PRESENTATION

BIG DATA INTELLIGENCE

METHODS AND TECHNOLOGIES

A.K.A. MACHINE LEARNING I

RICARDO ALER MUR (ALER@INF.UC3M.ES). 2.2B29 (LEGANÉS)

MASTER IN BIG DATA ANALYTICS

GOALS

1. To introduce **Machine Learning** basics: training, testing, models, hyper-parameter tuning, etc., and some advanced methods (Gradient Boosting, …)

2. Machine Learning in a **Big Data** context

3. To apply them in practice with current **tools** (scikit-learn and Spark-ML)

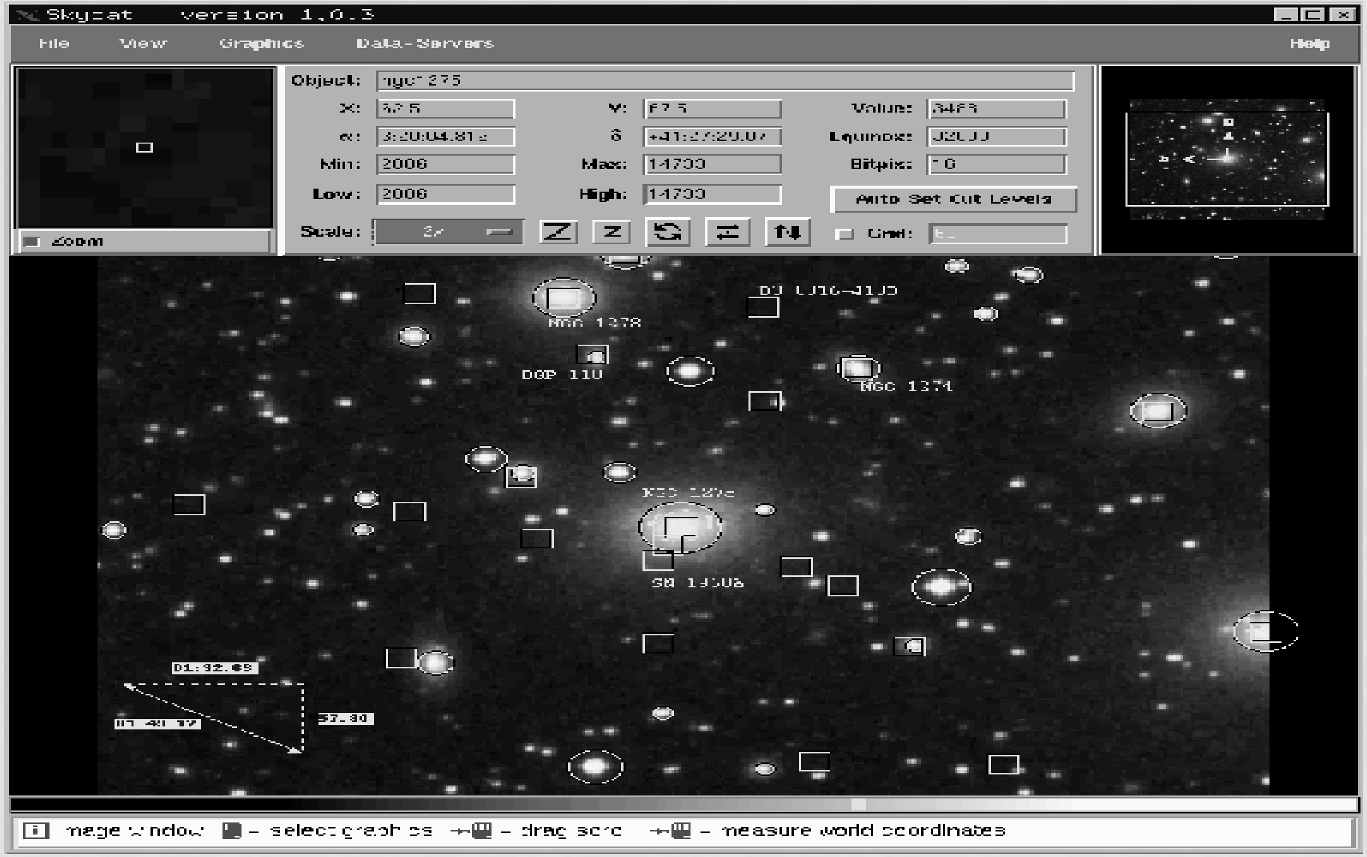
MACHINE LEARNING

- Formally, it's a subfield of **Artificial Intelligence** that tries to make computers and machines learn

- In practice, it tries to create models from data (data is the experience out of which machine learning methods learn a model from)

WHAT IS MACHINE LEARNING

- Example: Skycat: AUTOMATIC CLASSIFICATION OF OBJECTS IN THE SKY



Skycat version 1.0.3

Object: ngc 275

X: 32.5 Y: 67.5 Value: 3483

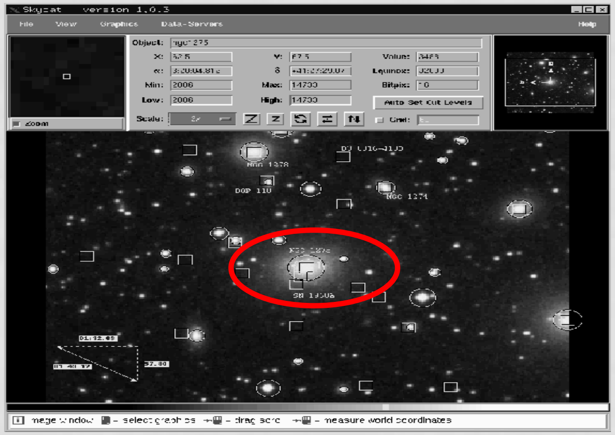
α: 3:20:04.812 δ: +41:27:20.07 Equinox: J2000

Min: 2006 Max: 14700 Bitpix: 16

Low: 2006 High: 14700

Scale: 2x

□ Grid:



Skycat version 1.0.3

File View Graphics Data-Servers Help

Object: ngc1275

X: 32.5 Y: 67.5 Value: 3483

α: 3:20:04.81s δ +41:27:29.07 Equinox: J2000

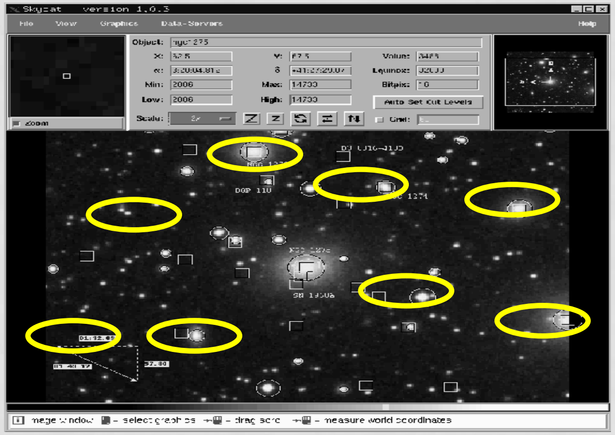
Min: 2006 Max: 14733 Bitpix: 16

Low: 2006 High: 14733 Auto Set Cut Levels

Scale: 2x □ Grid:

?

Training data (labeled pictures of sky objects: galaxies, stars, nebulae, …)



Skycat version 1.0.3

File View Graphics Data-Servers Help

Object: ngc1275

X: 32.5 Y: 67.5 Value: 3483

α: 3:20:04.81s δ +41:27:29.07 Equinox: J2000

Min: 2006 Max: 14733 Bitpix: 16

Low: 2006 High: 14733 Auto Set Cut Levels

Scale: 2x □ Grid:

Pictures in the catalog have been labeled by a human expert (astronomer)

ML Algorithm

Model

Spiral galaxy

RECOMMENDATION SYSTEMS

- Example: **Santander Product Recommendation**
  - https://www.**kaggle**.com/c/santander-product-recommendation/data
  - Prize: $60000. 1787 teams.
  - Data 1.5 years of customer behavior: products bought (saving accounts, credit card, funds, …) and demographic data (wages, age, gender, location, …)
- The goal was to predict what new products the customer would buy the last month (June 2016)

New customer:

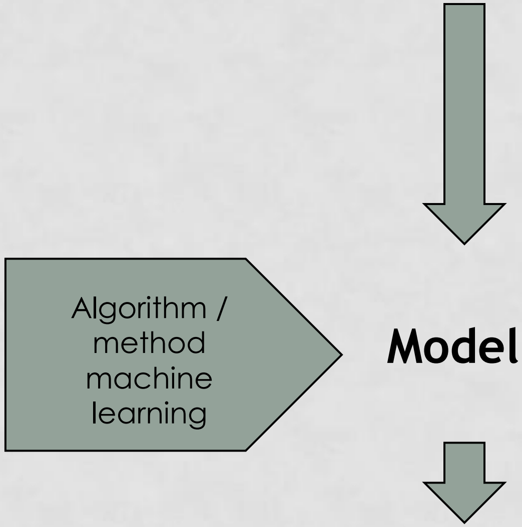
- specific data: age=50 years, gender=female, location=22500, …
- Products bought / used: credit card, savings account (up to May 2016)

?

Training data

Bank database, for every customer:

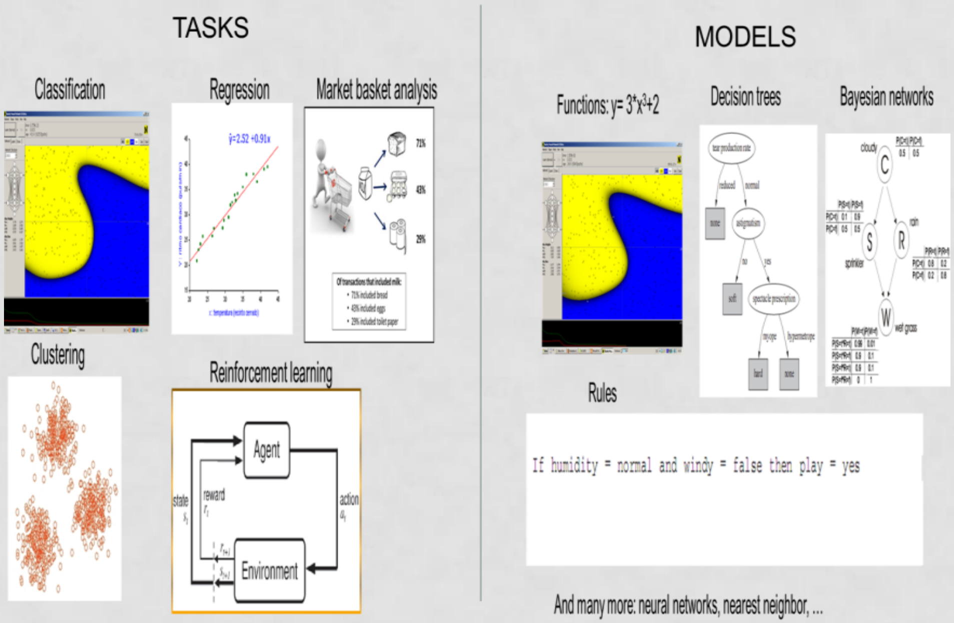
- Specific data: age, gender, location, …
- Products bought / used: credit card, funds, shares, savings account, …



She will buy Telefónica shares in June 2016

SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models
2. Basic methods for training classification and regression models:
3. Methodology
4. Methods for preprocessing (imputation, feature selection, ...)
5. Advanced training methods based on ensembles of models
6. Large Scale Machine Learning. Big Data
7. Advanced topics
8. Software tools



If humidity = normal and windy = false then play = yes

And many more: neural networks, nearest neighbor, ...

SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models

2. Basic methods for training classification and regression models:

- Nearest Neighbour (KNN)
- Classification / regression trees & rules

3. Methodology

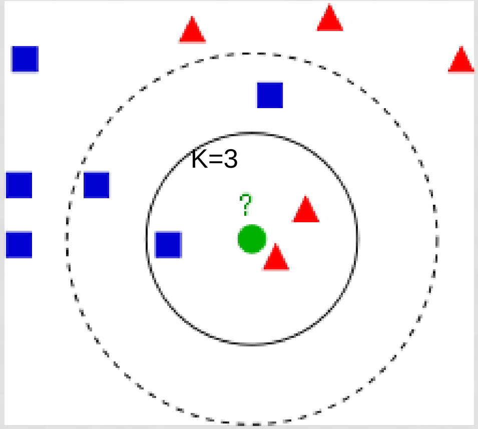
4. Methods for preprocessing

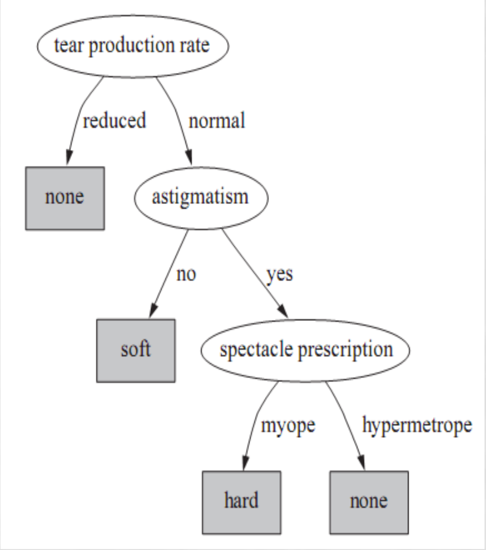
5. Advanced training methods based on ensembles of models

6. Large Scale Machine Learning. Big Data

7. Advanced topics

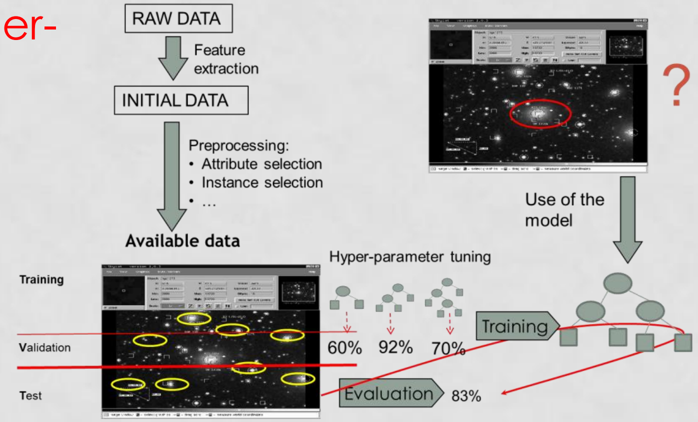
8. Software tools





SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models
2. Basic methods for training classification and regression models:
3. Methodology (the Machine Learning workflow): hyper-parameter tuning, model evaluation, preprocessing, …
4. Methods for preprocessing
5. Advanced training methods based on ensembles of models
6. Large Scale Machine Learning. Big Data
7. Advanced topics
8. Software tools



RAW DATA

Feature extraction

INITIAL DATA

Preprocessing:
- Attribute selection
- Instance selection
- …

Available data

Training

Validation

Test

Hyper-parameter tuning

60% 92% 70%

Training

Evaluation

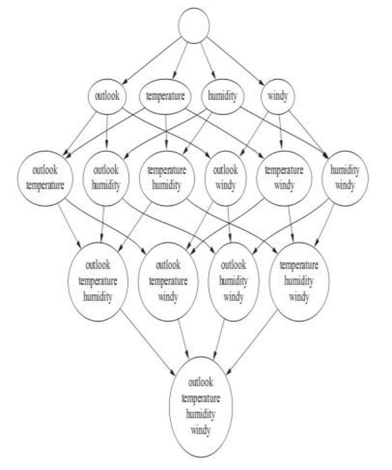
83%

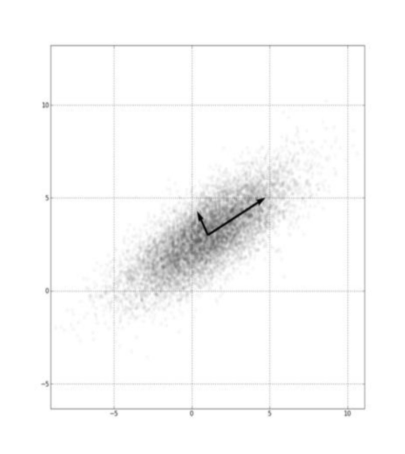
?

Use of the model

SYLLABUS

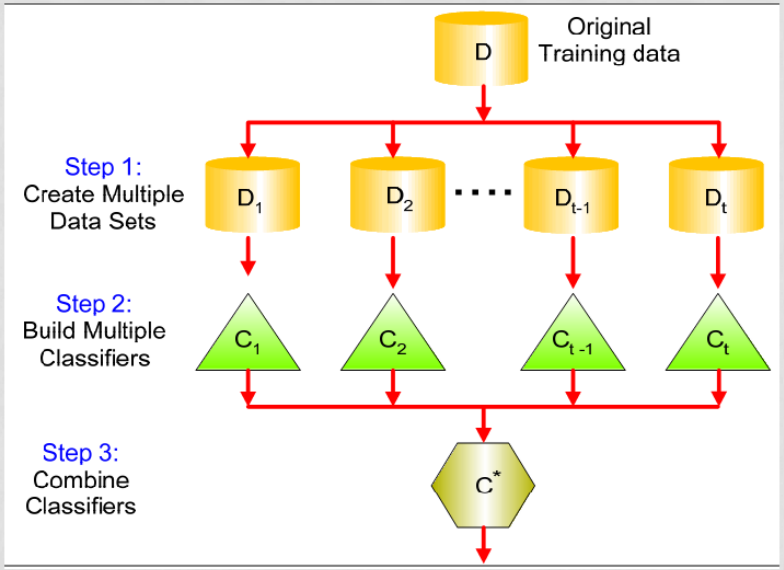
1. Introduction to Machine Learning: tasks, algorithms & models
2. Basic methods for training classification and regression models:
3. Methodology
4. Methods for preprocessing: imputation, categorical encoding, **feature selection**, ...
5. Advanced training methods based on ensembles of models
6. Large Scale Machine Learning. Big Data
7. Advanced topics
8. Software tools





SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models
2. Basic methods for training classification and regression models:
3. Methodology
4. Methods for preprocessing
5. Advanced training methods based on ensembles of models: bagging, boosting, stacking
6. Large Scale Machine Learning. Big Data
7. Advanced topics
8. Software tools


D
Original Training data
Step 1: Create Multiple Data Sets
D_1
D_2
....
D_{t-1}
D_t
Step 2: Build Multiple Classifiers
C_1
C_2
C_{t-1}
C_t
Step 3: Combine Classifiers
C^*

SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models
2. Basic methods for training classification and regression models:
3. Methodology
4. Methods for pre-processing
5. Advanced training methods based on ensembles of models
6. Large Scale Machine Learning. Big Data
   - Map-reduce & Spark (streaming)
7. Advanced topics
8. Software tools

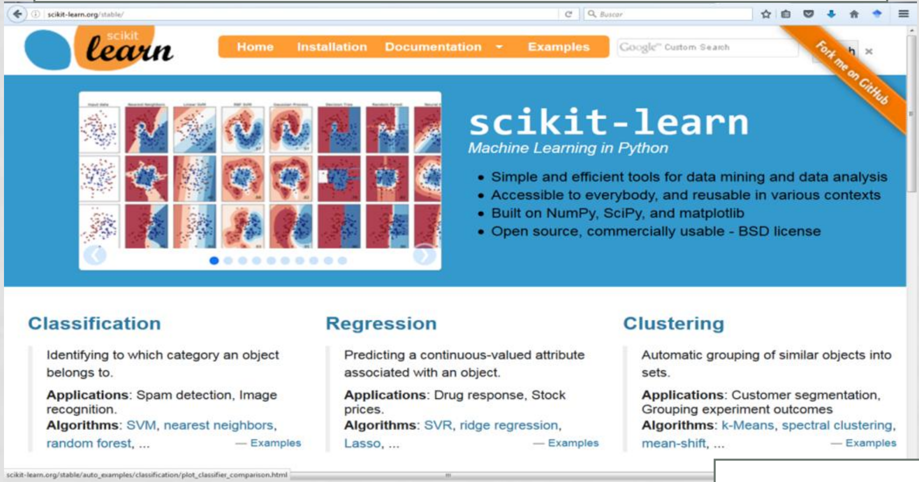
SYLLABUS

1. Introduction to Machine Learning: tasks, algorithms & models
2. Basic methods for training classification and regression models:
3. Methodology
4. Methods for pre-processing
5. Advanced training methods based on ensembles of models
6. Large Scale Machine Learning. Big Data
7. Advanced topics: imbalanced problems, probability prediction/calibration, metric learning, ...
8. Software tools

SYLLABUS:

7. SOFWARE TOOLS

FOR MACHINE LEARNING BASICS:
Python + scikit-learn



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

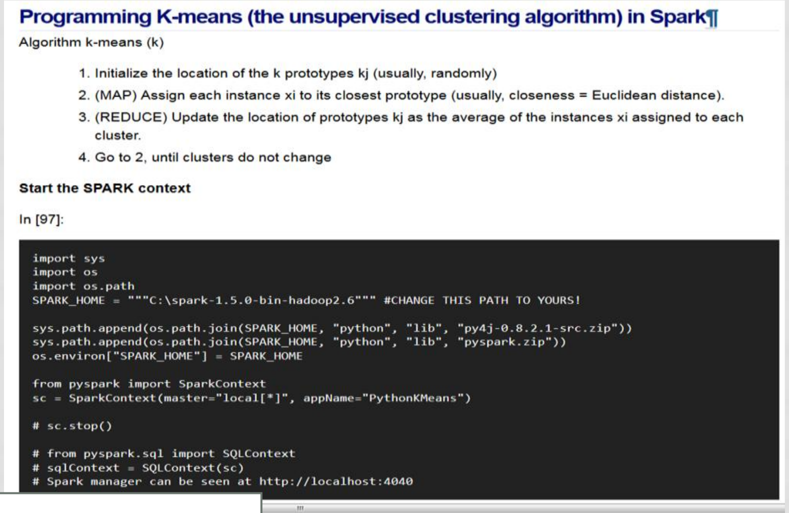
Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Pyspark + MLIB



Programming K-means (the unsupervised clustering algorithm) in Spark¶

Algorithm k-means (k)

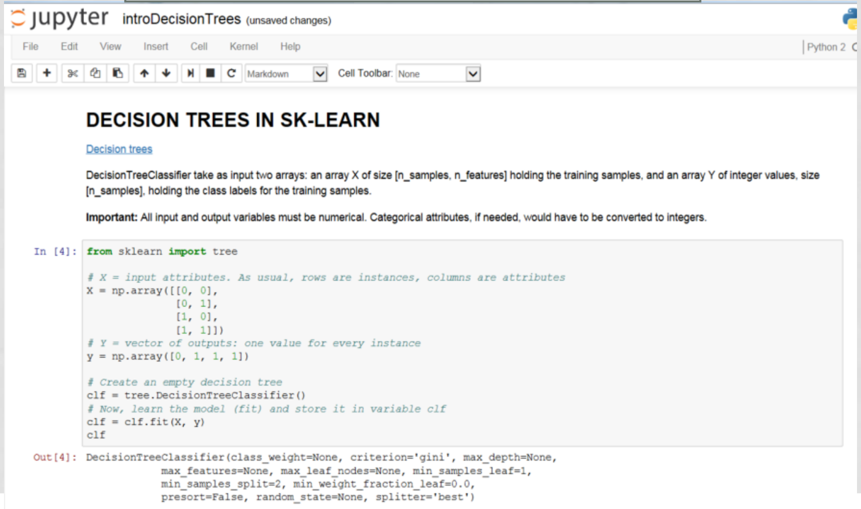
1. Initialize the location of the k prototypes kj (usually, randomly)
2. (MAP) Assign each instance xi to its closest prototype (usually, closeness = Euclidean distance).
3. (REDUCE) Update the location of prototypes kj as the average of the instances xi assigned to each cluster.
4. Go to 2, until clusters do not change

Start the SPARK context

In [97]:

```
import sys
import os
import os.path
SPARK_HOME = """C:\spark-1.5.0-bin-hadoop2.6""" #CHANGE THIS PATH TO YOURS!

sys.path.append(os.path.join(SPARK_HOME, "python", "lib", "py4j-0.8.2.1-src.zip"))
sys.path.append(os.path.join(SPARK_HOME, "python", "lib", "pyspark.zip"))
os.environ["SPARK_HOME"] = SPARK_HOME

from pyspark import SparkContext
sc = SparkContext(master="local[*]", appName="PythonKMeans")

# sc.stop()

# from pyspark.sql import SQLContext
# sqlContext = SQLContext(sc)
# Spark manager can be seen at http://localhost:4040
```

IPYTHON NOTEBOOKS



DECISION TREES IN SK-LEARN

Decision trees

DecisionTreeClassifier take as input two arrays: an array X of size [n_samples, n_features] holding the training samples, and an array Y of integer values, size [n_samples], holding the class labels for the training samples.

Important: All input and output variables must be numerical. Categorical attributes, if needed, would have to be converted to integers.

```
In [4]: from sklearn import tree

        # X = input attributes. As usual, rows are instances, columns are attributes
        X = np.array([[0, 0],
                      [0, 1],
                      [1, 0],
                      [1, 1]])
        # Y = vector of outputs: one value for every instance
        y = np.array([0, 1, 1, 1])

        # Create an empty decision tree
        clf = tree.DecisionTreeClassifier()
        # Now, learn the model (fit) and store it in variable clf
        clf = clf.fit(X, y)
        clf

Out[4]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                    max_features=None, max_leaf_nodes=None, min_samples_leaf=1,
                    min_samples_split=2, min_weight_fraction_leaf=0.0,
                    presort=False, random_state=None, splitter='best')
```

GRADING

- A = 30% FINAL EXAM
- B = 70% ASSIGNMENTS. Groups with two members
  (Scikit-learn, Pyspark / MLLIB)

- Pass if A+B>=50% (no mínimum grade in the exam)

TASKS

- What can be done? Tasks:
  - Supervised ML
    - Classification
      - Probability estimation
    - Regression
  - Unsupervised ML
    - Clustering
    - Association
  - Semi-supervised ML
  - Reinforcement learning → Time is one of the main variables involved in the model. Its very complex

TASKS

- **Supervised ML:**
  - Classification
    - Probability prediction
  - Regression
    - Quantile regression, prediction intervals
- Unsupervised ML:
  - Clustering
  - Association
- Semi-supervised ML
- Reinforcement learning

TASKS

- **Supervised learning:**
  - **Classification:**
  - Regression
- Semi-supervised learning
- Unsupervised learning:
  - Clustering
  - Association
- Reinforcement learning

STEPS:
-Training the model
-Deploying the model

CLASSIFICATION TASK. AN EXAMPLE:

- Bank credit approval:
  - An Internet bank owns a large data base with information about clients who either defaulted or not on a loan
  - The banks requires a model to determine if a new customer will repay the loan or not
  - Instances (client records in the database):
    - Input attributes : credit time-length (years), amount, overdue accounts?, own house?
    - Class: yes/no
  - Rule-based model:
    - **IF** (overdue accounts > 0) **THEN** repay loan = no
    - **IF** (overdue accounts = 0) **AND** ((salary > 2500) **OR** (years > 10)) **THEN** repay loan = yes

SUPERVISED MACHINE LEARNING CLASSIFICATION TASK. AN EXAMPLE:

future data

Years	Amount	Salary	Own house?	Overdue accounts?	Repay loan
10	50000	3000	Yes	0	??

T = training set

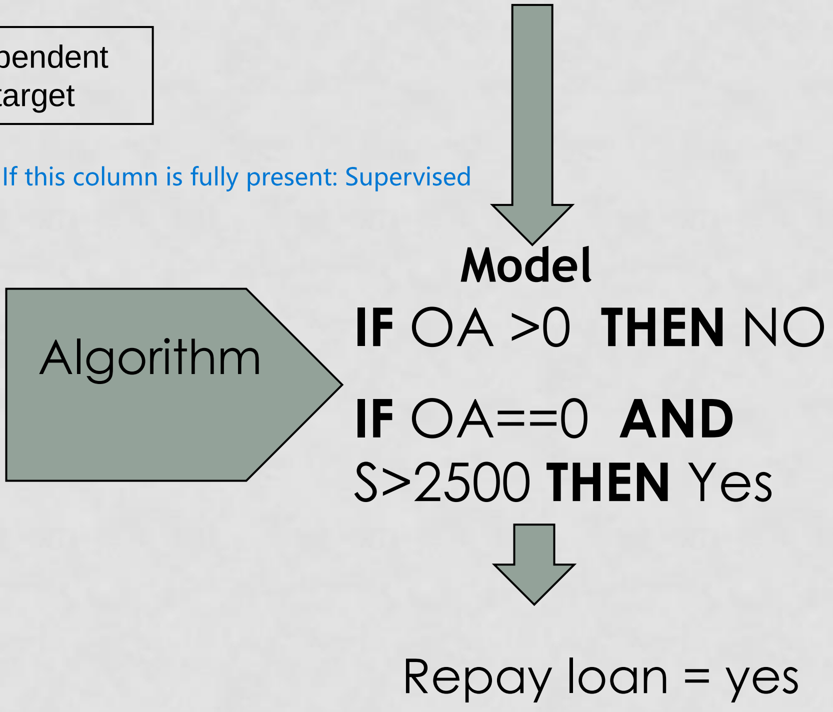
Attributes, **features**, predictors, Input variables, Independent variables, explanatory variables

Label, **class**, output variable, dependent variable, **response**, predictand, target

If this column is fully present: Supervised

Instances, examples, data points

Years	Amount	Salary	Own house?	Overdue accounts?	Repaid loan
15	60000	1900	Yes	2	No
2	30000	3500	Yes	0	Yes
9	9000	1700	Yes	1	No
15	18000	3000	No	0	Yes
10	24000	2100	No	0	No
	…	…	…	…	…



Model

IF OA >0 **THEN** NO

IF OA==0 **AND** S>2500 **THEN** Yes

Repay loan = yes

OTHER CLASSIFICATION PROBLEMS

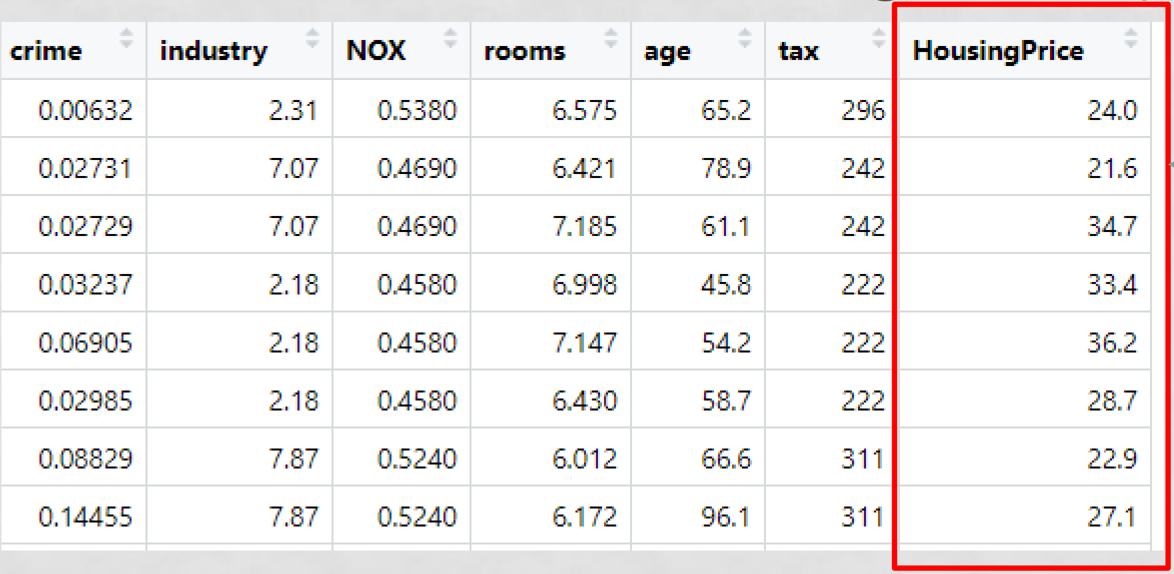
- Finances and banking
  - Credit default prediction
  - Credit card fraud detection
  - Banking products recommendation (https://www.kaggle.com/c/santander-product-recommendation)
- Insurance:
  - Expensive clients
- Education:
  - Prediction of school dropouts
- Medicine:
  - Illness diagnosis
  - Illness prediction from DNA analysis
  - Prediction if a new substance causes cancer
- Internet:
  - Spam detection

TASKS

- **Supervised learning:**
  - Classification
  - **Regression**
- Semi-supervised learning
- Unsupervised learning:
  - Clustering
  - Association
- Reinforcement learning

REGRESSION

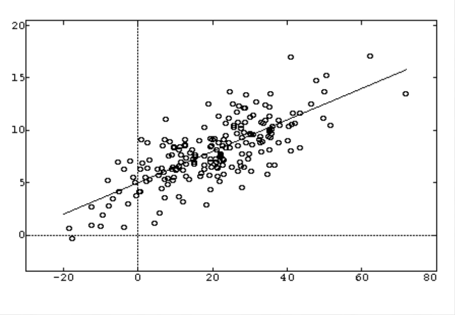
- If the class is continuous, it is a **regression** problem



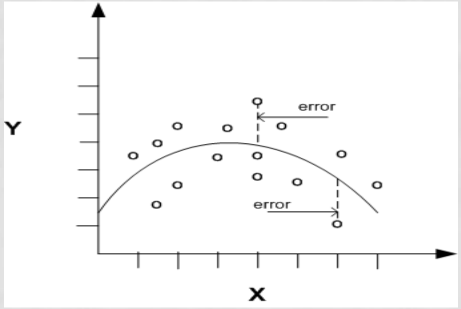
crime	industry	NOX	rooms	age	tax	HousingPrice
0.00632	2.31	0.5380	6.575	65.2	296	24.0
0.02731	7.07	0.4690	6.421	78.9	242	21.6
0.02729	7.07	0.4690	7.185	61.1	242	34.7
0.03237	2.18	0.4580	6.998	45.8	222	33.4
0.06905	2.18	0.4580	7.147	54.2	222	36.2
0.02985	2.18	0.4580	6.430	58.7	222	28.7
0.08829	7.87	0.5240	6.012	66.6	311	22.9
0.14455	7.87	0.5240	6.172	96.1	311	27.1

Response

Linear: y = ax+b



Non linear



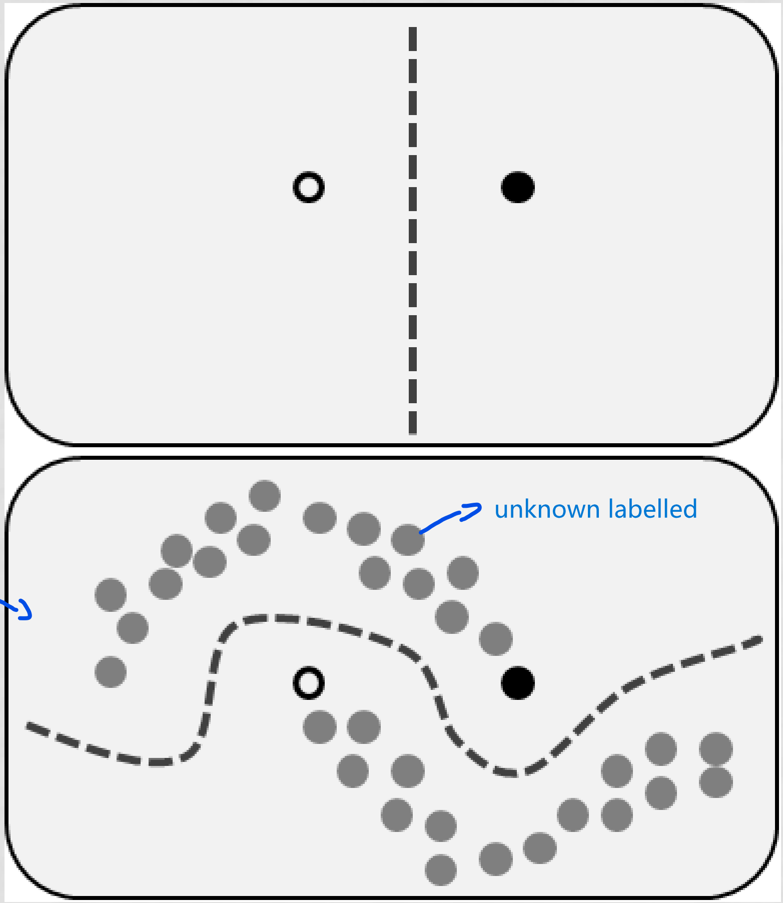
TASKS

- Supervised learning:
  - Classification
  - Regression
- **Semi-supervised learning**
- Unsupervised learning:
  - Clustering
  - Association
- Reinforcement learning

SEMISUPERVISED LEARNING

- When both labelled and unlabelled instances are available
- Why: labelling instances may be costly (ex: to perform a biopsy to determine if a person has cancer)

X1	X2	Y
-1	0	White
+1	0	Black
-2.3	0.1	?
-3	-0.1	?
+2.5	0.2	?
+2.7	-0.3	?
…	…	…



First approach: remove non labeled instances (Supervised). Problem: we reduce too much our dataset.

Second approach

TASKS

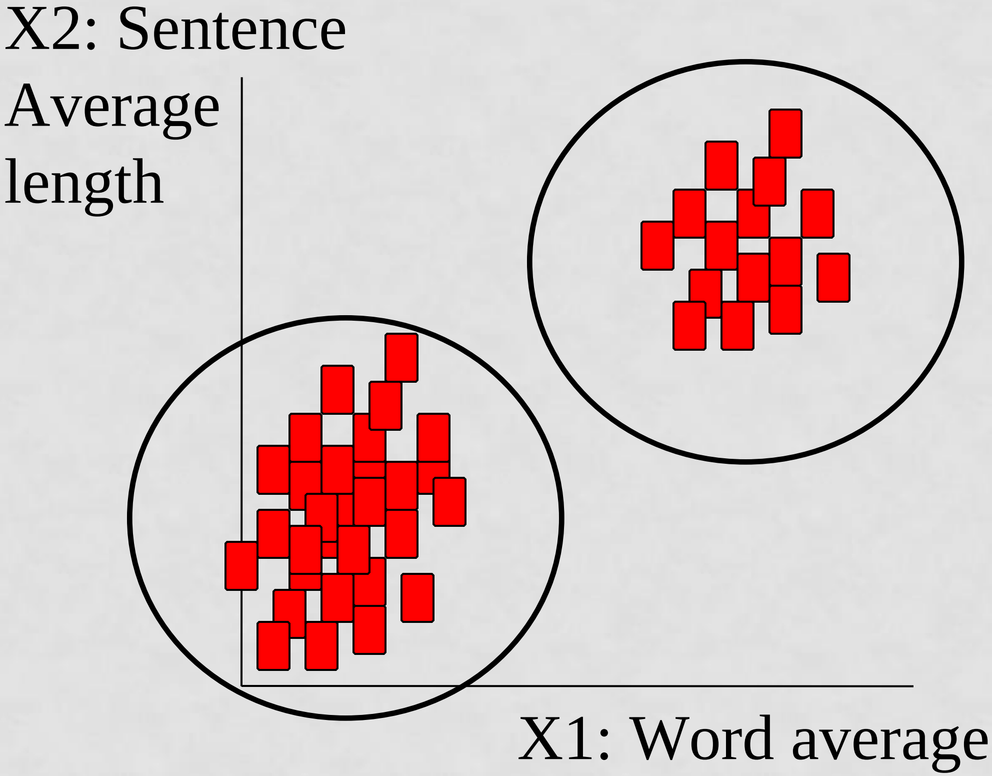
- Supervised learning:
  - Classification
  - Regression
- Semi-supervised learning
- **Unsupervised learning:**
  - **Clustering**
  - Association
- Reinforcement learning

UNSUPERVISED LEARNING (NO OUTPUT VARIABLE): CLUSTERING

- To determine natural clusterings in instance space, based on the input attributes (no labels)
- Real-world example: Market segmentation

books

WAL	SAL	
1.3	2.7	
2.5	6.7	
2.9	3.1	



Example: each data point is a different book. 2 groups:

* Long words and sentences (philosophy?)

* Short words and sentences (best-sellers?)

UNSUPERVISED LEARNING (NO LABELS): CLUSTERING

BEST PRICE
NEW
BIG OFFER
SPECIAL PRICE
SPECIAL OFFER
HOT SALE
I ♥ EAT

- Personalized publicity
- Solution: customer segmentation
  - 4 groups identified:

    Healthy, gourmets, junk food, families with children
  - Special offers, new products, …

https://medium.com/@cansuozcan/real-life-examples-of-association-analysis-clustering-analysis-text-mining-and-web-usage-mining-10eabe4a9590

TASKS

- Supervised learning:
  - Classification
  - Regression
- Semi-supervised learning
- **Unsupervised learning:**
  - Clustering
  - **Association**
- Reinforcement learning

MARKET BASKET ANALYSIS (ASSOCIATION)

- A supermarket needs to know customer behavior.
  - Ex: if customer buys X then s/he also buys Y
- Service might be improved (putting together products bought together, etc.)

TRAINING DATA (CUSTOMER BASKETS)

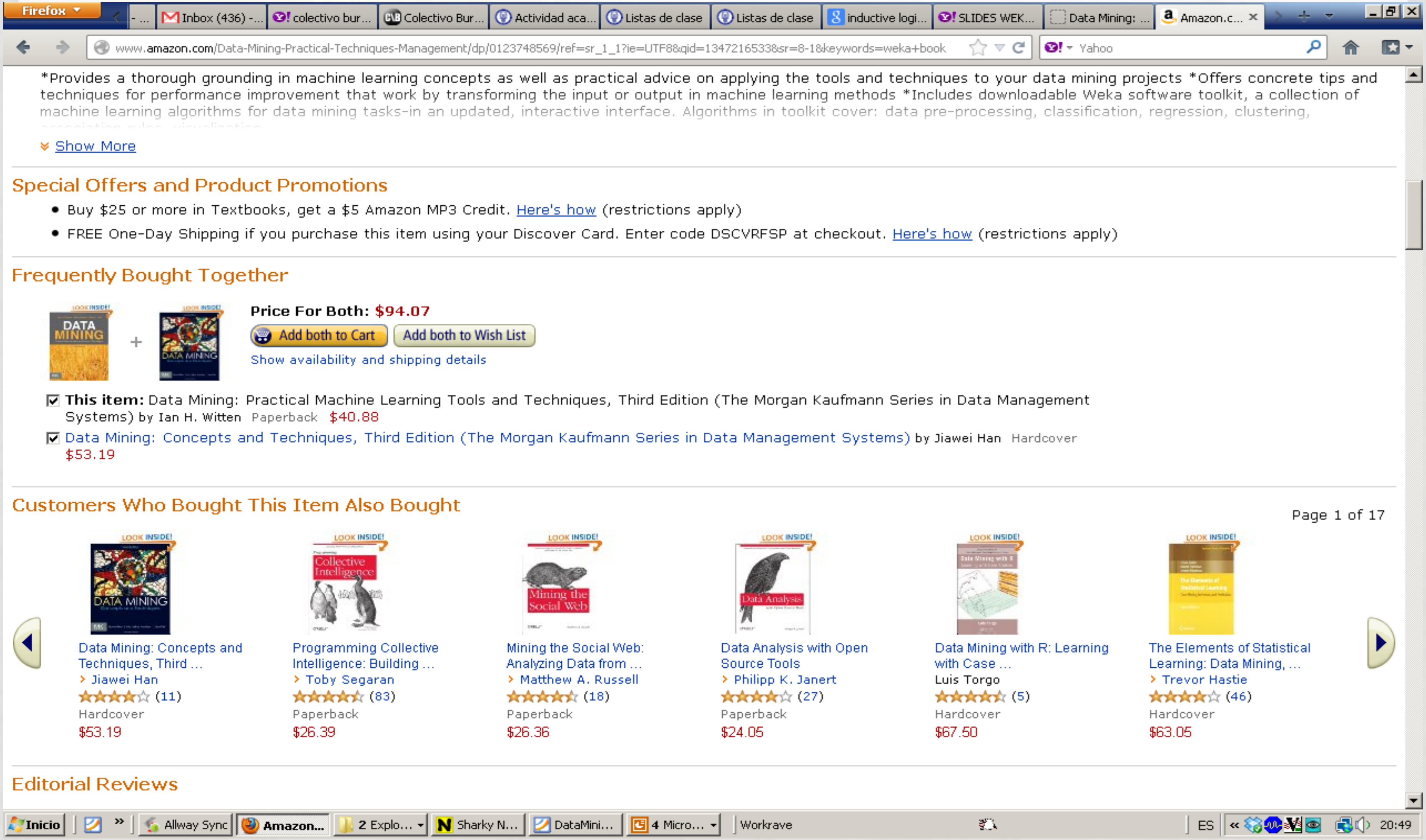
Id	Eggs	Oil	Napies	Wine	Milk	Butter	Salmon	Lettuce	...
1	Yes	No	No	Yes	No	Yes	Yes	Yes	...
2	No	Yes	No	No	Yes	No	No	Yes	...
3	No	No	Yes	No	Yes	No	No	No	...
4	No	Yes	Yes	No	Yes	No	No	No	...
5	Yes	Yes	No	No	No	Yes	No	Yes	...
6	Yes	No	No	Yes	Yes	Yes	Yes	No	...
7	No	No	No	No	No	No	No	No	...
8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	...
…	…	…	…	…	…	…	…	…	.. .

MODEL

- Rules **IF** At_1=a AND At_2=b **THEN** At_n=c, At_4=D
  - **IF** nappies=Yes **THEN** milk=Yes
  - **IF** butter = Yes **AND** salmon = Yes **THEN** wine = Yes, eggs = Yes

Service might be improved (putting together nappies and milk, etc.)

ASSOCIATION



*Provides a thorough grounding in machine learning concepts as well as practical advice on applying the tools and techniques to your data mining projects *Offers concrete tips and techniques for performance improvement that work by transforming the input or output in machine learning methods *Includes downloadable Weka software toolkit, a collection of machine learning algorithms for data mining tasks-in an updated, interactive interface. Algorithms in toolkit cover: data pre-processing, classification, regression, clustering, association rules, visualization

Show More

Special Offers and Product Promotions

- Buy $25 or more in Textbooks, get a $5 Amazon MP3 Credit. Here's how (restrictions apply)
- FREE One-Day Shipping if you purchase this item using your Discover Card. Enter code DSCVRFSP at checkout. Here's how (restrictions apply)

Frequently Bought Together

Price For Both: $94.07

Add both to Cart | Add both to Wish List

Show availability and shipping details

☑ **This item:** Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) by Ian H. Witten Paperback $40.88

☑ Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) by Jiawei Han Hardcover $53.19

Customers Who Bought This Item Also Bought

Page 1 of 17

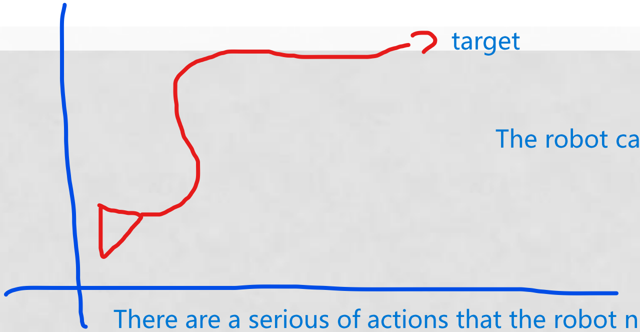
Data Mining: Concepts and Techniques, Third ...	Programming Collective Intelligence: Building ...	Mining the Social Web: Analyzing Data from ...	Data Analysis with Open Source Tools	Data Mining with R: Learning with Case ...	The Elements of Statistical Learning: Data Mining, ...
Jiawei Han	Toby Segaran	Matthew A. Russell	Philipp K. Janert	Luis Torgo	Trevor Hastie
★★★★☆ (11)	★★★★☆ (83)	★★★★☆ (18)	★★★★☆ (27)	★★★★★ (5)	★★★★☆ (46)
Hardcover	Paperback	Paperback	Paperback	Hardcover	Hardcover
$53.19	$26.39	$26.36	$24.05	$67.50	$63.05

Editorial Reviews

TASKS

- Supervised learning:
  - Classification
  - Regression
- Semi-supervised learning
- Unsupervised learning:
  - Clustering
  - Association
- **Reinforcement learning**

TASK: REINFORCEMENT LEARNING

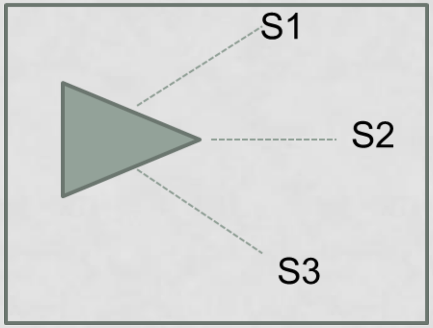


The robot cant fail any decision

There are a serious of actions that the robot need to perform in time

- Robotics, videogames, ...
- The goal of learning is a policy π so that the agent (robot) knows what to do at each situation.
- Actions:
  - forward
  - turn left
  - turn right

Distance obstacle sensors: S1, S2, S3



Π(S1, S2, S3) = action?

TASK: REINFORCEMENT LEARNING

- In principle, it is difficult to formulate it as a supervised problem, because it would be time consuming to create the training table:

output

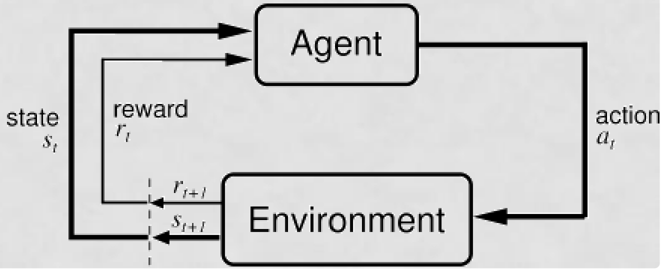
S1	S2	S3	π
1.3	0.5	7	? At the beginning we dont know the answer
10	8.7	5	?
0.5	0.5	0.6	?
…	…	…	…

S1, S2, S3 the distance that the robot is from the sensor on each step.

We capture data through simulations

We provide positive or negative feedback to the robot

- The policy Π is learned by allowing the agent to explore a simmulated world, receiving from time to time, positive and negative rewards.


Agent
state s_t
reward r_t
action a_t
r_{t+1}
s_{t+1}
Environment