Semantix

Data Science Course

Data Science Course

Contents:

- Main ideas about Data Science;
- Data Science Workflow;
- Landscape of Machine Learning;
- Types of learning
- Hands-on

Ricardo Klein Sercundes

Data scientist @ Semantix

PhD in Statistics and Agricultural Experimentation

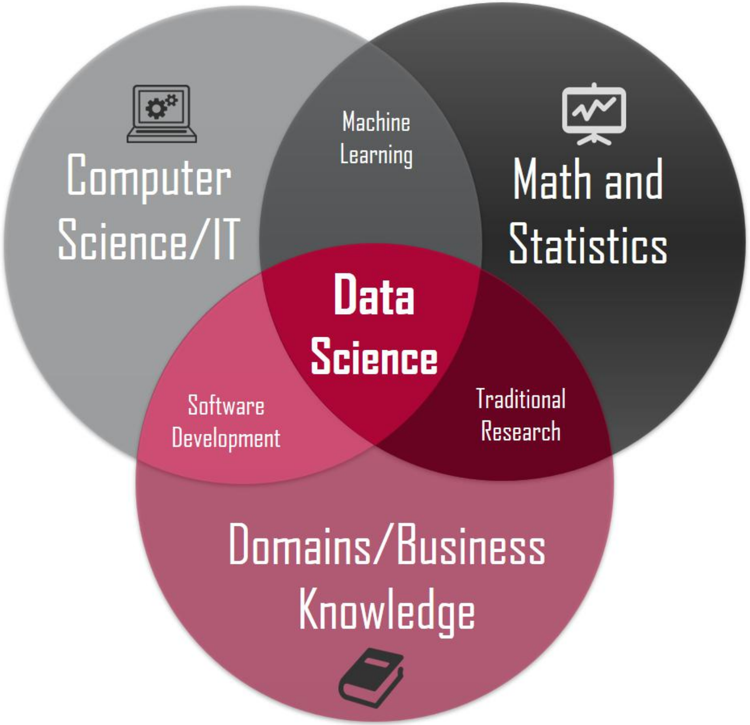
Phone: +55 (19) 99861-3769

email: ricardo.klein@semantix.com.br

What is Data Science?

“Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.”

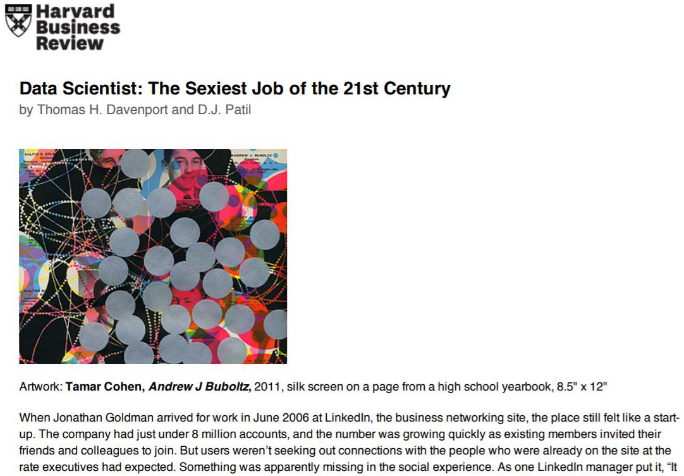
Dhar, V. and Leek, J. (2013)



What is a Data Scientist?

> “A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.”

Wills, J. (2012)



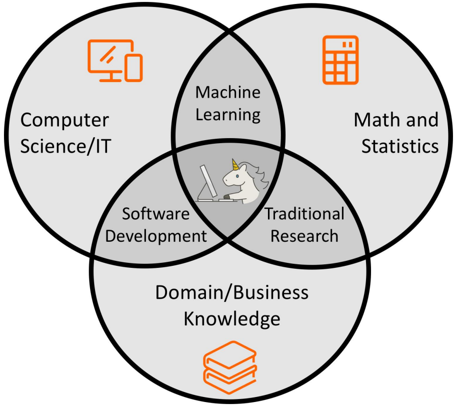
Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Artwork: **Tamar Cohen, *Andrew J Buboltz,*** 2011, silk screen on a page from a high school yearbook, 8.5" x 12"

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren’t seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, “It



Computer Science/IT

Math and Statistics

Domain/Business Knowledge

Machine Learning

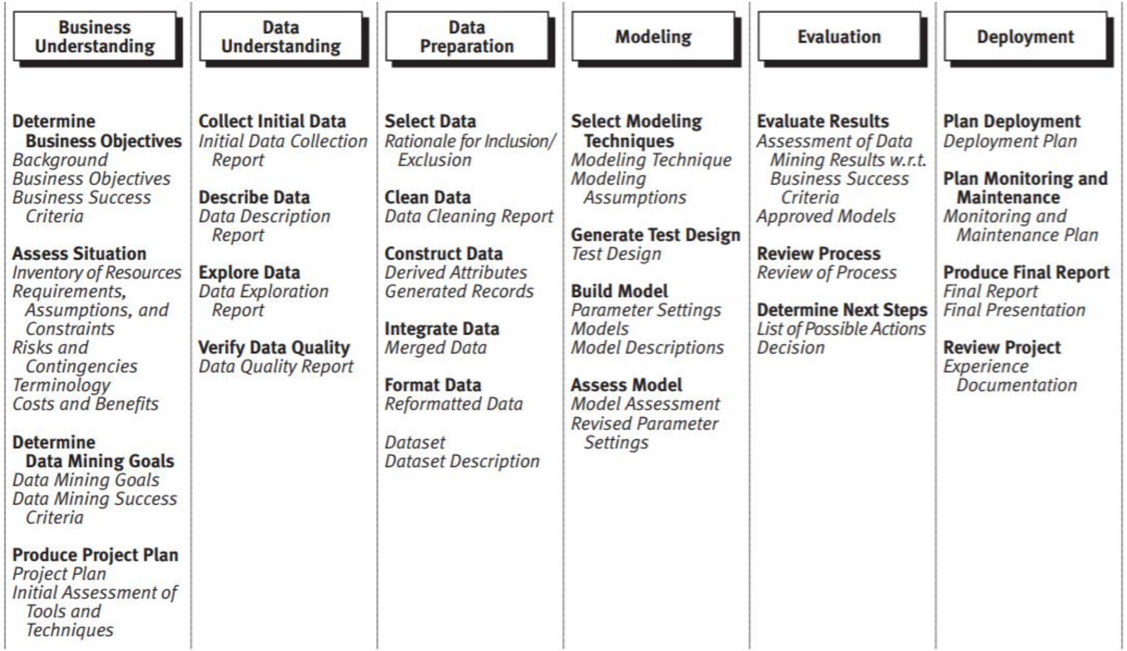
Software Development

Traditional Research

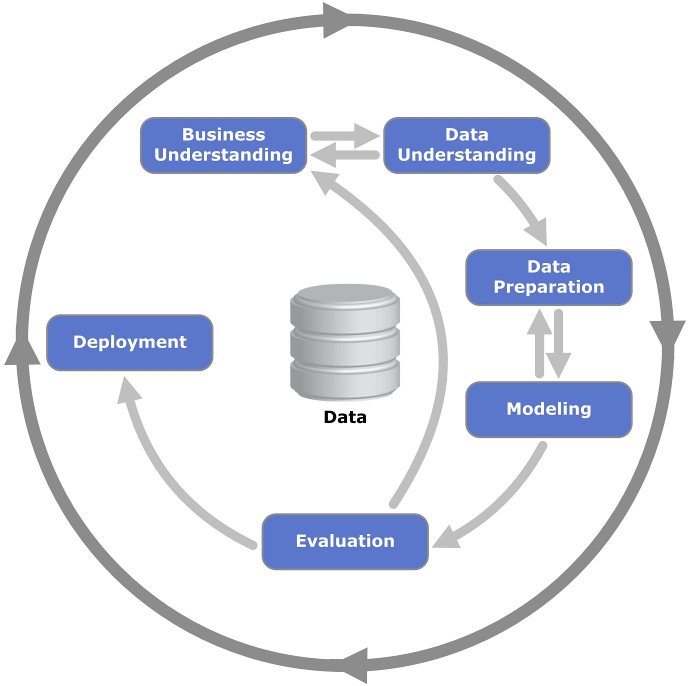
Semantix

Data Science workflow

CRISP-DM methodology



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

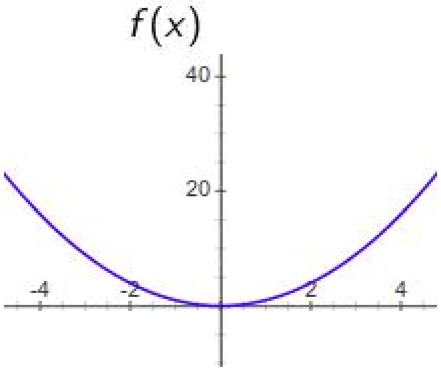


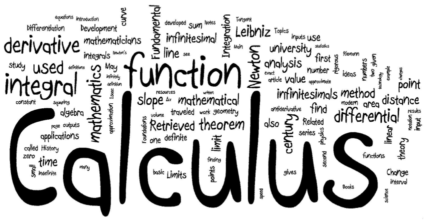
How machine learning algorithms actually “learn”?

The “learning” process involves to fit a theoretical model to the data. To do it, we usually need to maximize a cost function.

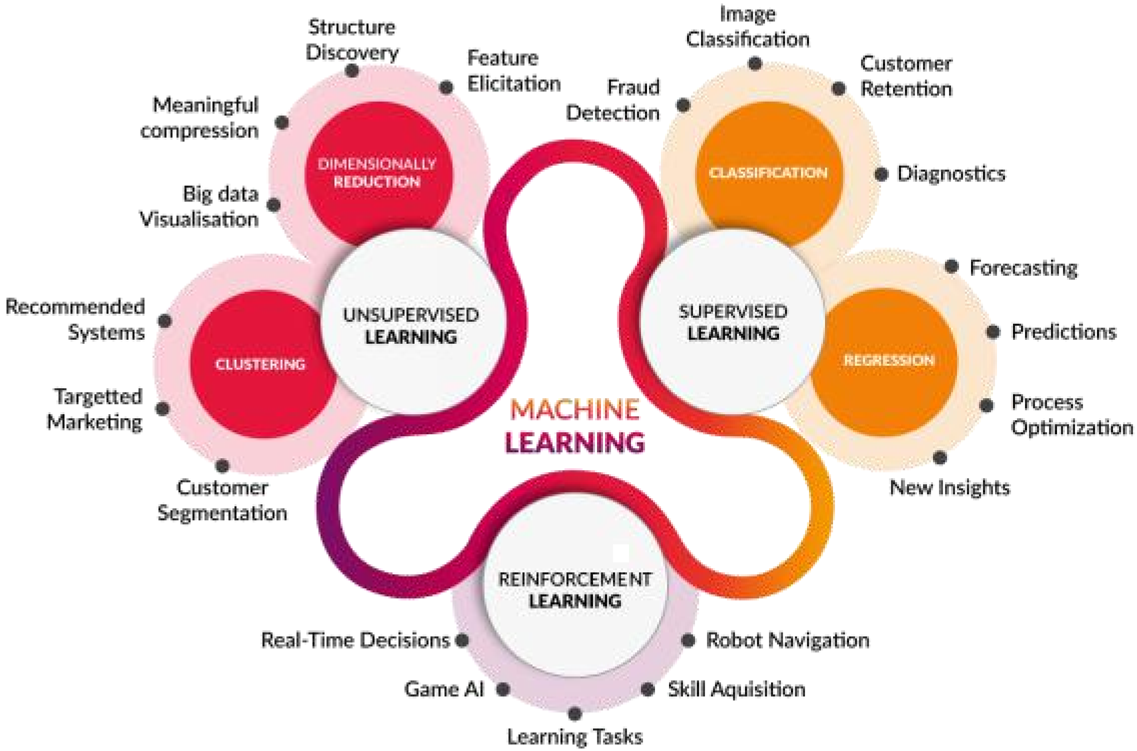
E.g.:

$$\begin{aligned} f(x) &= x^2 \\ f'(x) &= 2x \\ 2x &= 0 \\ x &= 0 \end{aligned}$$





Landscape of Machine Learning



Unsupervised Learning

In unsupervised learning we have a set of variables, also called features but there is no response variable, also called label.

Goals of unsupervised learning

- Find patterns;
- Find clusters, groups of similar observations;
- Dimensionality reduction;
- Find unusual observations (outliers);
- Generate hypotesis about the data.

These types of learning do not use stochastic models, i.e. no hypothesis tests and CIs.

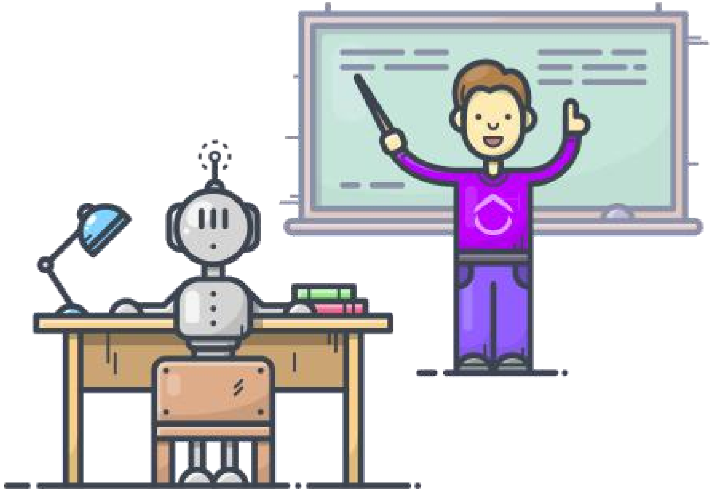
Unsupervised Learning

In this course, we will foccus on the main unsupervised learning methods:

- Elementary methods;
- Clustering algorithms:
  - Hierarchical cluster;
  - K-means.
- Principal component analysis (PCA).

Supervised Learning

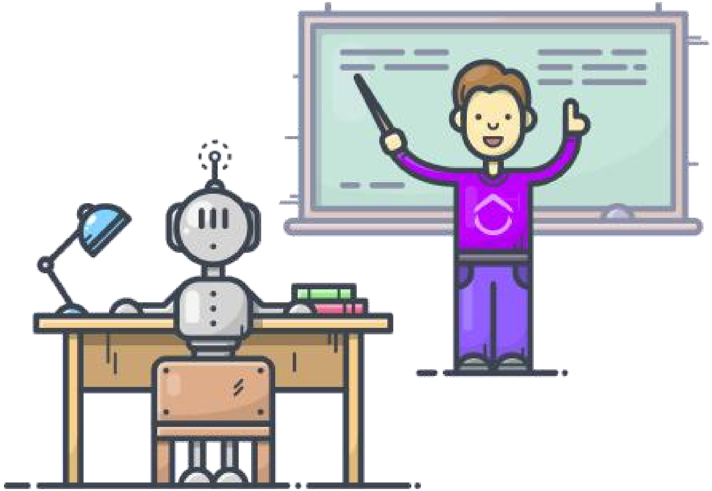
In supervised learning we have a set of variables, also called features and a response variable, also called label.



Goals of supervised learning

- Describe patters;
- Forecast;
- Measure probabilities;
- Test hypotesis;
- Build confidence intervals.

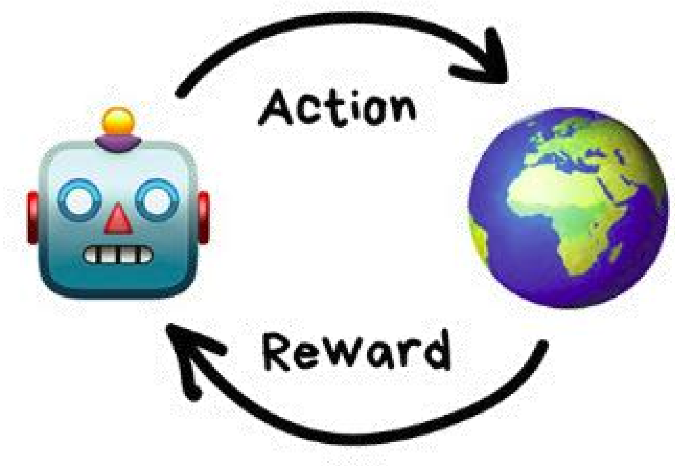
Supervised Learning



In this course, we will foccus on the main supervised learning methods:

- Generalized linear models:
  - Linear regression;
  - Poisson, Quasi-Poisson and Negative Binomial;
  - Logistic regression;
  - Variable selection and likelihood ratio tests.
- Tree-based methods:
  - Classification and Regression trees;
  - Random forests.
- Artificial Neural Networks

Reinforcement Learning



In reinforcement learning, the algorithm tries several events, giving greater weight to the cases that promote success.

Goals of supervised learning

- Find patterns to do an action;
- Learn customers behaviour;
- Etc.

These types of learning are used in navigation, to control traffic lights and mimic balance.

Some questions so far?

Unsupervised Learning

Elementary methods

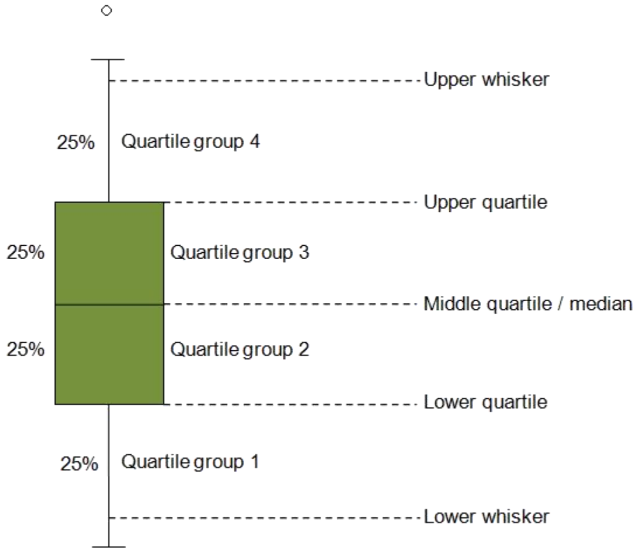
These methods are very simple and will help us to describe a multivariate data.

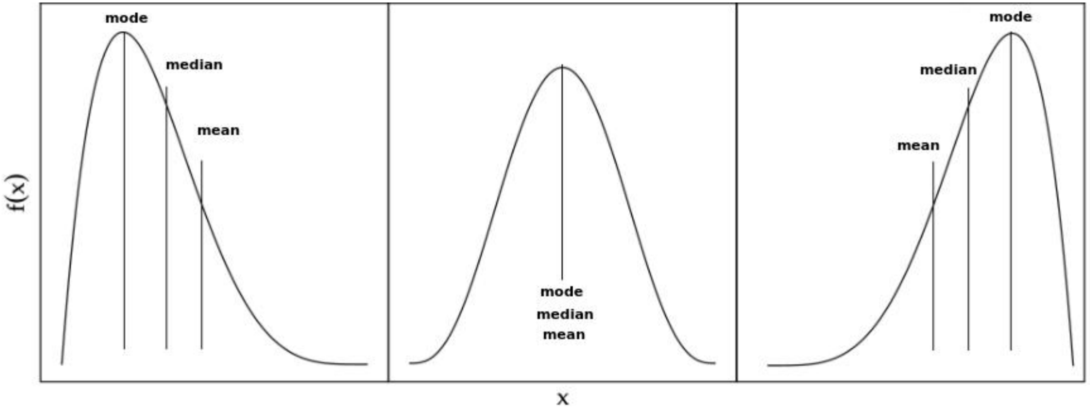
- Box-plots;
- Correlation matrix;
- Chernoff faces & Stars;

Unsupervised Learning

Elementary methods

- Box-plots



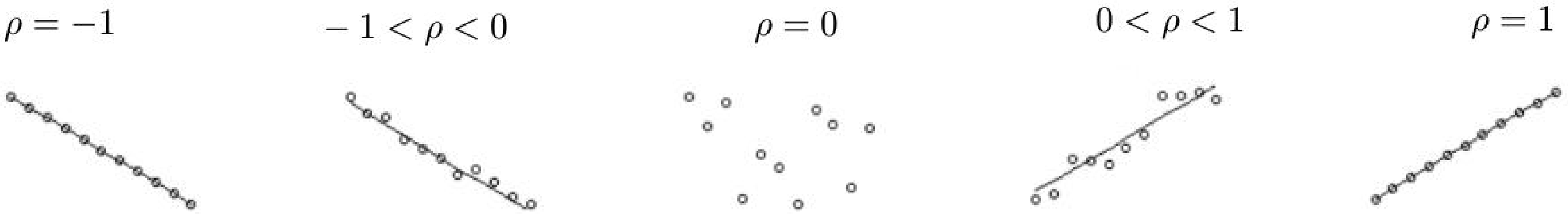


Unsupervised Learning

Elementary methods

- Pearson correlation:

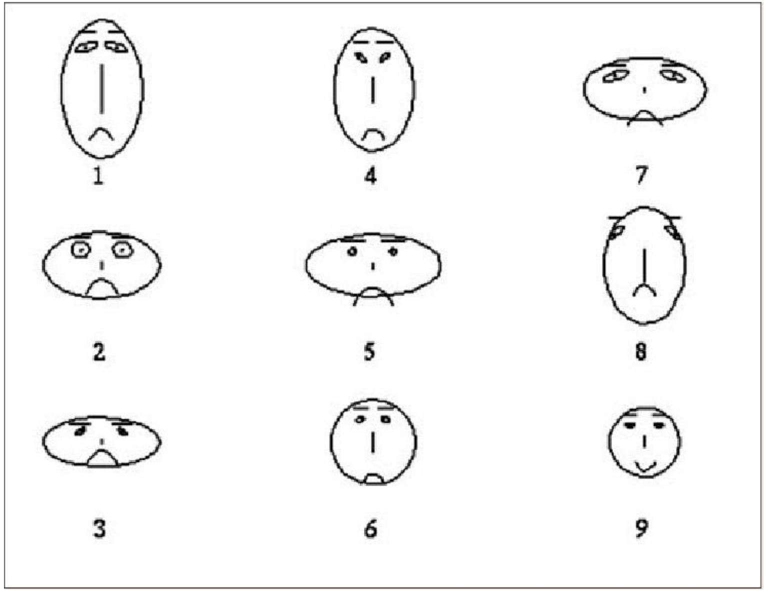
$$\rho = Corr(X, Y) = \frac{n\left(\sum xy\right) - \left(\sum x \sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

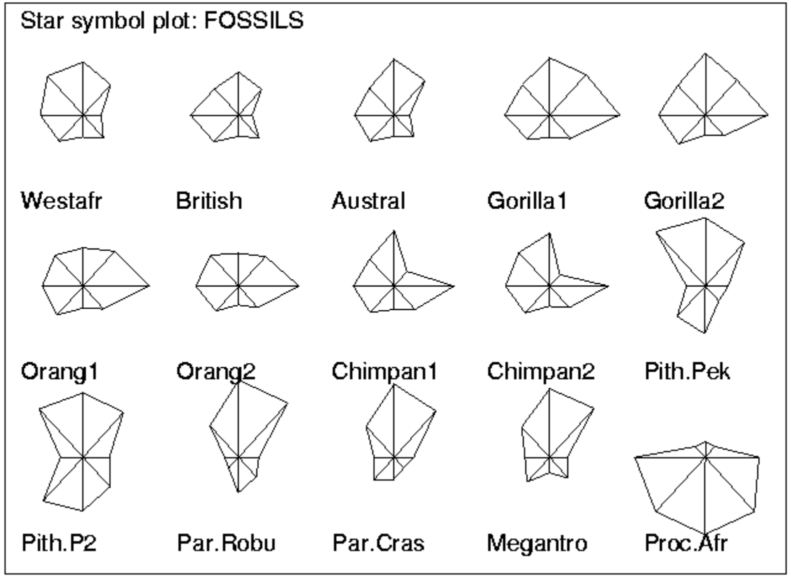


Unsupervised Learning

Elementary methods

- Chernoff faces & Stars





Dataset: Prehistoric dogs from Thailand

Excavations of prehistoric sites in northeast Thailand have produced a collection of canid (dog) bones covering a period from about 3500 B.C. to the present. With these bones, several measures of jaws and teeths were collected. This type of data is usually used to estimate similarities between species. Using this information, how the prehistoric dog seems to relate to the other groups?

Modern dog

Golden jackal

Chinese wolf

Indian wolf

Cuon

Dingo

?

Prehistoric dog

Dataset: Prehistoric dogs from Thailand

Group	x1	x2	x3	x4	x5	x6
Modern dog	9,7	21,0	19,4	7,7	32,0	36,5
Golden jackal	8,1	16,7	18,3	7,0	30,3	32,9
Chinese wolf	13,5	27,3	26,8	10,6	41,9	48,1
Indian wolf	11,5	24,3	24,5	9,3	40,0	44,6
Cuon	10,7	23,5	21,4	8,5	28,8	37,6
Dingo	9,6	22,6	21,1	8,3	34,4	43,1
Prehistoric dog	10,3	22,1	19,1	8,1	32,2	35,0

Note: X1 = breadth of mandible; X2 = height of mandible below the first molar; X3 = length of the first molar; X4 = breadth of the first molar; X5 = length from first to third molar, inclusive; and X6 = length from first to fourth premolar, inclusive.
Source: Adapted from Higham, C.F.W. et al. (1980), J. Archaeological Sci., 7, 149–165.

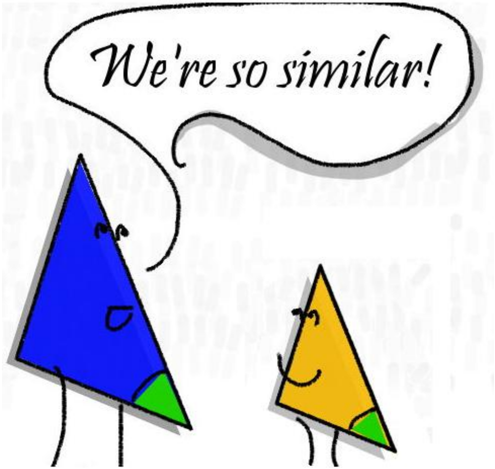
Unsupervised Learning

Clustering algorithms

The goal is create groups based in some measures of similarity.

How can I measure similarity?

- Distances:
  - Euclidean;
  - Manhattan;
  - Mahalanobis;
  - Etc.



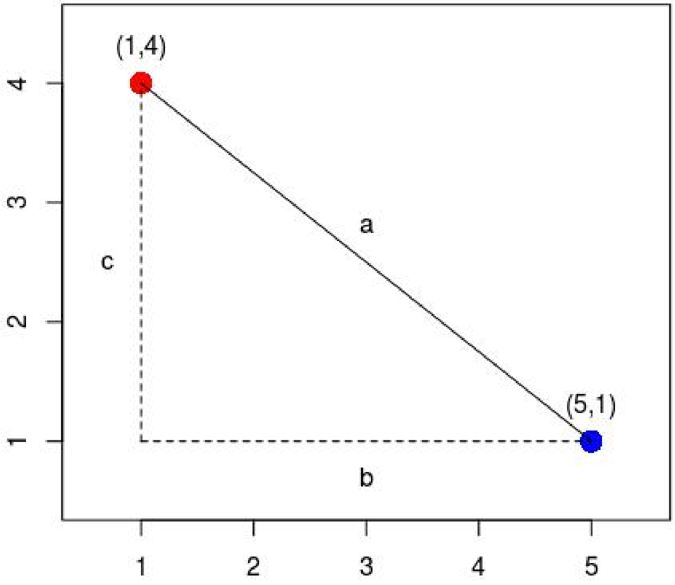
Unsupervised Learning

Clustering algorithms

- Pythagoras' Theorem: $a = \sqrt{b^2 + c^2}$
- Euclidean distance:

$$\begin{bmatrix} 5 & 1 \\ 1 & 4 \end{bmatrix}$$

$$\begin{aligned} d(i,j) &= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \\ &= \sqrt{(5-1)^2 + (4-1)^2} \\ &= \sqrt{16+9} \\ &= 5 \end{aligned}$$



Unsupervised Learning

Clustering algorithms

Note:

- If the variables are measured in different units or have different scales then it is advisable to standardise variables prior to computing distances.

Distribution standardization

$$z_i = \frac{x_i - \mu}{\sigma}$$

Range standardization

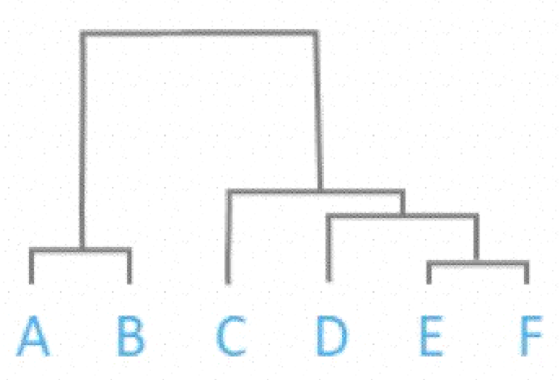
$$z_i = \frac{x_i - x_{(min)}}{x_{(max)} - x_{(min)}}$$

Unsupervised Learning

Hierarchical clustering (HC)

Hierarchical clustering it is a very nice tool to find how similar some measurements are. The result of this method is an attractive tree-based representation of the observations (dendrogram).

Here, we will foccus on the most common type of hierarchical clustering, also called bottom-up or agglomerative clustering.



Unsupervised Learning

Hierarchical clustering (HC)

Steps to build a HC:

You have n objects. You wish to form k clusters.

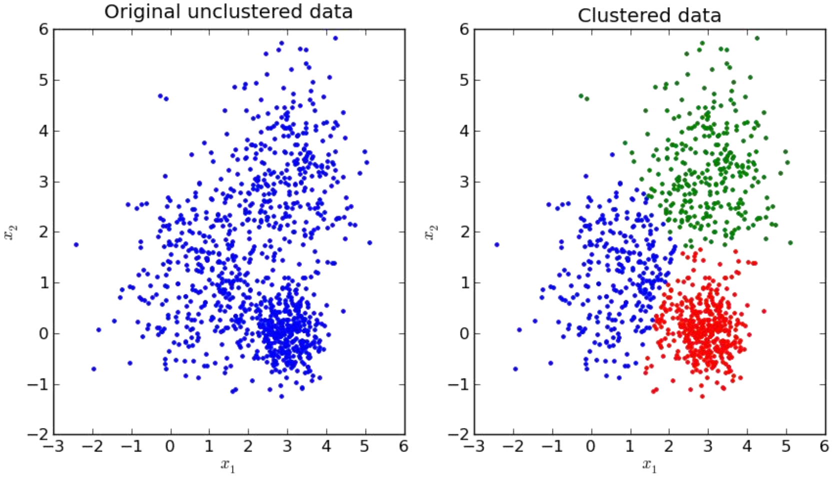
Construct a distance matrix D.

- 1. Start with m = n clusters.
- 2. Find the closest pair of clusters and merge them. Now there are m= n-1 clusters.
- 3. Repeat step 2 until m = k

Unsupervised Learning

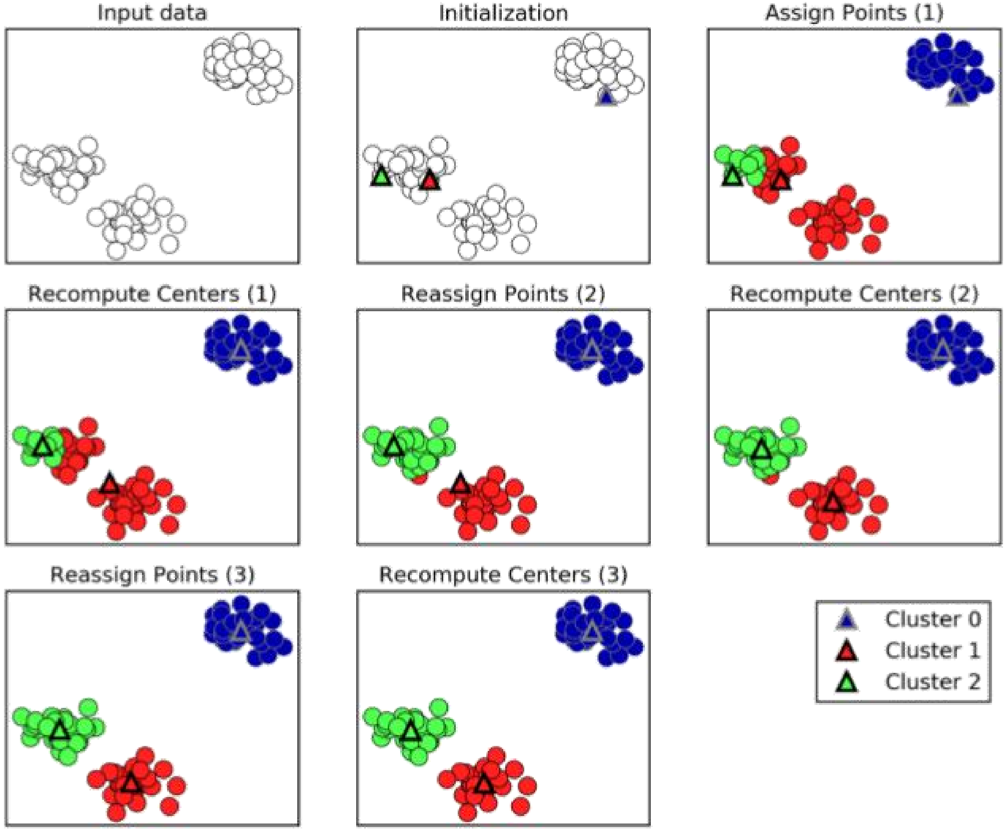
K-means

This is a simple and elegant approach for partitioning a data set into K distinct, non-overlapping clusters. To perform K-means clustering, we must first specify the desired number of clusters K; then the K-means algorithm will assign each observation to exactly one of the K clusters.



Unsupervised Learning

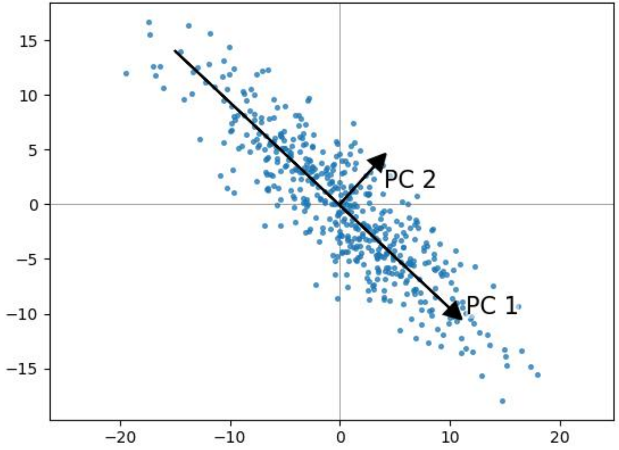
K-means



Unsupervised Learning

Principal components analysis (PCA)

For data with many variables/dimensions it is difficult to comprehend or visualize the associations between them. Thus, PCA ‘re-expresses’ the data to account for most of the information in the data through a few linear combinations of the original variables.



Unsupervised Learning

Principal components analysis (PCA)

- Dimensionality reduction

$$\begin{array}{c}
\begin{array}{cccc} x_1 & x_2 & & x_n \end{array} \\
\begin{bmatrix} 10.3 & 15.2 & \dots & 55.6 \\ 8.3 & 11.5 & \dots & 67.7 \\ \vdots & \vdots & \ddots & 20.3 \\ 10.4 & 7.6 & \dots & 70.2 \end{bmatrix}
\end{array}$$

$$\begin{aligned} Z_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ Z_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \end{aligned}$$

Some questions so far?

Semantix

Supervised Learning

Generalized linear models

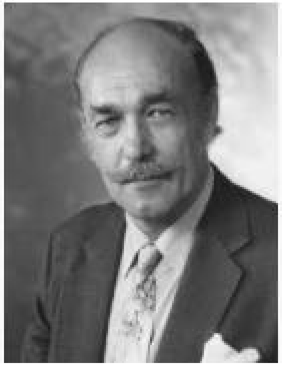
Unifying framework for much statistical modelling. First introduced by Nelder & Wedderburn (1972) as an extension to the standard normal theory linear model.

The GLM is defined by three components:

- Response variable or label that belong to the exponential

family of distributions (Gaussian, Poisson, Binomial, Gamma, etc);

- Linear predictor: covariates or features;
- Link function: function to link the outcome with the covariates.



Nelder

Wedderburn

Supervised Learning

Generalized linear models

- Parameter estimation: Maximum likelihood:

$$\mathbf{L}(\boldsymbol{\beta};\mathbf{y}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\beta})$$

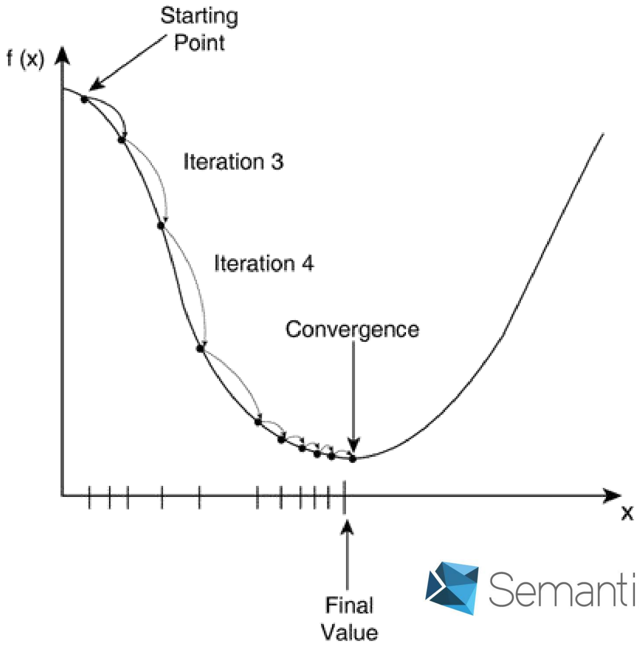
$$= \prod_{i=1}^{n} \cancel{\frac{n_i!}{y_i!\,(n_i - y_i)!}} \pi_i^{y_i}(1-\pi_i)^{n_i - y_i}$$

$$= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{n_i - y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\mathrm{X}_i\boldsymbol{\beta}}}{1+e^{\mathrm{X}_i\boldsymbol{\beta}}}\right)^{y_i} \left(\frac{1}{1+e^{-\mathrm{X}_i\boldsymbol{\beta}}}\right)^{n_i - y_i}$$

E.g.: Binary outcomes

$$\boxed{\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \boldsymbol{X}_i \beta \; - \log\left(1 + e^{\mathrm{X}_i \boldsymbol{\beta}}\right) \right]}$$



Supervised Learning

Generalized linear models

Normal models - Linear regression

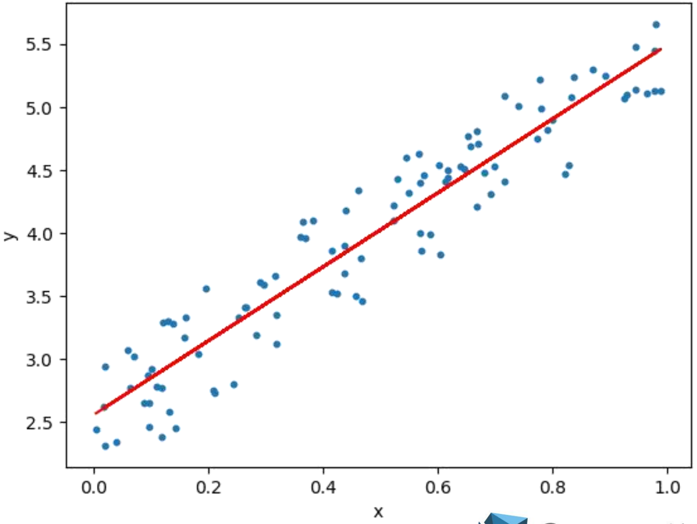
Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple.

- The meaning of “linear” is related with the parameters;
- Continuous response variable.

$$Y_i \sim N(\mu_i, \sigma^2), \quad i = 1, \dots, n$$

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} = \beta^T \mathbf{x}_i$$

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



Supervised Learning

Generalized linear models

Normal models - Linear regression

Goodness-of-fit: R^2

The coefficient of determination is a measure used in statistical analysis that assesses how well a model explains the outcomes. It is indicative of the level of explained variability in the data set.

$$R^2 = \frac{\text{SQReg}}{\text{SQtotal}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Drawbacks: R^2 increases as the number of predictors increases.

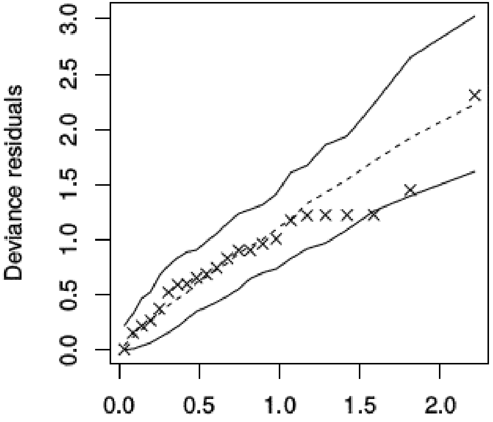
Supervised Learning

Generalized linear models

Normal models - Linear regression

Goodness-of-fit: Half-normal plot with simulated envelope

It detects outliers and indicates whether the error distribution was specified appropriately.



Supervised Learning

Generalized linear models

Normal models - Linear regression

Model selection: AIC and BIC

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are estimators of the relative quality of statistical models for a given set of data.

$$AIC = -2\log\left(L(\beta, y)\right) + 2[(p+1)+1]$$

$$BIC = -2\log\left(L(\beta, y)\right) + [(p+1)+1]\log(n)$$

In both criterion, small values are preferred.

Dataset: Advertising

Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product. The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

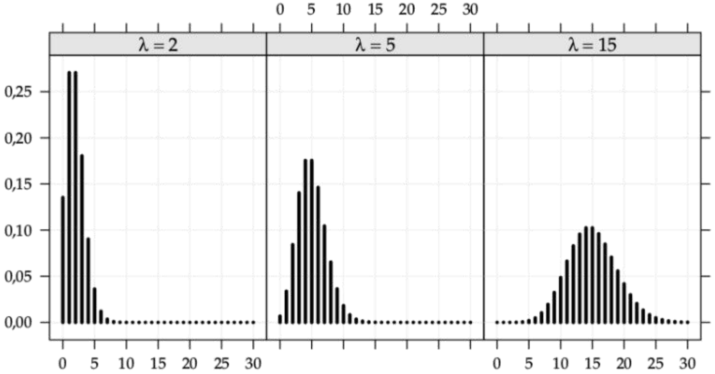
Supervised Learning

Generalized linear models

Poisson, Quasi-Poisson and Negative Binomial regression models

All these models are used when the outcomes are counts. They come from Poisson distribution.

$$f(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2 \ldots, \quad \lambda_i > 0$$



Supervised Learning

Generalized linear models

Poisson, Quasi-Poisson and Negative Binomial regression models

All these models are used when the outcomes are counts. The difference between then is the variance function.

$$E(Y_i) = \lambda_i$$

Poisson distribution	Quasi-Poisson	Negative Binomial
$Var(Y_i) = \lambda_i$	$Var(Y_i) = \phi\lambda_i$	$Var(Y_i) = \lambda_i + \lambda_i^2\phi$

Dataset: *Diaphorina citri*

In an experiment to assess the effect of three agricultural oils on the oviposition of *Diaphorina citri*, 70 Orange Jessamine (*Murraya paniculata*) plants were sprayed with seven solutions of the mineral oils. The experiment used the oils in a completely randomized design with ten replicates. Following treatment, when the plants were dry, ten pregnant females of *D. citri* were released on each plant. After 5 days, the insects were removed and the total number of eggs on each plant was observed.

Supervised Learning

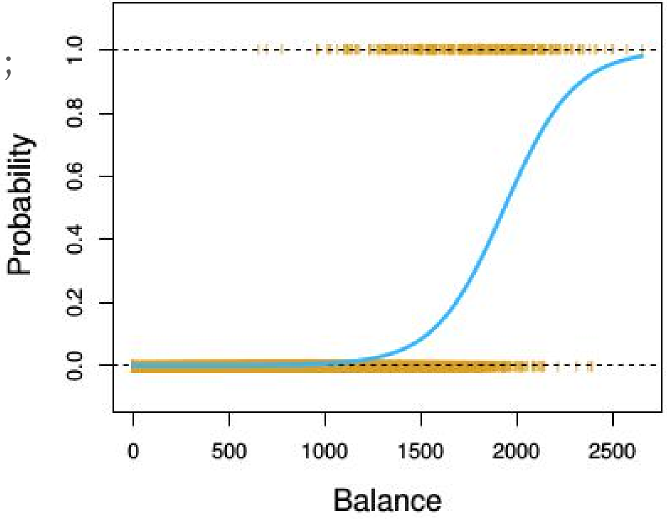
Generalized linear models

Logistic regression

The outcomes are binary (sucess/failure)

- Logistic regression always gives probabilities between 0 and 1;
- The decision boundary is 0.5.

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



Dataset: Loan bank's prospection

Banks and other financial institutions have used loan as a very profitable product. Models to predic the probability of a client acquire these type of products are often used. In this dataset, the financial institutions would like to know the relationship between take a loan, income and balance.

Supervised Learning

Generalized linear models

Variable selection and likelihood ratio tests in parametric models

There are several ways to do a variable selection. Here, we will focus on methods based on likelihood ratio tests such as:

Backward

$$M_1: \quad y_i = \frac{1}{1 + e^{-(\boldsymbol{\beta}_0)}}$$

$$M_2: \quad y_i = \frac{1}{1 + e^{-(\boldsymbol{\beta}_0 + \beta_1 X_1)}}$$

Forward

$$M_3: \quad y_i = \frac{1}{1 + e^{-(\boldsymbol{\beta}_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

$$M_4: \quad y_i = \frac{1}{1 + e^{-(\boldsymbol{\beta}_0 + \beta_2 X_2 + \cdots + \cancel{\beta_n X_n})}}$$

Stepwise

$$M_5: \quad y_i = \frac{1}{1 + e^{-(\boldsymbol{\beta}_0 + \beta_1 X_1)}}$$

$$M_6: \quad y_i = \frac{1}{1 + e^{-(\boldsymbol{\beta}_0 + \beta_1 X_1 + \beta_2 X_2)}}$$

$$M_7: \quad y_i = \frac{1}{1 + e^{-(\boldsymbol{\beta}_0 + \beta_2 X_2)}}$$

Supervised Learning

Generalized linear models

Variable selection and likelihood ratio tests in parametric models

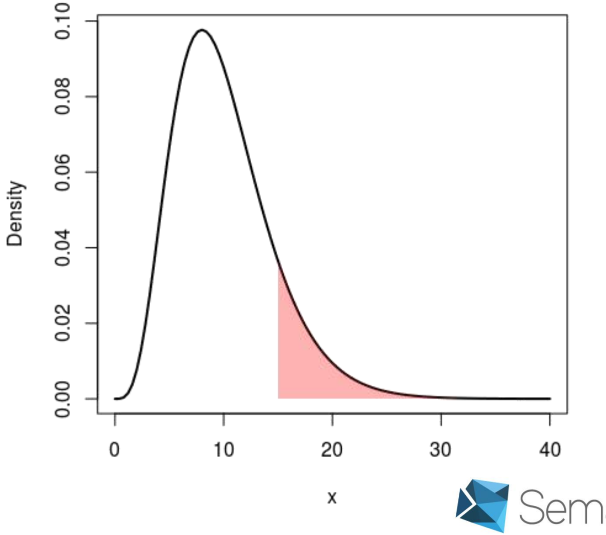
How does it work? $M_1: \quad y_i = \frac{1}{1+e^{-(\boldsymbol{\beta}_0)}}$ vs $M_2: \quad y_i = \frac{1}{1+e^{-(\boldsymbol{\beta}_0+\beta_1 X_1)}}$

Hypotesis: $H_0 : \beta_1 = 0 \quad vs \quad H_a : \beta_1 \neq 0$

$$\boldsymbol{\Lambda} = -2\left(\ell(\widehat{\boldsymbol{\beta}}^{(r)}) - \ell(\widehat{\boldsymbol{\beta}}^{(p)})\right),$$

with $p > r$

Check with $\chi^2_{(p-r)}$



Semantix

Supervised Learning

Tree-based methods

Classification and Regression trees

- Non-parametric methods;
- These methods stratifying the predictor space into a number of simple regions;
- The predictor space can be summarized into a tree diagram (easy to understand);
- Handle classification and regression tasks;
- They don’t need data transformation.

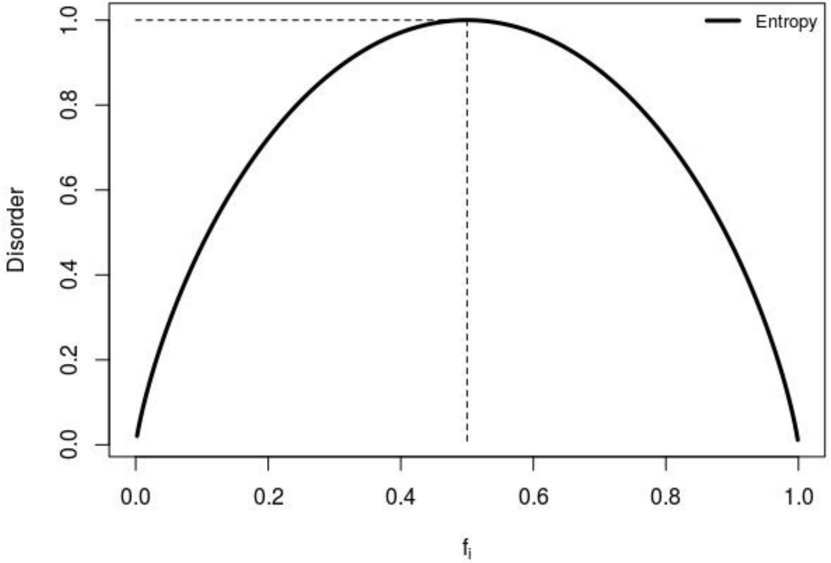
Supervised Learning

Tree-based methods

Classification and Regression trees

Cost functions:

- Classification: Entropy or Gini index;
- Regression: Variance.

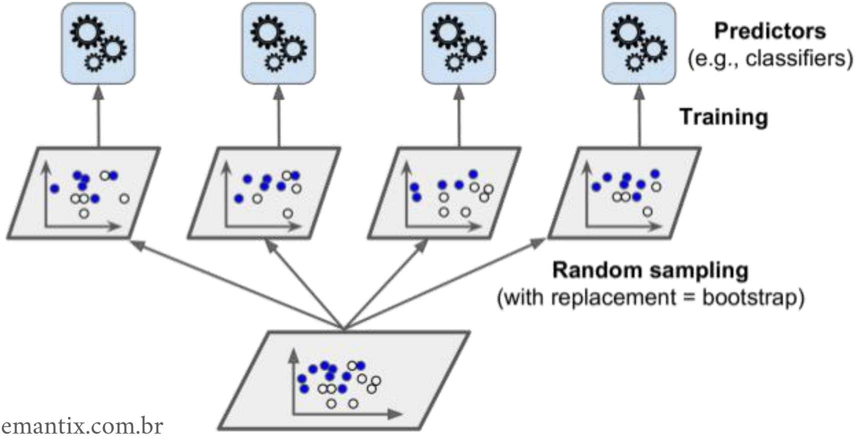


Supervised Learning

Tree-based methods

Random forests

Basic tree-based methods suffer from high variance (ie. if we reperatedly fitted the model to different training sets we expect a large variation in results). To reduce this drawback, we can obtain a number of bootstrapped training sets, build a tree on each one (without pruning) and average the resulting predictions for each observation.



Dataset: Iris flower

The data set consists of 50 samples from each of three species of Iris (*Iris setosa, Iris virginica and Iris versicolor*). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. If we find an Iris flower, how can we classify it based on this informations?

I. setosa

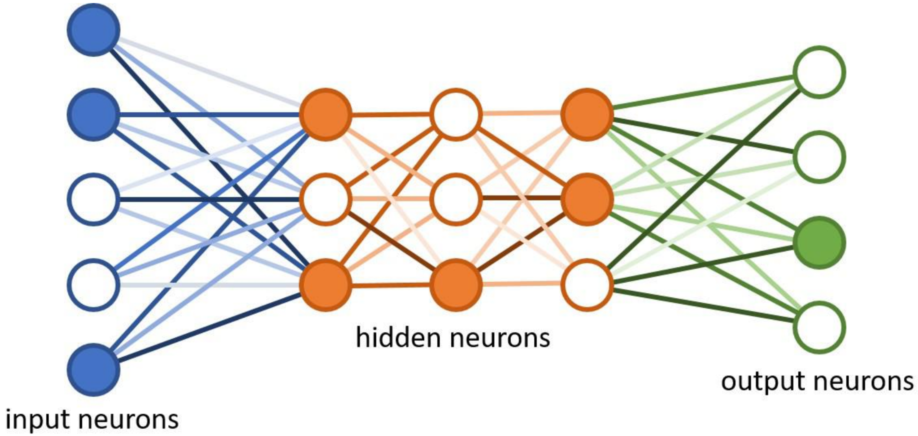
I. versicolor

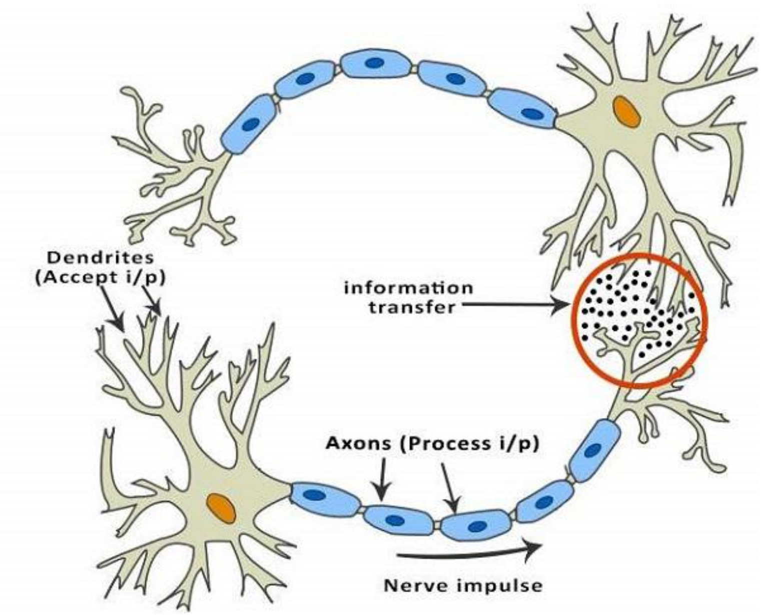
I. virginica

Supervised Learning

Atificial Neural Networks

Artificial neural networks are one of the main tools used in machine learning. As the “neural” part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn.





Thanks!

Semantix

Dataset: California housing price

This dataset presents a census survey conducted in California. Through covariates, try to understand which factors affect house prices and in which regions they are most expensive.