Introduction to Modelling and to Statistical Learning with Application

Mathias Bourel
mbourel@fing.edu.uy

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL),
Facultad de Ingeniería, Universidad de la República, Uruguay

March 29, 2023

Plan

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 unlabeled and in this case we use non supervised approaches. Some descriptive techniques are:

- Clustering : find existing groups on data
- Factorial Analysis : find factors, i.e. groups of variables or groups of observations.
- Dimensional Reduction: Principal Component Analysis, Multidimensional Scalling, ISOMAP, etc.
- Density Estimation

Examples :

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

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
- Text Mining

The Role of the Statistician

Statistics plays a central role in data mining:

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

Leo Breiman, 2001, The Two Cultures

Leo Breiman. Statistical Learning: The Two Cultures. Statistical Science. 16 (3): 199-231, 2001

- There are two cultures in the use of statistical modeling to reach conclusions from data.
- One assumes that the data are generated by a given stochastic data model (*The Data Modeling Culture*).
- The other uses algorithmic models and treats the data mechanism as unknown (*The Algorithmic Modeling Culture*).
- The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.

Machine Learning

- 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 of predictive model: 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.
- If Y is the response:

$$\text{modelisation: } Y = f(X) + \epsilon$$

$$\text{prediction: } \hat{Y} = \hat{f}(X)$$

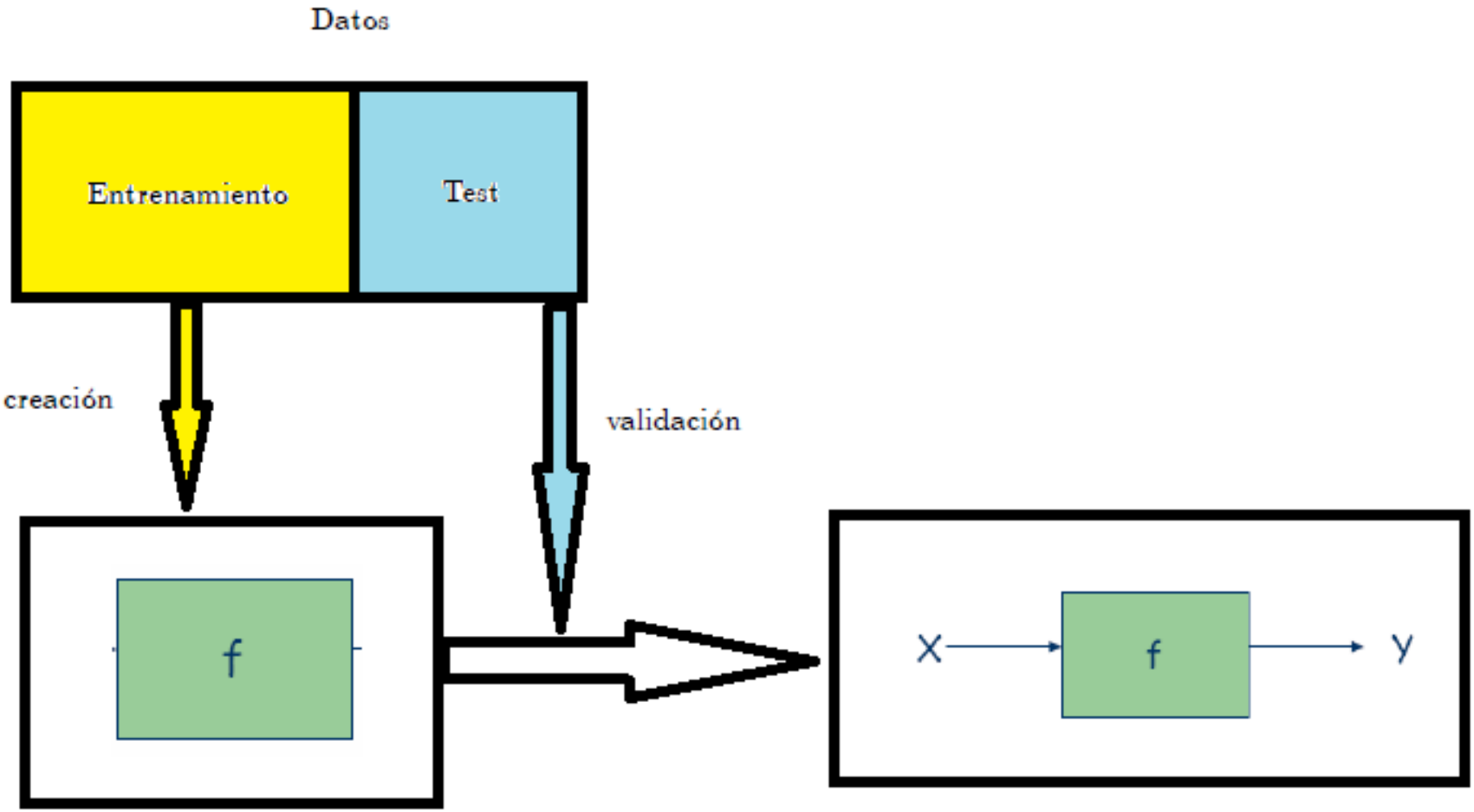
$$\text{we want: } \hat{Y} \approx Y$$

  - Data Modeling Culture: f has a given form (linear or logistic regression) and we estimate parameters from the data. Work for the model. Validation is (generally) about goodness of fit.
  - Algorithmic Modeling Culture: f is an algorithm. Validation is measured by predictive accuracy.

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

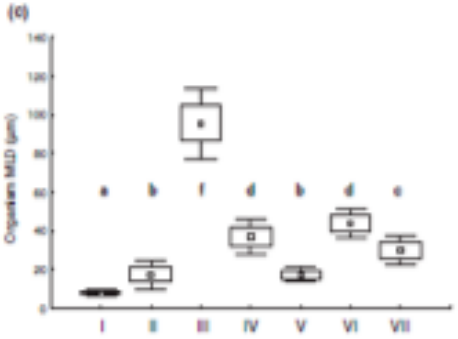
  - Clustering
  - Density estimation

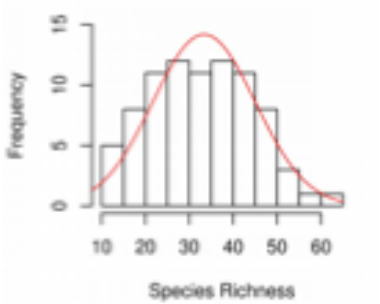
  In all cases, the sample $\mathcal{L}$ is a collection of n independents realization of a multivariate random variable (X, Y) or X

When you are working with data...

When you think about a model to use, it is useful to remember that:

- Data Modeling Culture - Algorithm Modeling Culture
- Supervised - Unsupervised
- Supervised: Clasification - Regression
- Types of variables, Missing Values
- Accuracy of the method is important but it is even more important over *test data*.
- Multiplicity of good models: aggregation methods
- Occam's razor dilemna: simpler is better? Simplicity vs Accuracy?
- Curse of Dimensionality? Handicap or blessing?
- *(The focus)..is on solving the problem instead of asking what data model (they can create). The best solution could be an algorithmic model, or may be a data model, or may be a combination. But the trick to being a scientist is to be open to using a wide variety of tools*, Breiman, The two cultures.
- *All models are wrong, but some are useful*, Georges Box (1919-2013)

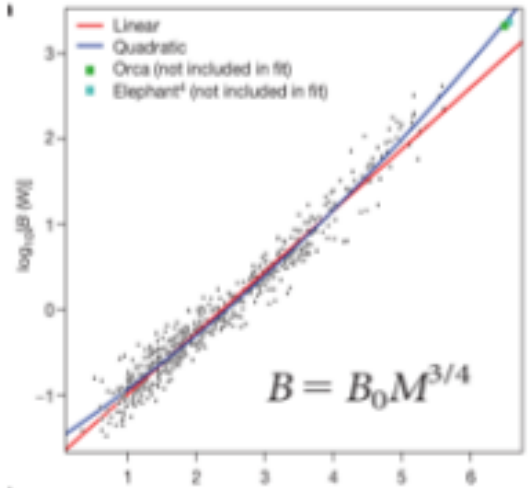




Describe

Explain

Predict



Kolokotrones et al 2010

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.

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, 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, 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, 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, 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, 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, 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.**

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.

Plan

Classification setting

$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $y_1, \ldots, y_n$ are qualitative in $\{1, \ldots, K\}$. Let $\hat{f}$ an estimator of f and $\hat{y}_i = \hat{f}(x_i)$.

- Training error rate: $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}$ (fraction of incorrect classification)

Classification setting

$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $y_1, \ldots, y_n$ are qualitative in $\{1, \ldots, K\}$. Let $\hat{f}$ an estimator of f and $\hat{y}_i = \hat{f}(x_i)$.

- Training error rate: $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}$ (fraction of incorrect classification)
- Test error rate: set of test observations (x_0, y_0): $Ave(\mathbf{1}_{\{y_0 \neq \hat{y}_0\}})$
  A good classifier is one for which the test error is smallest.

Classification setting

$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $y_1, \ldots, y_n$ are qualitative in $\{1, \ldots, K\}$. Let $\hat{f}$ an estimator of f and $\hat{y}_i = \hat{f}(x_i)$.

- Training error rate: $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}$ (fraction of incorrect classification)
- Test error rate: set of test observations (x_0, y_0): $Ave(\mathbf{1}_{\{y_0 \neq \hat{y}_0\}})$
  A good classifier is one for which the test error is smallest.
- **Bayes Classifier**. The test error rate given above is minimized, on average, by the classifier that assigns each observation to the most likely class, given its predictor values.
  Observation x_0 is assigned to class k^* if $\mathbb{P}(Y = k|X = x_0)$ is the largest:

$$\hat{f}(x_0) = k^* \Leftrightarrow k^* = \underset{k \in \{1,\ldots,K\}}{\text{Argmax}} \ \mathbb{P}(Y = k|X = x_0)$$

Classification setting

$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $y_1, \ldots, y_n$ are qualitative in $\{1, \ldots, K\}$. Let $\hat{f}$ an estimator of f and $\hat{y}_i = \hat{f}(x_i)$.

- Training error rate: $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}$ (fraction of incorrect classification)
- Test error rate: set of test observations (x_0, y_0): $Ave(\mathbf{1}_{\{y_0 \neq \hat{y}_0\}})$
  A good classifier is one for which the test error is smallest.
- **Bayes Classifier**. The test error rate given above is minimized, on average, by the classifier that assigns each observation to the most likely class, given its predictor values.
  Observation x_0 is assigned to class k^* if $\mathbb{P}(Y = k|X = x_0)$ is the largest:

$$\hat{f}(x_0) = k^* \Leftrightarrow k^* = \underset{k \in \{1,\ldots,K\}}{\text{Argmax}} \ \mathbb{P}(Y = k|X = x_0)$$

  It is impossible to compute this optimal classifier because we don't know the distribution of Y given X.

Classification setting

$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $y_1, \ldots, y_n$ are qualitative in $\{1, \ldots, K\}$. Let $\hat{f}$ an estimator of f and $\hat{y}_i = \hat{f}(x_i)$.

- Training error rate: $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}$ (fraction of incorrect classification)
- Test error rate: set of test observations (x_0, y_0): $Ave(\mathbf{1}_{\{y_0 \neq \hat{y}_0\}})$
  A good classifier is one for which the test error is smallest.
- **Bayes Classifier**. The test error rate given above is minimized, on average, by the classifier that assigns each observation to the most likely class, given its predictor values.
  Observation x_0 is assigned to class k^* if $\mathbb{P}(Y = k|X = x_0)$ is the largest:

$$\hat{f}(x_0) = k^* \Leftrightarrow k^* = \underset{k \in \{1,\ldots,K\}}{\text{Argmax}} \ \mathbb{P}(Y = k|X = x_0)$$

  It is impossible to compute this optimal classifier because we don't know the distribution of Y given X.
- In binary problem, generally,

$$\hat{f}(x_0) = 1 \Leftrightarrow \mathbb{P}(Y = 1|X = x_0) > 0.5$$

Classification setting

$\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $y_1, \ldots, y_n$ are qualitative in $\{1, \ldots, K\}$. Let $\hat{f}$ an estimator of f and $\hat{y}_i = \hat{f}(x_i)$.

- Training error rate: $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}$ (fraction of incorrect classification)
- Test error rate: set of test observations (x_0, y_0): $Ave(\mathbf{1}_{\{y_0 \neq \hat{y}_0\}})$
  A good classifier is one for which the test error is smallest.
- **Bayes Classifier**. The test error rate given above is minimized, on average, by the classifier that assigns each observation to the most likely class, given its predictor values.
  Observation x_0 is assigned to class k^* if $\mathbb{P}(Y = k|X = x_0)$ is the largest:

$$\hat{f}(x_0) = k^* \Leftrightarrow k^* = \underset{k \in \{1,\ldots,K\}}{\text{Argmax}} \ \mathbb{P}(Y = k|X = x_0)$$

  It is impossible to compute this optimal classifier because we don't know the distribution of Y given X.
- In binary problem, generally,

$$\hat{f}(x_0) = 1 \Leftrightarrow \mathbb{P}(Y = 1|X = x_0) > 0.5$$

Bayes Decision boundary

$$\{x : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 0|X = x)\}$$

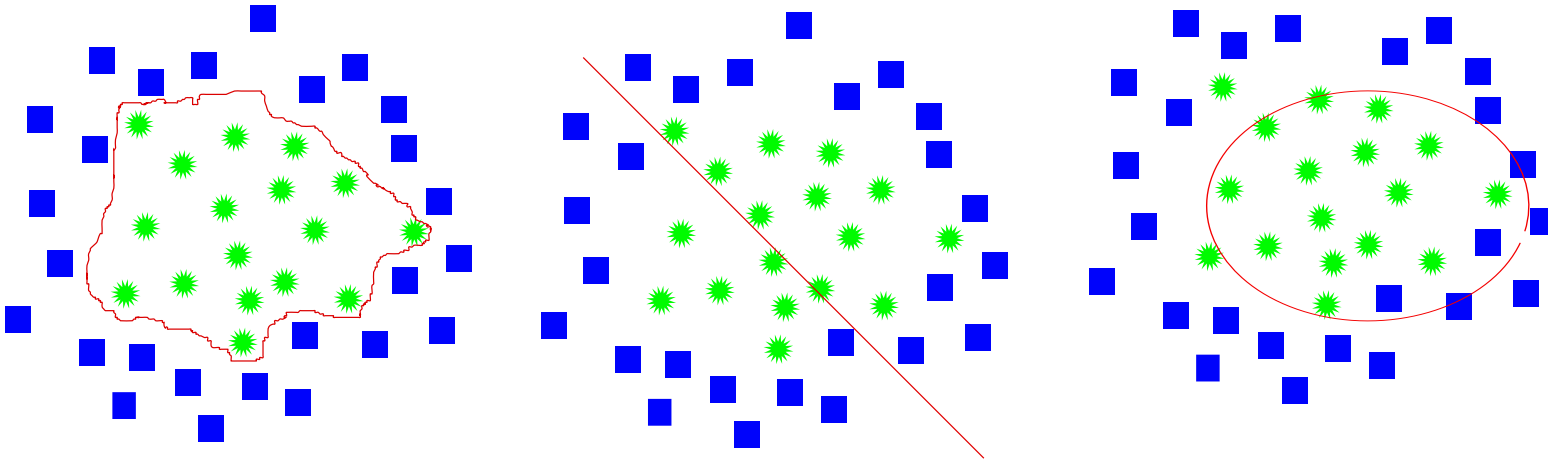
Plan

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.

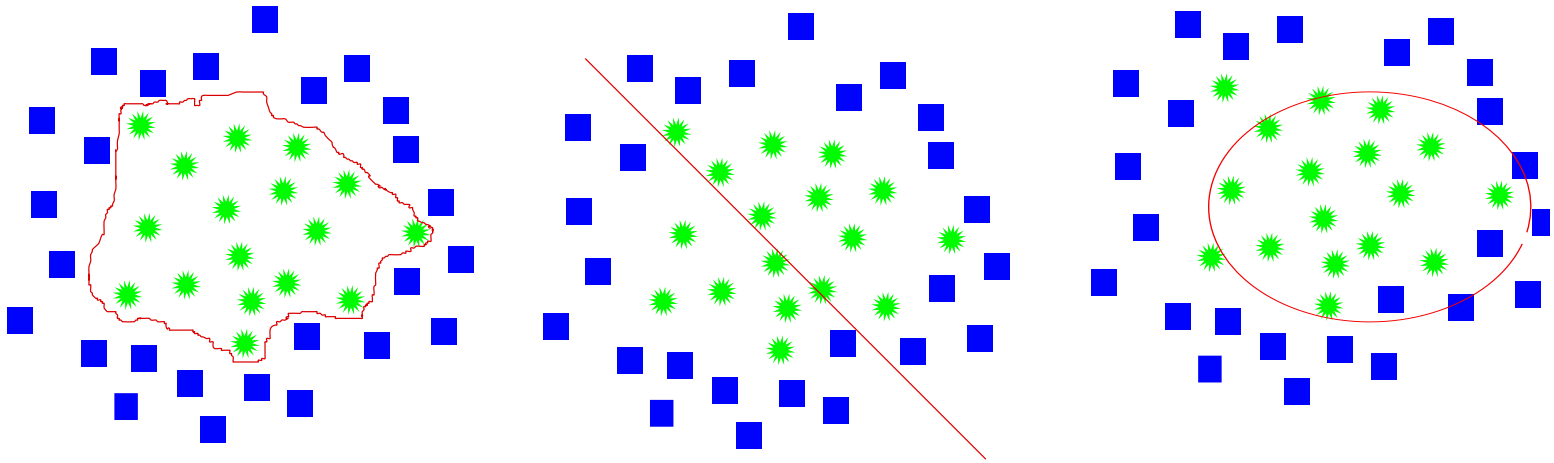
Overfitting

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

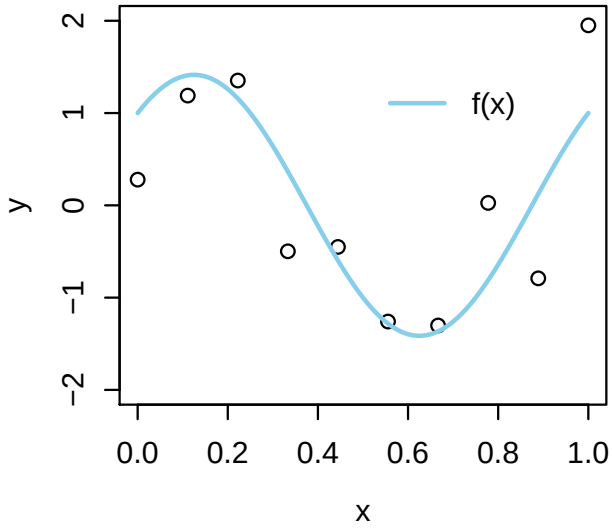
Notice that a very simple model will probably have a high modelling error and we will not learn too much from the data (underfitting) whereas a model with many parameters will have a high statistical error (overfitting).
We must achieve a compromise between both errors, in such a way that the "generalization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

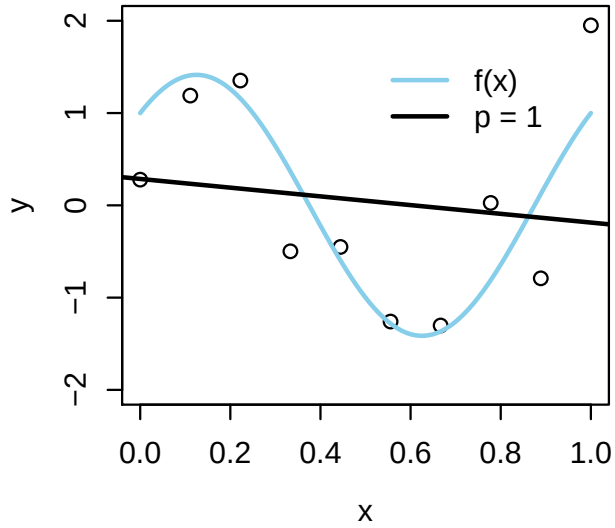


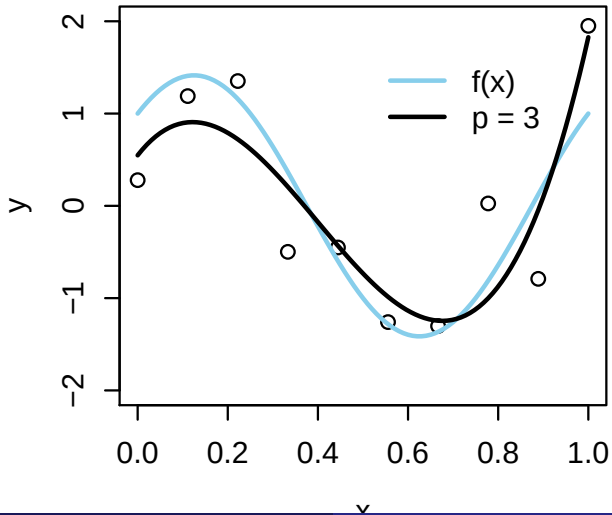
To avoid overfitting, the predictor performance (classification error, mean quadratic error) is evaluated with a new sample called the evaluation sample, independent of the training sample.

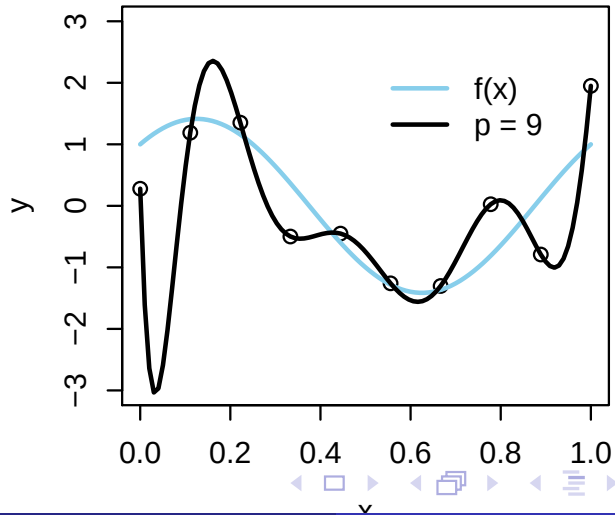
Other ways to evaluate the predictor: cross validation, bootstrap.

Overfitting









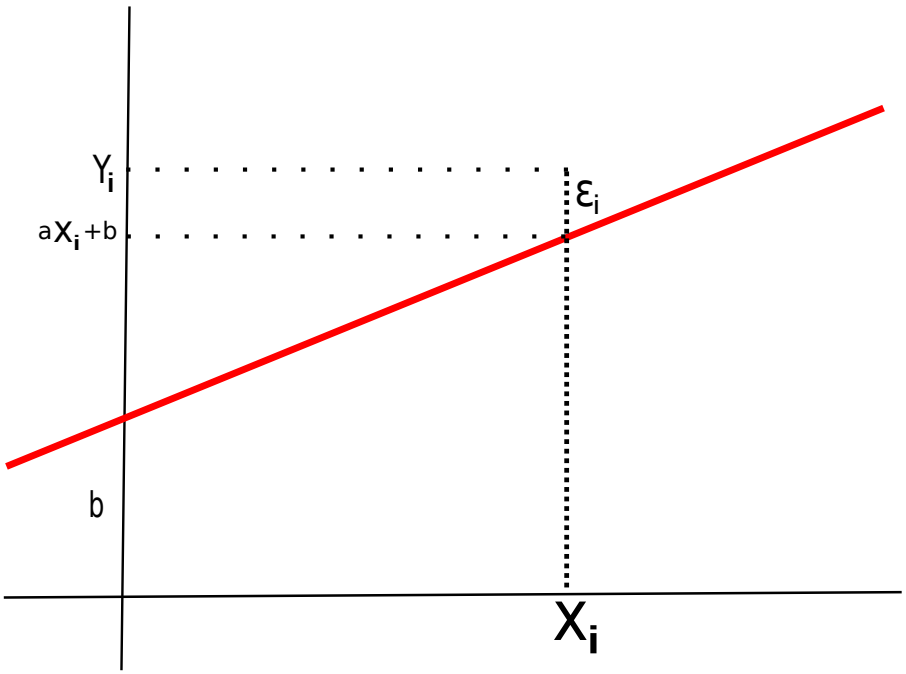
Plan

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

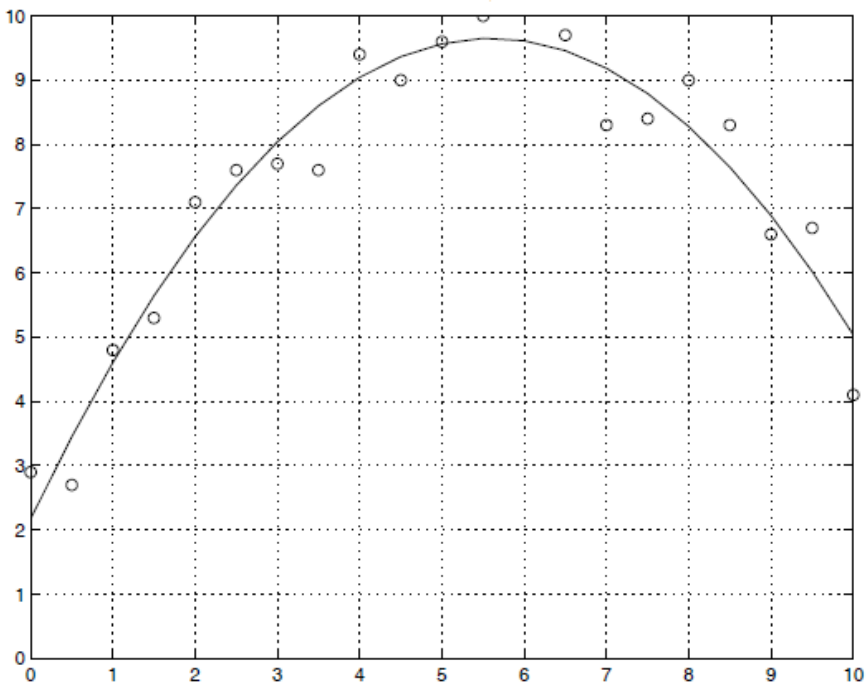
The above method can be easily extended.

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 predict a real random variable $Y \in \mathbb{R}$ from d real variables $X_1, \ldots, X_d$. We consider model:

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

As in simple linear regression, if $\mathcal{L} = \{(\mathbf{x_1}, y_1,), \ldots, (\mathbf{x_n}, y_n)\}$ 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

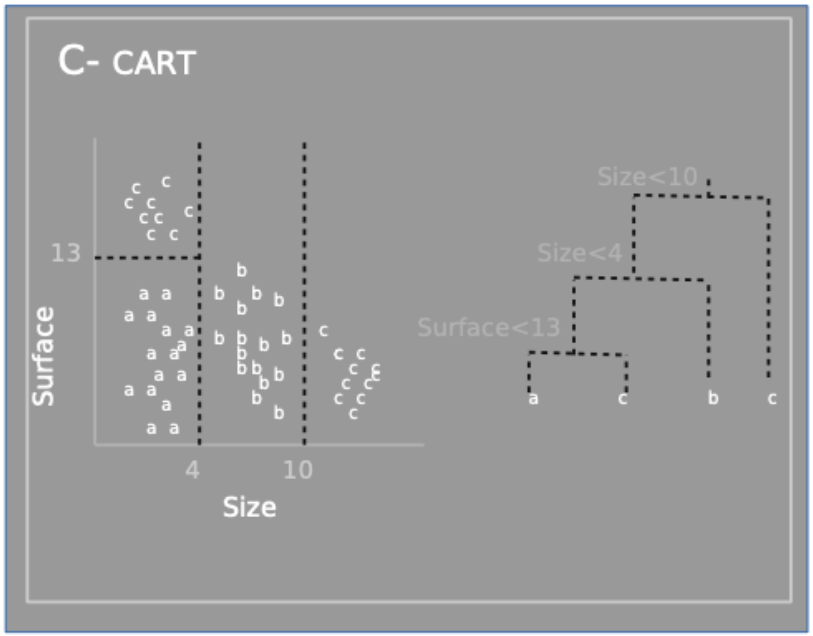


Figure: CART, Breiman et al. 1984 Bourel y Segura 2018. Multiclass classification methods in ecology. Ecological Indicators 85 (2018) 1012-1021

Figure: CART, Thaís Bourel

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.

Random Forest

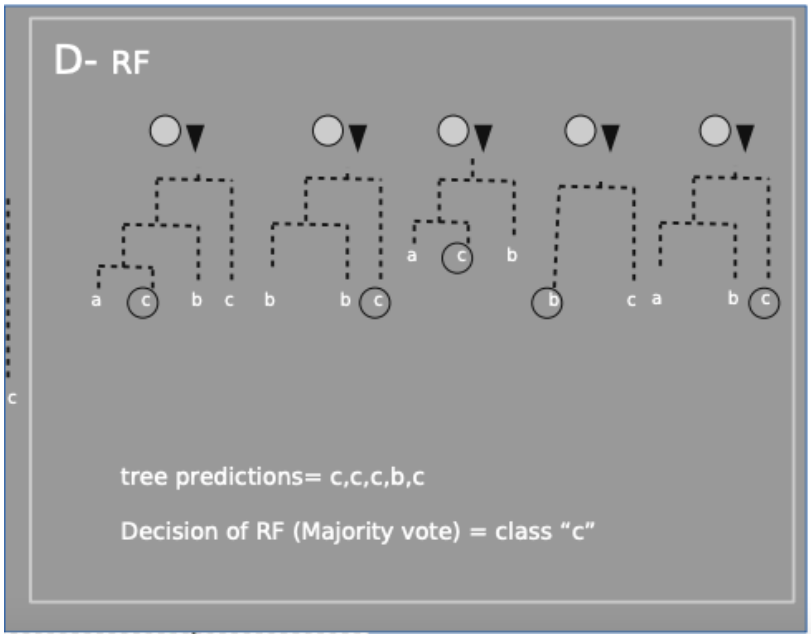


Figure: Random Forest, Breiman 2001
Bourel y Segura 2018. Multiclass classification methods in ecology. Ecological Indicators 85 (2018) 1012-1021

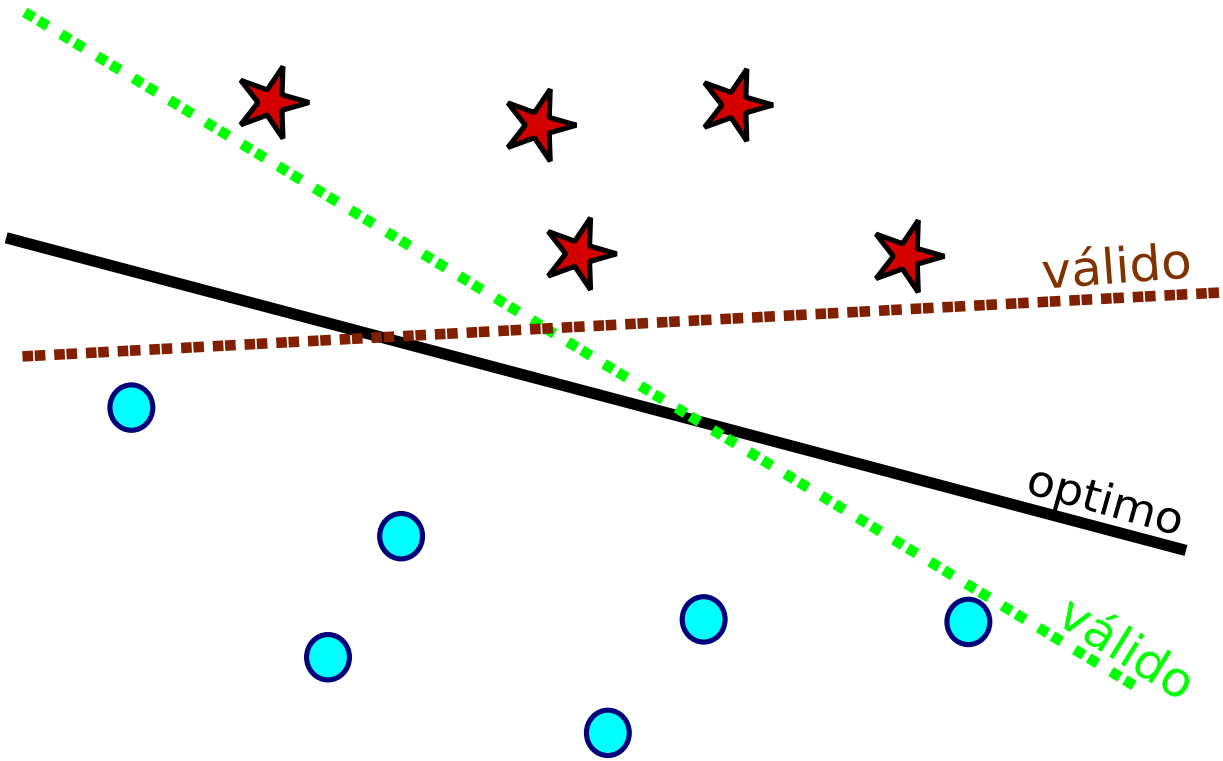
Figure: Random Forest, Thaís Bourel

Support Vector Machines (SVM)

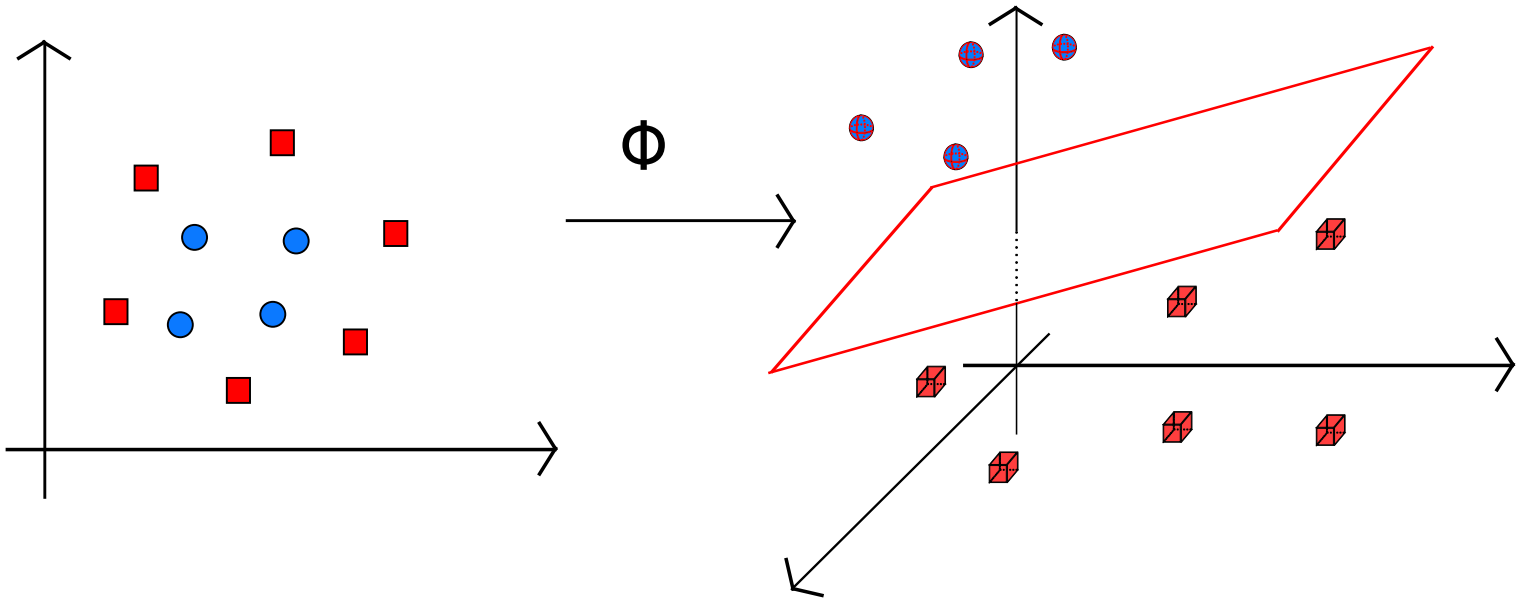
In the classification context, SVM (Vapnik, 1995) is a method that consists of finding a curve that separates the data as best as possible.

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:



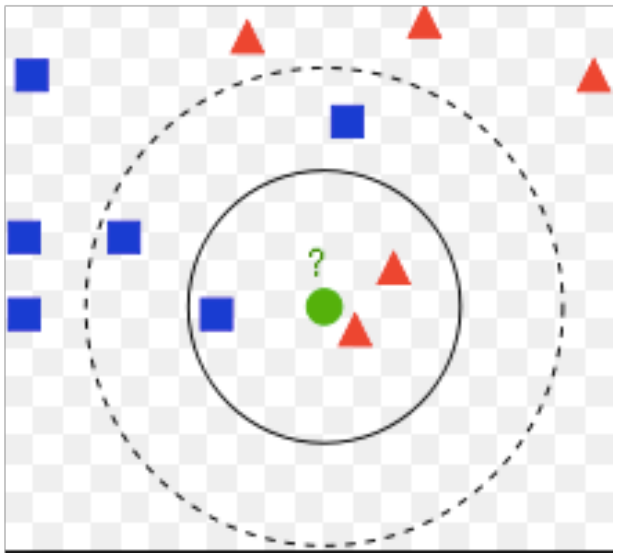
If the data are not linearly separable, we transform them to a space where they are:



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.

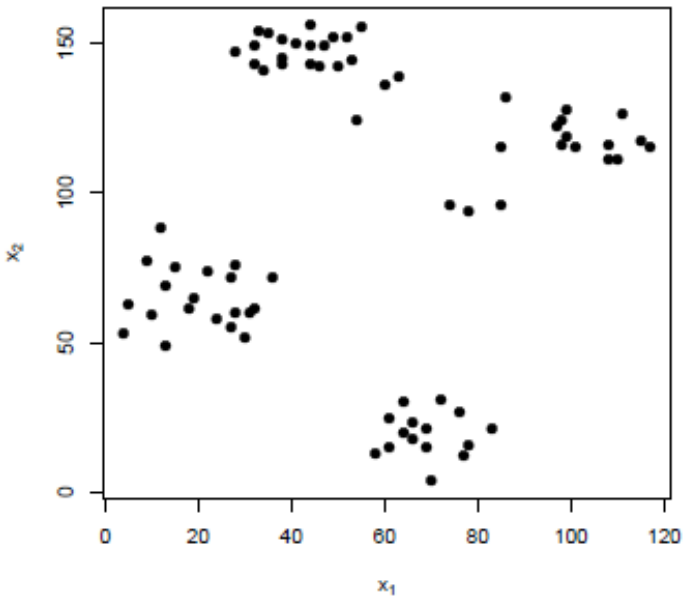


Unsupervised Learning - Clustering

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.

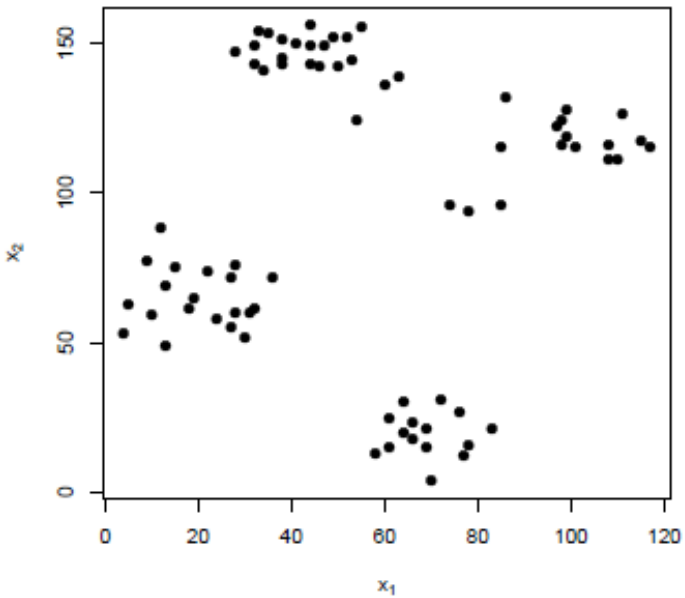
Unsupervised Learning - Clustering

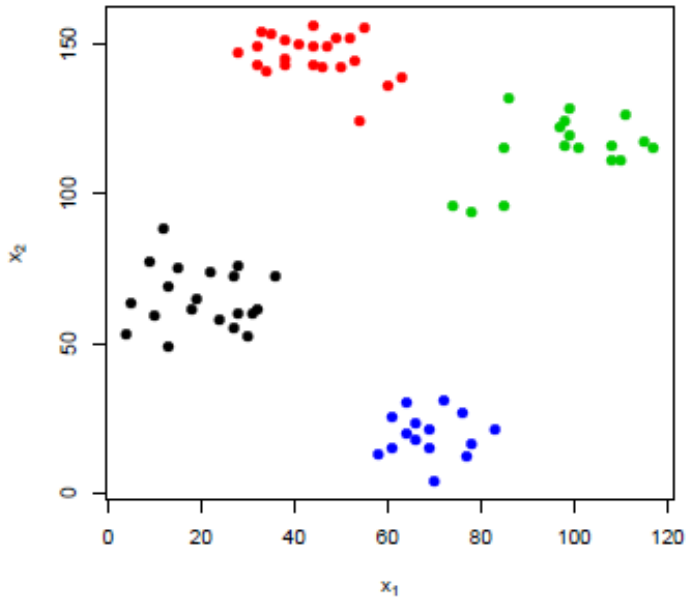
Here we have a data set but without output, that is, $\mathcal{L} = \{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^d$ and we want to create K different homogeneous groups.



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.





Organización del seminario

- Lunes 13/3: Ernesto. Generalidades Aprendizaje Estadístico. Cap 2.
- Lunes 20/3: Mathias. Generalidades Aprendizaje Estadístico. Cap 2.
- Miércoles 29/3: Mathias. Generalidades Aprendizaje Estadístico. Cap 2.
- Miércoles 12/4: Rodrigo y Antonio. Regresión lineal. Cap 3.
- Miércoles 19/4: Rodrigo y Antonio. Regresión lineal. Cap 3.
- Miércoles 26/4: Nacha. Clasificación. Cap 4.
- Miércoles 3/5: Nacha. Clasificación. Cap 4.
- Miércoles 10/5: Verónica. Selección de variables. Cap 6.
- Miércoles 17/5: Verónica. Selección de variables. Cap 6.
- Miércoles 24/5: Mauricio. Arboles de clasificación y regresión. Cap 8.
- Miércoles 31/5: Rodrigo y Antonio. Bagging y Random Forest. Cap 8.
- Miércoles 7/6: Mauricio. Support Vector Machines. Cap 9.
- Miércoles 14/6: Mauricio. Support Vector Machines. Cap 9.
- Miércoles 21/6: Rodrigo y Antonio. Clustering.Cap 10.
- Miércoles 28/6: Colchón