find a and b that minimize the sum of squared errors

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2$$

The simple linear regression model is

$$y_i = \underbrace{ax_i + b}_{y_{\text{est}}} + \varepsilon_i, \ \forall\, i = 1, \ldots, n$$

Linear Model: method of least squares

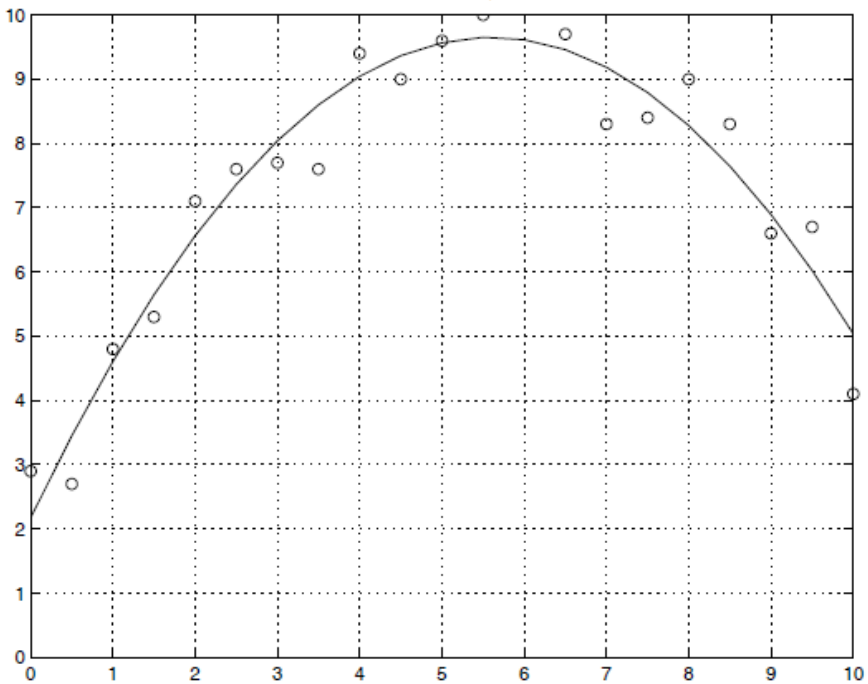
The above method can be easily extended.

Linear Model: method of least squares

The above method can be easily extended.
For example the parabola that adjusts a set of points:

Linear Model: method of least squares

The above method can be easily extended.
For example the parabola that adjusts a set of points:



$$y = a + bx + cx^2$$

(linear model on the coefficients!)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

Multiple Linear Regression

Now we want to predicto a real random variable $Y \in \mathbb{R}$ from d real variables $X_1, \dots, X_d$. We consider model:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d$$

As in simple linear regression, if $\mathcal{L} = \{(\mathbf{x_1}, y_1,), \dots, (\mathbf{x_n}, y_n)\}$ is the data set, we look at a vector $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ \vdots \\ \beta_d \end{pmatrix} \in \mathbb{R}^{d+1}$ that minimizes

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id})\right)^2$$

Observe that $\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id})\right)^2 = ||Y - X\beta||^2$ so we have a linear algebra problem:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix}_{n \times (d+1)}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

whose solution is given by $(X^t X)\beta = X^t \mathbf{y}$.

Classification and Regression Trees (CART)

Classification **A**nd **R**egression **T**rees (Breiman 1984).

Two types of trees: regression trees to predict continuous variables and classification trees to predict categorical variables.

The tree is constructed from binary partitions with respect to the coordinates of the data. For example if the variables are $X_1, \ldots, X_d$, the cut condition for the data will be of type $X_2 < c$ or $X_2 \geq c$ if X_2 is continuous or $X_2 \in \mathcal{A}$ or $X_2 \notin \mathcal{A}$ if X_2 is categorical.

Three steps:

1. Binary separation of the data of each node in two subnodes according to some criterion;

Classification and Regression Trees (CART)

Classification **A**nd **R**egression **T**rees (Breiman 1984).

Two types of trees: regression trees to predict continuous variables and classification trees to predict categorical variables.

The tree is constructed from binary partitions with respect to the coordinates of the data. For example if the variables are $X_1, \ldots, X_d$, the cut condition for the data will be of type $X_2 < c$ or $X_2 \geq c$ if X_2 is continuous or $X_2 \in \mathcal{A}$ or $X_2 \notin \mathcal{A}$ if X_2 is categorical.

Three steps:

1. Binary separation of the data of each node in two subnodes according to some criterion;
2. Decision of the size of the tree: stop and prune criteria

Classification and Regression Trees (CART)

Classification **A**nd **R**egression **T**rees (Breiman 1984).

Two types of trees: regression trees to predict continuous variables and classification trees to predict categorical variables.

The tree is constructed from binary partitions with respect to the coordinates of the data. For example if the variables are $X_1, \ldots, X_d$, the cut condition for the data will be of type $X_2 < c$ or $X_2 \geq c$ if X_2 is continuous or $X_2 \in \mathcal{A}$ or $X_2 \notin \mathcal{A}$ if X_2 is categorical.

Three steps:

1. Binary separation of the data of each node in two subnodes according to some criterion;
2. Decision of the size of the tree: stop and prune criteria
3. Assigning a class or value to terminal nodes.

CART

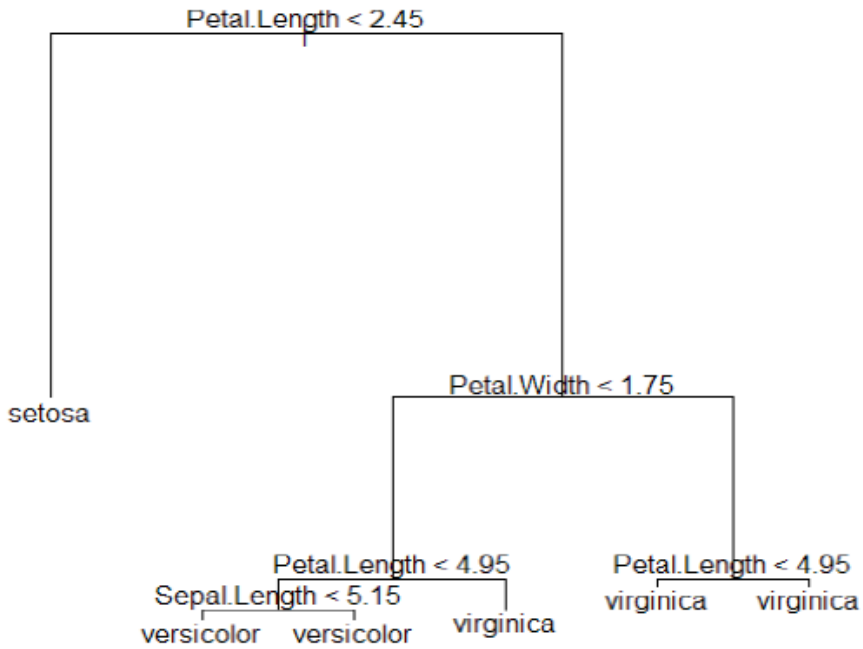
Example: Iris

Goal: Predict the species of the iris flower.

Data: 150 flowers

Dependent variable: Species (setosa, virginica, versicolor)

Independents variables: Sepal Length, Petal Length, Sepal Width, Petal Width



CART

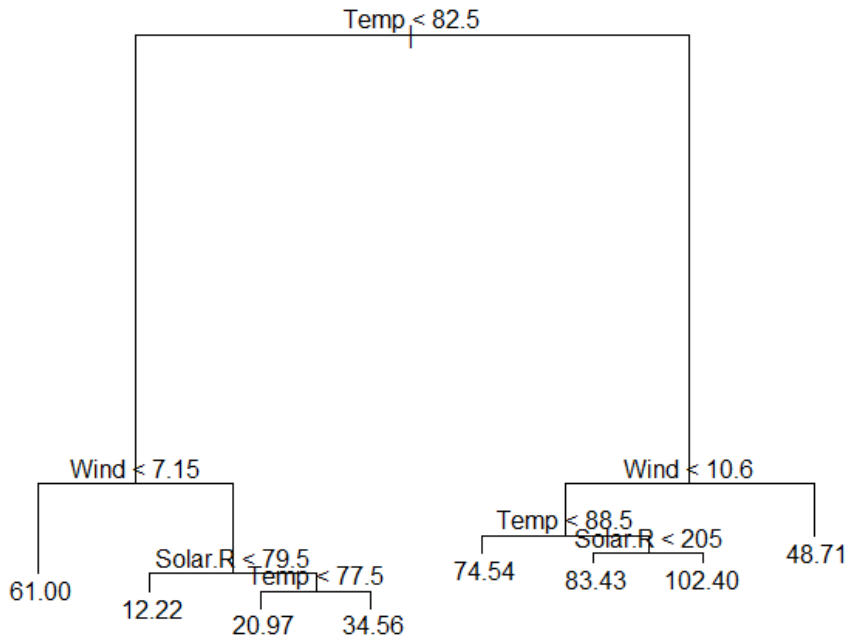
Example: airquality

Goal: Predict the ozone level in New York.

Data: 153 days

Dependent Variable: ozone level

Independents variables Date, Solar Radiation, Wind and Temperature



CART

Easy to interpret, but ... very unstable: a small change in the sample leads to completely different results.

CART

Easy to interpret, but ... very unstable: a small change in the sample leads to completely different results.

Aggregation Methods:

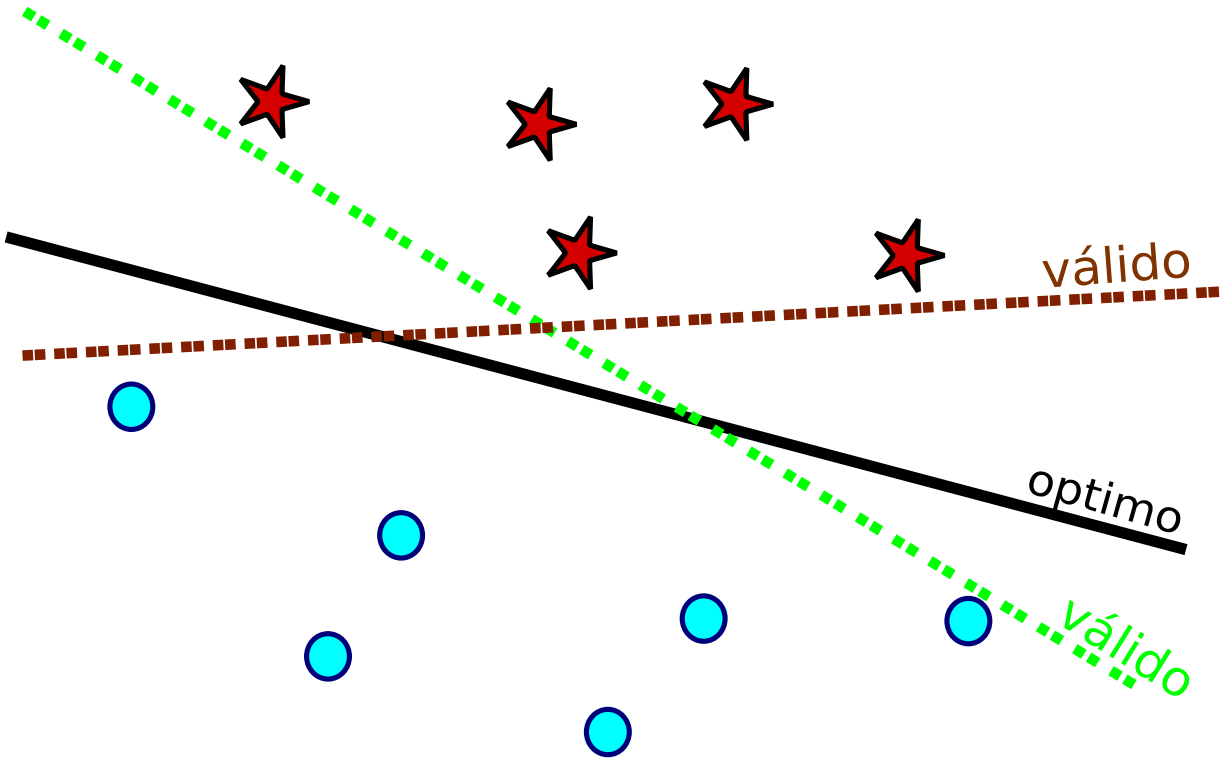
1. **Bagging** (Breiman, 1996): average of several trees based on data re-samples.
2. **Random Forests** (Breiman, 2001): combines the Bagging and CART algorithms.
3. **Boosting** (Freund and Shapire, 1997): weighted average of trees. The weighting takes into account the performance of each tree in each stage of the algorithm.

Support Vector Machines (SVM)

In the classification context, SVM (Vapnik, 1995) is a method that consists of finding a curve that separates the data as best as possible.

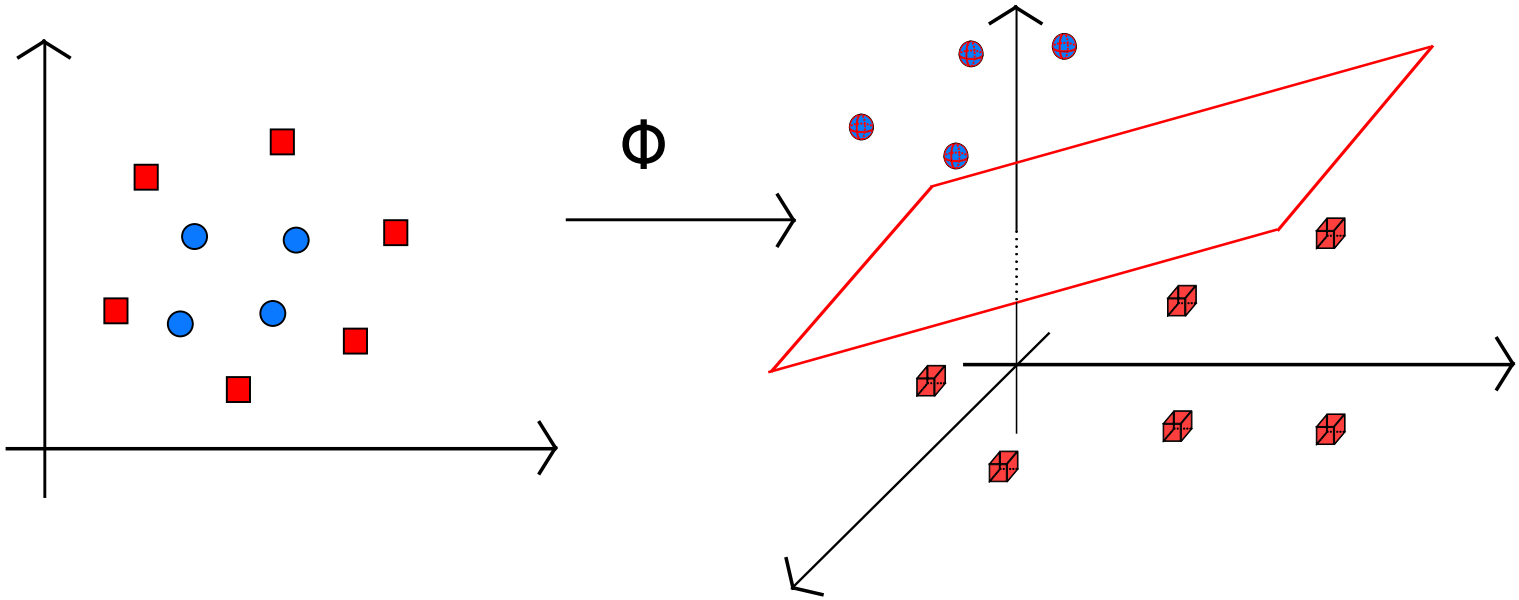
Support Vector Machines (SVM)

In the classification context, SVM (Vapnik, 1995) is a method that consists of finding a curve that separates the data as best as possible.
If the data are linearly separable:



Support Vector Machines

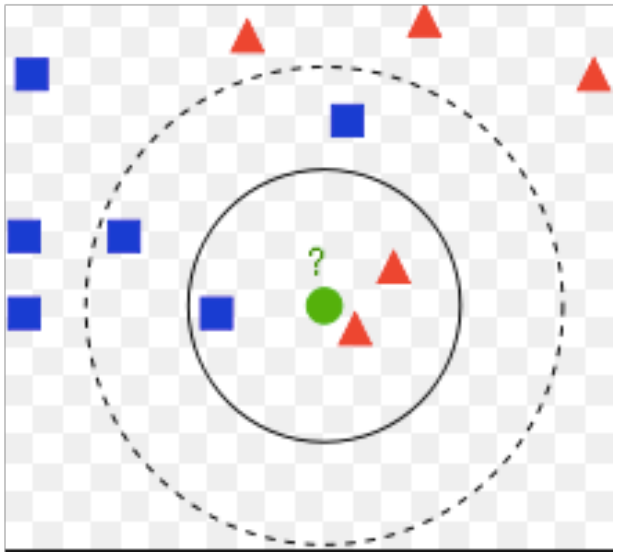
If the data are not linearly separable, we transform them to a space where they are:



k-Nearest Neighbor (k-NN)

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

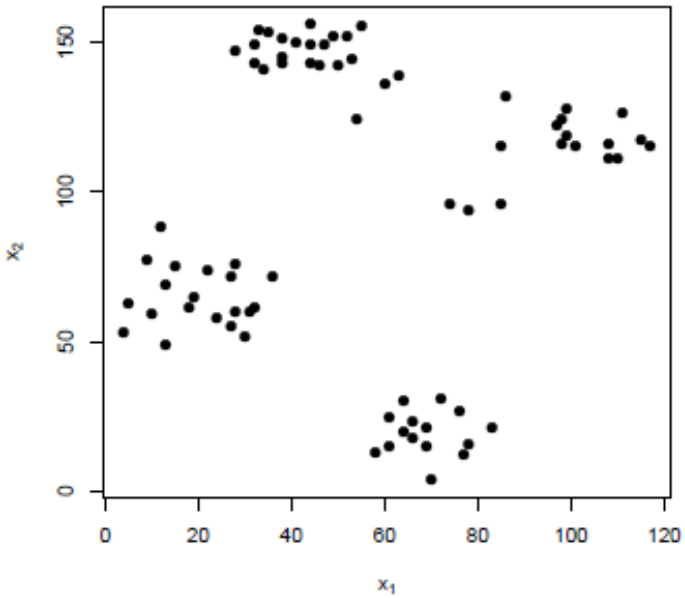


Unsupervised Learning - Clustering

Here we have a data set but without output, that is, $\mathcal{L} = \{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^d$ and we want to create K different homogeneous groups.

Unsupervised Learning - Clustering

Here we have a data set but without output, that is, $\mathcal{L} = \{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^d$ and we want to create K different homogeneous groups.



Unsupervised Learning - Clustering

Here we have a data set but without output, that is, $\mathcal{L} = \{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^d$ and we want to create K different homogeneous groups.