Machine Learning for Big Data: Introduction

Lionel Fillatre

fillatre@unice.fr

Topics

- Examination
- Data analysis
- Machine learning
- Programming tools
- Conclusion

1 Examination

Examination

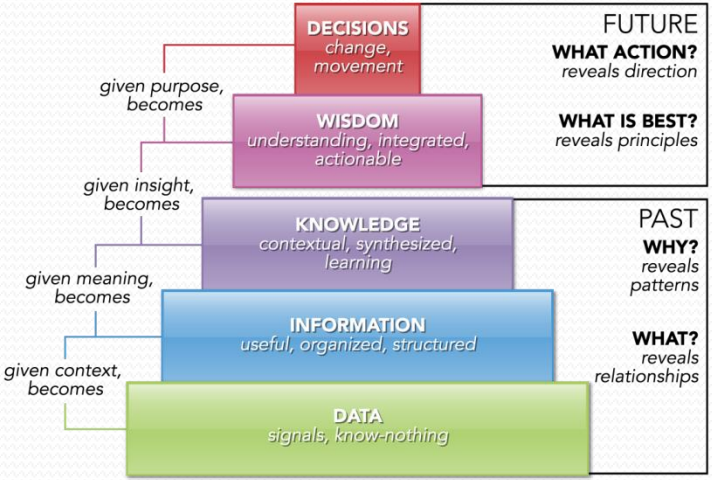
- Written midterm examination: concentrated on the same topics as the lectures and exercises and linked to the learning objectives
- Written final examination: concentrated on the same topics as the lectures and exercises and linked to the learning objectives
- Group reports : 2 reports on assignments with Python
- Final grade based on an overall assessment of the reports (2 grades, 20% each report) and written exams (2 grades, 30% each exam).

2 Data Analysis

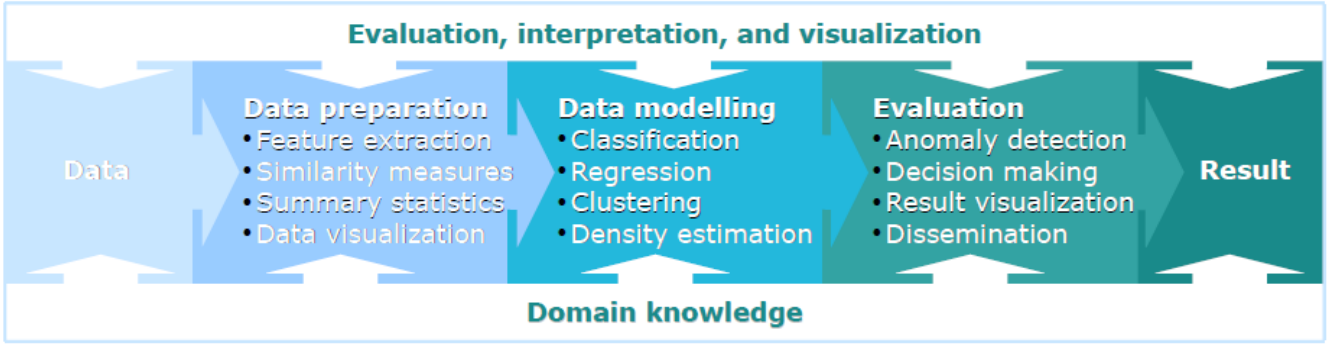
Basic Definitions

- Data (datum)
  - A single piece of information, as a fact, statistic, or code; an item of data.
  - When data are processed, organized, structured or presented in a given context so as to make them useful, they are called Information.
- Information
  - It consists of facts and data organized to describe a particular situation or condition
- Knowledge
  - It consists of facts, truths, and beliefs, perspectives and concepts, judgments and expectations, methodologies and know-how.
  - Knowledge is accumulated and integrated and held over time to handle specific situations and challenges.

Data Driven Decisions



Data modeling framework



Deluge of information

Every day, we create 2.5 quintillion (10^{18}) bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

Source: http://www-01.ibm.com/software/data/bigdata/

"If data had mass, the earth would be a black hole"

Stephen Marsland

"We are drowning in information and starving for knowledge"

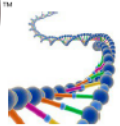
John Naisbitt

Big Data Era

- ~1 trillion webpages
(http://googleblog.blogspot.dk/2008/07/we-knew-web-was-big.html)

- One hour of video is uploaded to youtube every second resulting in 10 years of content every day
(source: youtube)

- We have sequenced more than 1000 peoples genome of $3.8 \cdot 10^9$ base pairs
(source: K. P. Murphy "Machine Learning")



- Walmart handles more than 1 mio. transactions per hour and has databases containing more than $2.5 \cdot 10^{15}$ bytes of information
(source: K. P. Murphy "Machine Learning")

- Each night the worlds astronomy laboratories store high-resolution of the night sky of around a terabyte (10^{12})
(source: Stephen Marsland "Machine Learning An Algorithmic Perspective")

- In total, the four main detectors at the Large Hadron Collider (LHC) produced 13 petabytes (10^{15}) of data in 2010
(source: wikipedia "Big Data")

- Facebook handles 40 billion photos from its user base.
(source: wikipedia "Big Data")

- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide
(source: wikipedia "Big Data")

Motivating Challenges

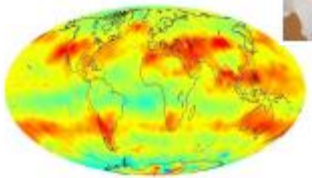
- **Scalability**
  - Datasets with sizes of gigabytes, terabytes and petabytes are becoming common
- **High Dimensionality**
  - It is now common to encounter data sets with hundreds or thousands of attributes
- **Heterogenous and Complex Data**
  - Attributes commonly of different types
  - Multiple types of datasets
- **Data Ownership and Distribution**
  - Data distributed across many locations/organizations
  - Security Issues, privacy preserving issues.
- **Non-traditional Analysis**
  - Traditional statistical approach: Hypothesize-and-test paradigm
  - Current data analysis tasks often require the generation and evaluation of thousands of hypotheses.
  - Data mining can automate this process to identify interesting hypotheses for formal testing.

Applications

- **Chemistry**
  - Spectrometry, Chemical sensors
- **Audio processing**
  - Spoken digit classification, Music gen classification
- **Image processing**
  - Hand-written digit recognition, Image tagging and classification, Number plate recognition
- **Informatics**
  - Collaborative filtering, Text corpus Spam filters, Computer games
- **Biomedical**
  - Micro-array gene analysis, Medical Imaging
- **Financial data mining**
  - Market predictions
- **Climate data**
  - Weather forecast

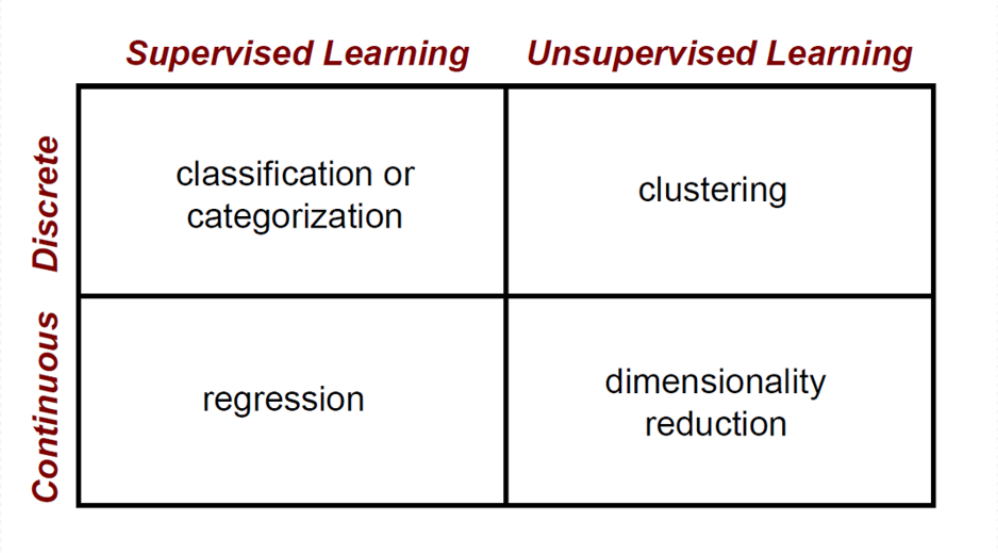
NETFLIX

amazon.com



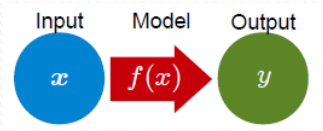
3 Machine learning

Machine Learning Problems

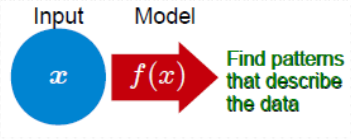


Supervised vs. Unsupervised

- Predictive tasks (Supervised learning)
  - Use some variables to predict unknown or future values of other variables
    - Classification (discrete output) : determine which class a new data object belongs to
    - Regression (continuous output) : determine the output value from the input variables

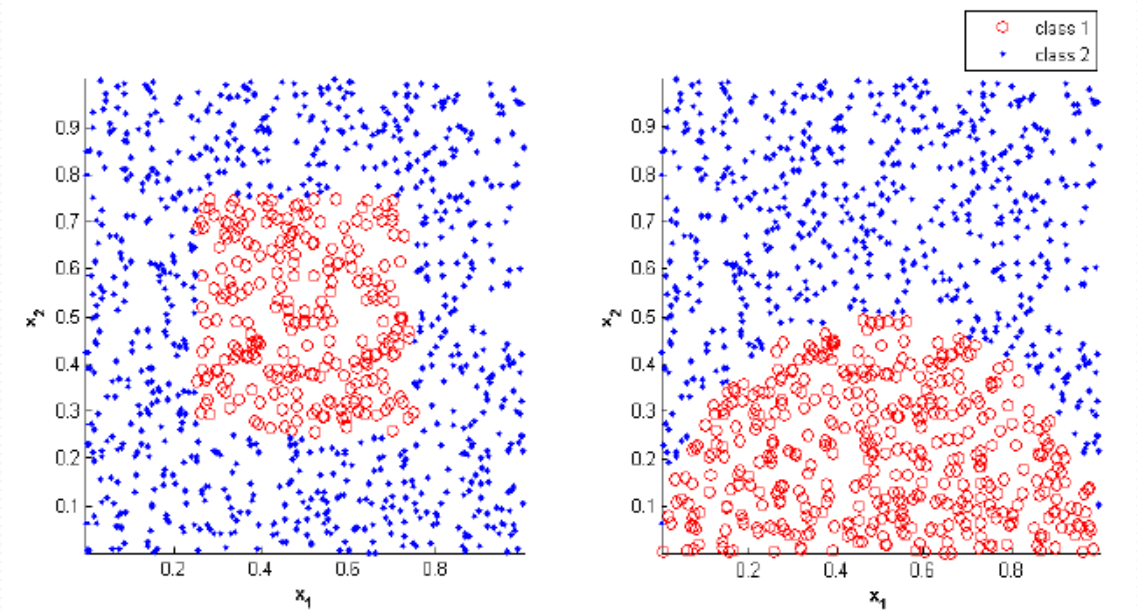


- Descriptive tasks (Unsupervised learning)
  - Find human-interpretable patterns that describe the data
    - Clustering : discover group structure in data
    - Association rule discovery : discover how data objects relate to each other
    - Anomaly detection : find data objects that are abnormal

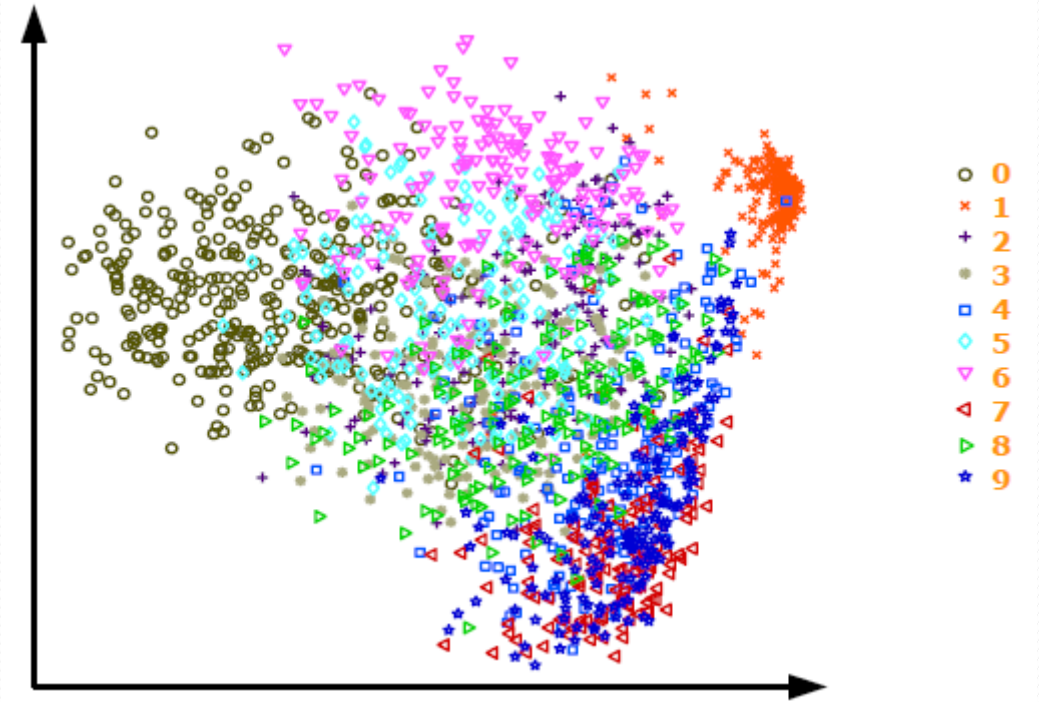


Binary classification (detection)

Goal: Assign a class label to a previously unseen object



Multiclass Classification



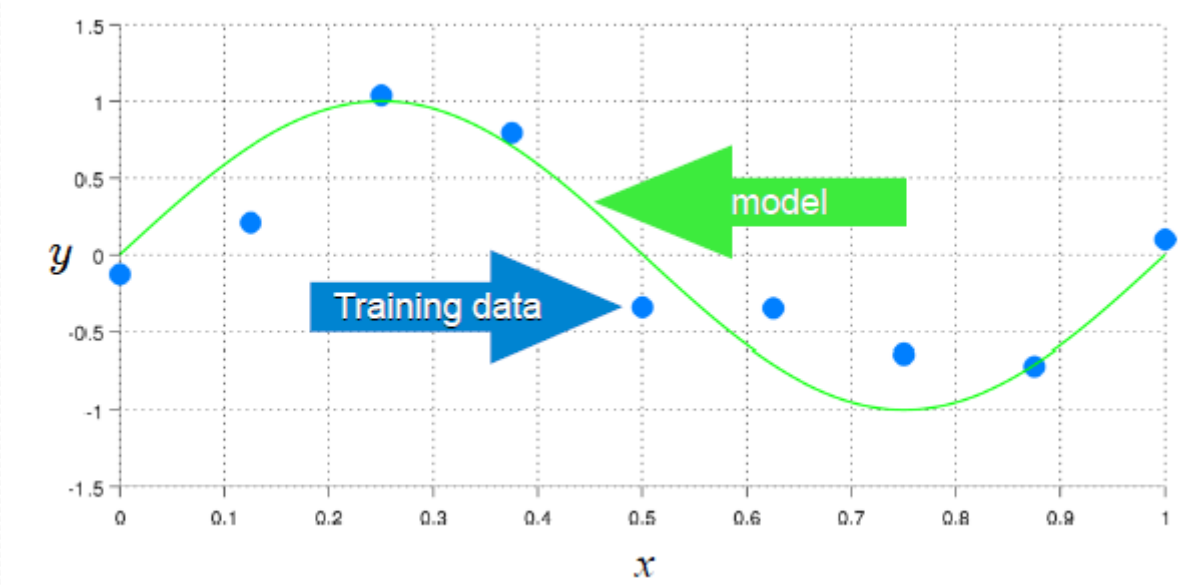
Clustering

Goal: Group the objects into clusters such that

- Objects within each cluster are similar
- Objects in separate clusters are less similar

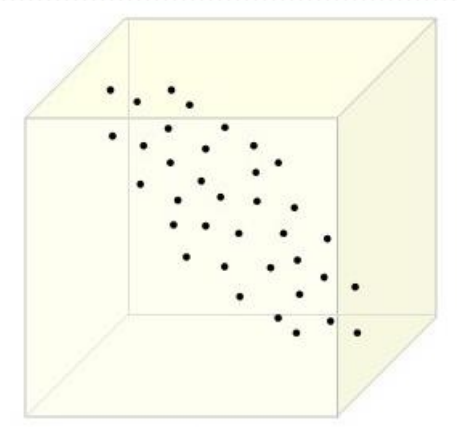
Regression

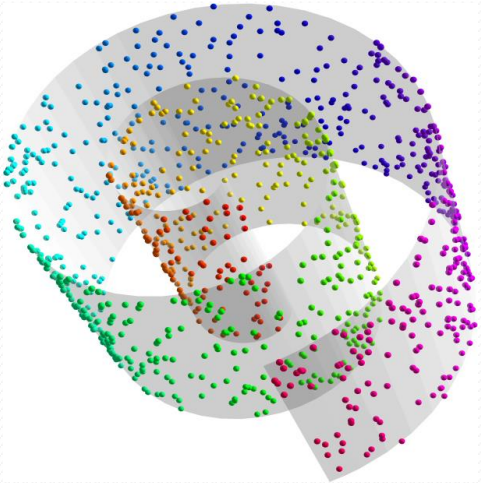
Goal: Predict the value of the variable for a previously unseen object.



Dimension Reduction

- Goal: Produce a compact low-dimensional encoding of a given high-dimensional data set.





4 Programming tools

Tools for Big Data and Data Science

- MLLIB: MLlib is Apache Spark's scalable machine learning library.
  - logistic regression, linear support vector machine (SVM), classification, random forest, clustering via k-means, singular value decomposition (SVD), principal component analysis (PCA), linear regression with L1, L2, hypothesis testing

- MAHOUT: Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms
  - Collaborative Filtering, Matrix Factorization, Classification, Logistic Regression, Naive Bayes, Random Forest, Hidden Markov Models, Multilayer Perceptron, Clustering, k-Means Clustering, Spectral Clustering, Dimensionality Reduction, Singular Value Decomposition, PCA

- And many others: Rhadoop, H2O, Scikit-learn, Theano, Weka, LibSVM, etc.

Useful programming languages

- SQL (1970): querying and namaging data
- Python (1991): data processing, productivity, good learning curve
- R (1995): data analysis, oriented toward statistical analysis, more difficult to learn, free alternative to SAS, huge community
- And others: Java, Scala, SAS, Matlab

Python for Practical Works

- We will use Anaconda (a Python distribution) with Spyder (an interactive Python development environment).
- We will also use Jupyter notebook. This is a web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text.
- Please visit the page https://www.continuum.io/downloads

Python: important libraries

- NumPy is an extension to the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays.

- SciPy is an open source Python library used for scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, signal and image processing, and other tasks common in science and engineering.

- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Qt, or GTK+.

- Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.

Python Tutorials

- Python 3:

https://docs.python.org/3/tutorial/
http://deusyss.developpez.com/tutoriels/Python/python-en-bref/

- Jupyter Notebook:

http://dichotomies.fr/2015/informatique/info1/cours/debuter-avec-les-notebooks/

- Numpy library:

http://www.python-course.eu/numpy_numerical_operations_on_numpy_arrays.php

- Scipy library:

http://docs.scipy.org/doc/scipy/reference/tutorial/

- Matplotlib/Pyplot library:

http://matplotlib.org/users/pyplot_tutorial.html

5 Conclusion

Conclusion

- How to manage and process data is important to produce knowledge.
- Numerous tools for Big Data Analytics and Data Science
- Tuning and reliability are often crucial

  => the basics of algorithms should be known