Data Mining - Overview

Dr. Jean-Michel RICHER

2018

jean-michel.richer@univ-angers.fr

Outline

1. Introduction

Definition 1

Data Mining (Minería de Datos) is the extraction of implicit, previously unknown, and potentially useful information from data.

Definition 1

Data Mining (Minería de Datos) is the extraction of implicit, previously unknown, and potentially useful information from data.

Definition 2

Data Mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

What do they have in common ?

Common terms

- discovery, extraction (not obvious)
- useful information, patterns, rules (form of the useful information)
- data, database (potentially a huge amount of data)
- statistics, machine learning, ... (methods)

What is data mining ?

Definition

Data Mining is a branch of computer science which is either

- a knowledge discovery process of hidden or non trivial information (ex: supermarket rules)
- or a knowledge synthesis process used to improve the understanding of raw data (ex: flower classification)

More generally

- given a set of Objects $\mathcal{X} = \{\vec{X_1}, \ldots, \vec{X_n}\}$ (persons, customers, patients, plants, ...)
- defined by a set of properties $P = \{P_1, \ldots, P_k\}$ (length, size, weight, color, ...)
- each P_i having discrete or continuous values
- we want to find
  - patterns / rules than can explain the relationships between objects and their properties
  - classify objects

What are properties ?

A note on properties

Properties can be

- **quantitative:** a numerical value
- **qualitative:** a text

Example

For temperatures:

- **quantitative and continuous:** $[-20°C, +40°C]$
- **qualitative and discrete:** very cold, cold, ..., hot, very hot

From data to knowledge

Objects	P_1	P_2	...	P_k
$\vec{X_1}$	x_1^1	x_1^2	...	x_1^k
...				
$\vec{X_n}$	x_n^1	x_n^2	...	x_n^k

if $(x_i^1 > 10)$ and $(x_i^3 == \text{true})$ then $x_i^7 = \text{humid}$

$$x_i^1 = 1.23 \times x_i^2 + 2.56 x_i^3$$

$$c_1 = \{\vec{X_1}, \vec{X_3}\}, c_2 = \{\vec{X_2}, \vec{X_4}, \vec{X_5}\}$$

Example (1/3)

Truth table

Consider the following truth table in logic (0 = *false*, 1 = *true*):

X	Y	Z	$f(X,Y,Z)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

How can you summarize the function $f(X,Y,Z)$?

Example (2/3)

Boolean expression

You can compute the boolean expression of $f(X, Y, Z)$ as :

$$f(X, Y, Z) = \bar{X}.Y.Z + X.\bar{Y}.Z + X.Y.\bar{Z} + X.Y.Z$$

Example (2/3)

Boolean expression

You can compute the boolean expression of $f(X, Y, Z)$ as :

$$f(X, Y, Z) = \bar{X}.Y.Z + X.\bar{Y}.Z + X.Y.\bar{Z} + X.Y.Z$$

Majority function

It is the majority function, such that $f(X, Y, Z) = 1$ if the number of variables (X, Y, Z) set to 1 is greater than the number of variables set to 0.

Example (3/3)

Majority function

With the majority function definition, you can extend the function with a greater number of input variables and keep the same definition :

- $f(X, Y, Z, W)$
- $f(X, Y, Z, W, T)$
- ...

2. What will we cover ?

We will cover

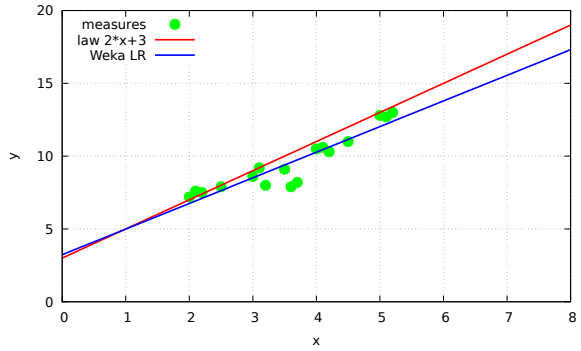
Theoretical and practical aspects of

- regression
- classification
- clustering
- decision tree
- neural networks

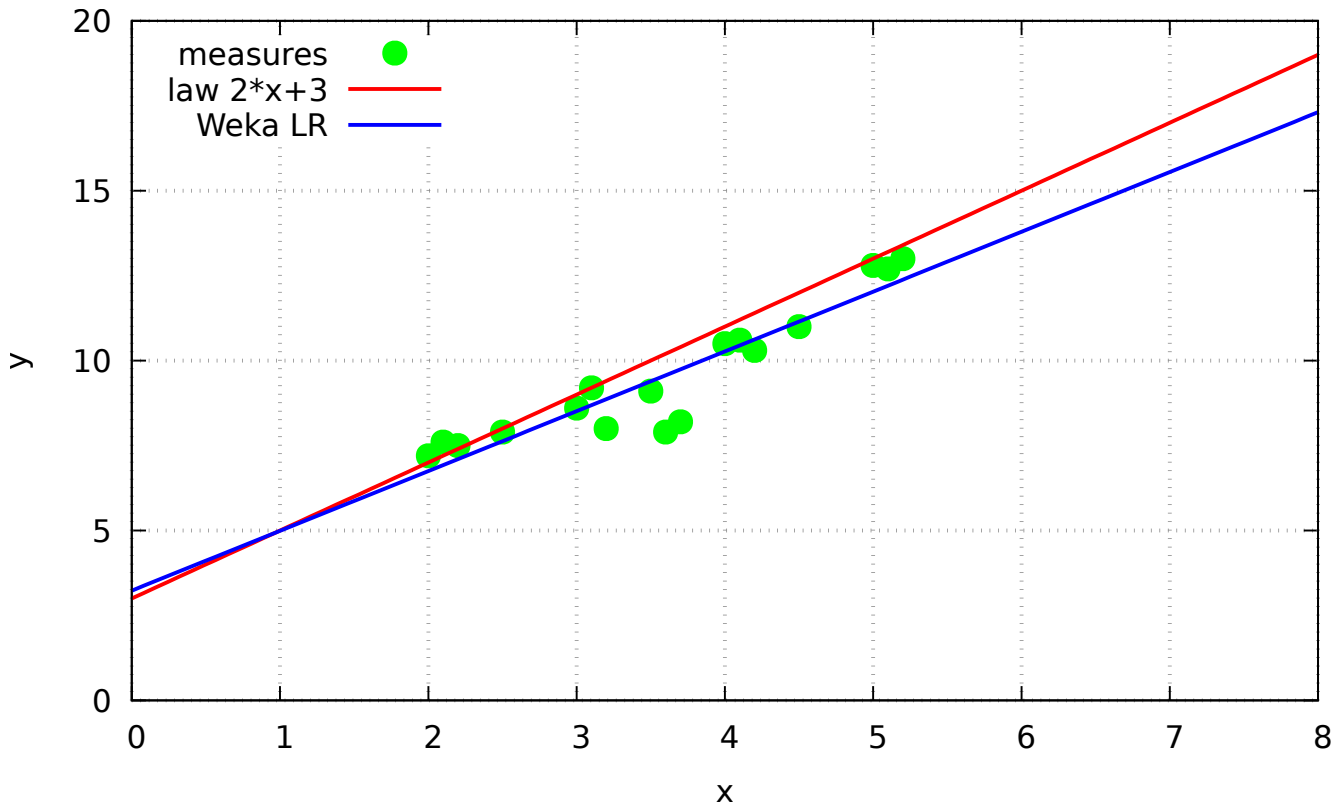
We will cover

Regression

- **Principle:** try to compute a property in function of other properties
- example: $weight = f(height, age, sex)$



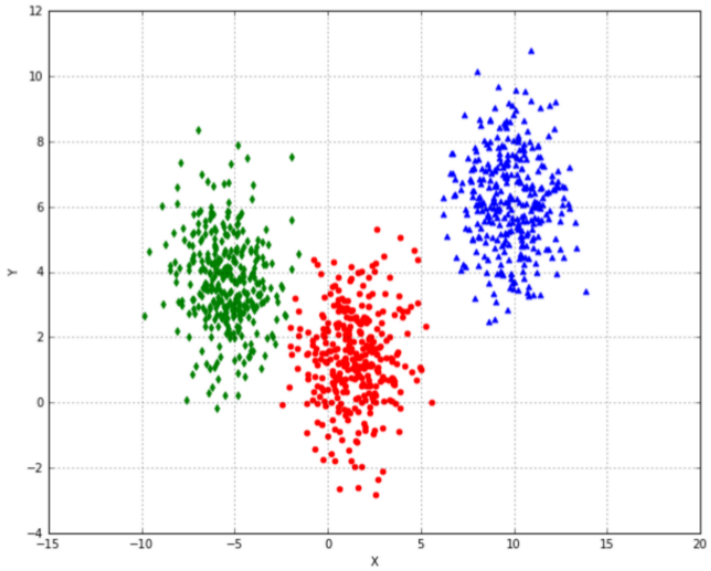
We will cover



We will cover

Classification

- **Principle:** try to split data into groups using the properties of objets, knowing the number of groups
- part of supervised learning



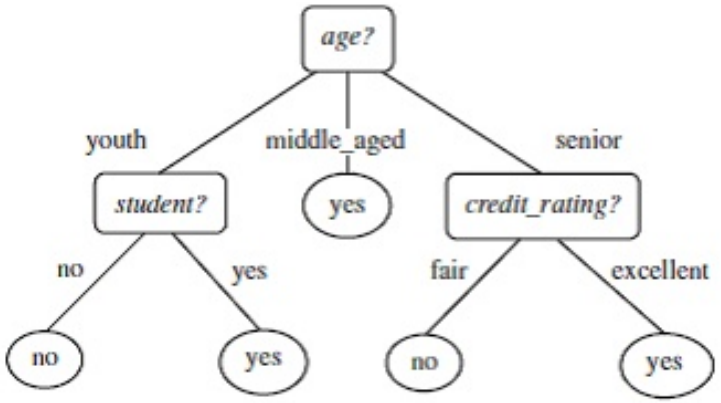
We will cover

Clustering

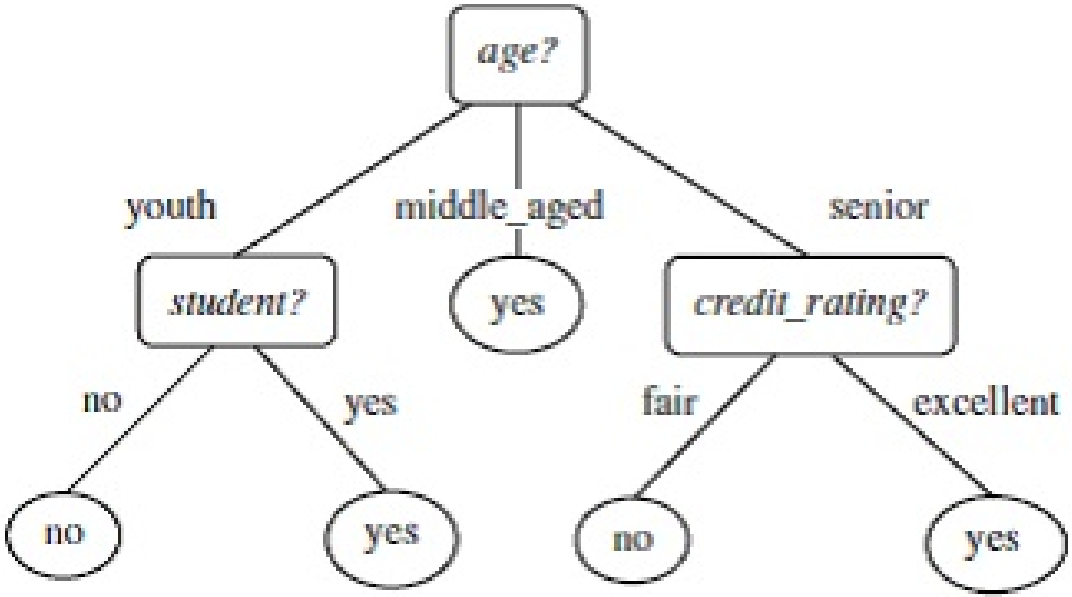
- **Principle:** try to create groups of objects based on their similarities, the number of groups is unknown
- part of unsupervised learning

Decision tree

- **Principle:** create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features
- a non-parametric supervised learning method used for classification and regression

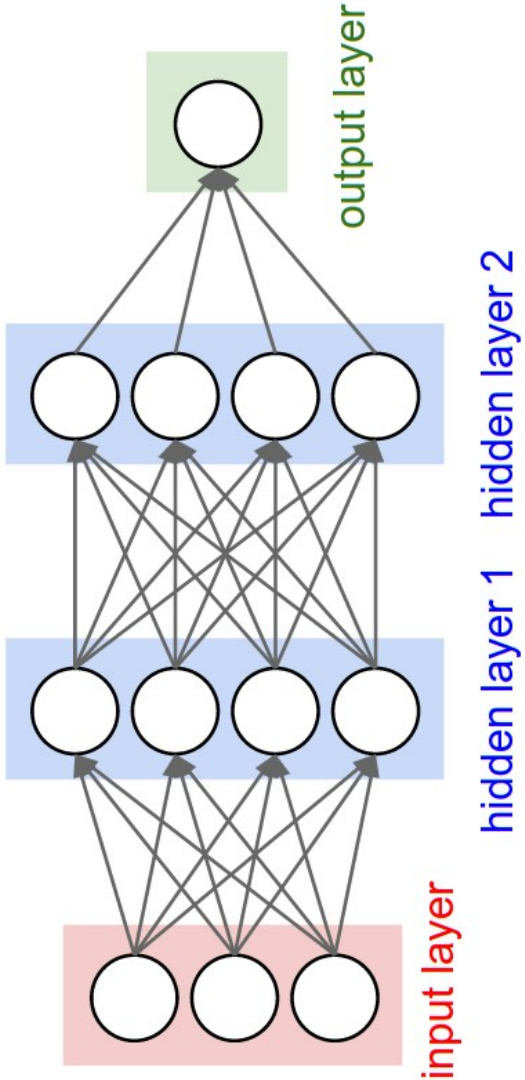


We will cover



Neural Networks

- also called Artificial Neural Networks (ANNs) or Connectionist Systems
- **Principle:** a computing device inspired by biological neural networks that has the ability to progressively learn by considering examples (train set)
- kind of a black-box



Software

We will use the following software:

- WEKA Waikato Environment for Knowledge Analysis, `http://www.cs.waikato.ac.nz/ml/weka`
- R (rio, dplyr, ggplot2)
- Python (numpy, scipy, sklearn, matplotlib)

3. Skills

Required skills

Skills you need to have

Before entering the course you should have

- mathematical background (matrix, statistics, probabilities)
- computer science background (Java, R or Python)

Skills to acquire

After this course you shouldbe able to

- give a definition and explain what is data mining
- define the different kind of analysis / analytics of Data Mining
- chose the correct method / algorithm to produce an analysis in function of the data and the question you have to answer

3. End

FACULTÉ
DES SCIENCES
Unité de formation
et de recherche
DÉPARTEMENT
INFORMATIQUE

UA - Angers
2 Boulevard Lavoisier 49045 Angers
Cedex 01
Tel: (+33) (0)2-41-73-50-72

université
angers
FACULTÉ
DES SCIENCES
Unité de formation
et de recherche
DÉPARTEMENT
INFORMATIQUE