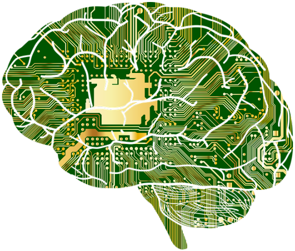
Introduction to AI: Data analysis and machine learning



Rodrigo Cabral, Lionel Fillatre and Michel Riveill

EUR DS4H-LIFE-SPECTRUM

cabral@unice.fr

Source: https://commons.wikimedia.org/wiki/File:Deep_learning.png. Some of these slides are based on lecture 1 from Machine Learning for Big Data (Polytech Nice-Lionel Fillatre)

Syllabus

MINOR

Introduction to machine learning

Coordinator

DIGITAL SYSTEMS FOR HUMANS

UNIVERSITÉ CÔTE D'AZUR

Syllabus

- Python → learn by ourself:
  - https://data-flair.training/blogs/python-machine-learning-tutorial/
- Rodrigo Cabral
  1. General introduction
     - The different problems of ML
     - The learning process

     ⇦ Today
  2. Regression with the linear model
  3. Classification - Régression logistique
  4. Test at the beginning of 4th class, 40% of total grade
     SVM
- Lionel Fillatre
  5. LDA / Naive Bayes
  6. CART / Decision Tree / Random Forest
- Michel Riveill
  7. Clustering (k-means, hclust)
  8. Test at the beginning of 8th class, 40% of total grade
  9. M - Dimension reduction (PCA, t-SNE)

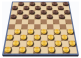
Outline

Machine learning

- General definition from Tom Mitchell (Carnegie Mellon 1997)

  “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

- Examples given by Tom Mitchell

	Checkers learning
Tasks T	Playing checkers
Performance measure P	Percent of games won against opponents
Training experience E	Playing practice games against itself

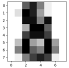
Source: https://commons.wikimedia.org/wiki/File:International_draughts.jpg

Machine learning

- General definition from Tom Mitchell (Carnegie Mellon 1997)

  "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

- Examples given by Tom Mitchell



	Handwritten recognition
Tasks T	Recognizing and classifying handwritten words within images
Performance measure P	Percent of words correctly classified
Training experience E	A database of handwritten words with given classifications

Source: http://scikit-learn.org/stable/auto_examples/datasets/plot_digits_last_image.html

Machine learning

- General definition from Tom Mitchell (Carnegie Mellon 1997)

  “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

- Examples given by Tom Mitchell

	Robot driving learning
Tasks T	Driving on public four-lane highways using vision sensors
Performance measure P	Average distance traveled before an error
Training experience E	A sequence of images and steering commands recorded while observing a human driver

Source: https://commons.wikimedia.org/wiki/File:Waymo_Chrysler_Pacifica_in_Los_Altos,_2017.jpg

Motivation

Progress at three levels :

data gathering

data storage

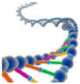
data processing

$\Downarrow$

Big data: E is now easy to get!

Big data

- ~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")

1. What is machine learning?
2. Machine learning applications
3. Materials: the data
4. Methods: algorithms and pipeline
5. Tools: computer software
6. Bibliography
7. Evaluation of the first part
8. Conclusions

Applications

- Audio processing
  spoken word classification and automatic translation, music generation and classification
- Image processing
  handwritten word classification, image tagging and classification, image forensics, object segmentation and classification 
- Text processing
  text classification, natural language processing, spam filtering
- Chemometrics
  molecule identification and quantification 

Sources: http://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html, https://commons.wikimedia.org/wiki/File:LTQ_OrbiTrap_XL-Thermo_Scientific_1.jpg

Applications

- Biomedical

  microarray gene analysis, medical imaging

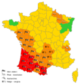
- Recommender systems

  collaborative filtering, information retrieval

NETFLIX

- Climate data

  weather forecast



- Defense and security

  target detection and classification

- Transportation

  autonomous vehicles

Sources: https://commons.wikimedia.org/wiki/File:Logo_Netflix.png,
https://commons.wikimedia.org/wiki/File:Carte_vigilance_Météo_France_24-01-2009_08h10.svg,
https://commons.wikimedia.org/wiki/File:Waymo_Chrysler_Pacifica_in_Los_Altos,_2017.jpg

Applications

- Transportation

  autonomous vehicles $\Longrightarrow$self-driving cars
- Multiple tasks of machine learning
- Convolutional neural network for detection, segmentation and classification:

  https://www.youtube.com/watch?v=OOT3UIXZztE
- Huge impact on society

Applications

- Text processing

  spam filtering

- Computer program = your mailing service

  1. Classed emails = data: analyze your already classed mails
  2. Learning or fitting: specify rules linking probability of a spam with frequency of words and email origin
  3. Prediction: apply rules to classify each new email you receive

Examples: “SAVE", ‘money" $\Longrightarrow$ higher prob. spam
“meeting", “problem" $\Longrightarrow$ lower prob. spam

Data

- *Datum*

  a characteristic or a number that may contain information about an objects, individuals, observations, populations

  e.g. Age [years] = 31

- *Data*

  multiple *datum* about one or multiple objects, individuals, *etc*

Without data $\Longrightarrow$ Without **E** $\Longrightarrow$ No machine learning!

Data

- *Datum*
  a characteristic or a number that may contain information about an objects, individuals, observations, populations

  $$\textit{e.g.}\ \text{Age [years]} \quad = \quad 31$$

- *Data*
  multiple *datum* about one or multiple objects, individuals, *etc*

  Without data $\Longrightarrow$ Without **E** $\Longrightarrow$ No machine learning!

- Multiple individuals or observations for the same quantity $\Longrightarrow$ variable, feature or attribute x
- Observed x for M individuals $x_1, \cdots, x_M \Longrightarrow$ feature vector $\mathbf{x}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}, \qquad \textit{e.g.}\ \text{Age in years of } M = 3 \text{ individuals } \mathbf{x}_{\mathrm{A}} = \begin{bmatrix} 31 \\ 23 \\ 32 \end{bmatrix}$$

Data - feature matrix

- Most cases we have multiple individuals **and** multiple variables
  $\Longrightarrow$ N feature vectors $\mathbf{x}_1, \cdots, \mathbf{x}_N \Longrightarrow$ feature matrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,N} \\ \vdots & \ddots & \\ x_{M,1} & & x_{M,N} \end{bmatrix}$$

- Columns are feature vectors
  Rows are observation vectors $\Longrightarrow$ $\mathbf{X}$ is $M \times N$ matrix

  e.g. Age in years $\mathbf{x}_\text{A}$ and weight $\mathbf{x}_\text{W}$ in kilos of 3 individuals

$$\mathbf{X} = [\mathbf{x}_\text{A}, \mathbf{x}_\text{W}] = \begin{bmatrix} 31 & 68 \\ 23 & 64 \\ 32 & 58 \end{bmatrix}$$

Data - inputs/outputs

- In some cases, a feature vector $\mathbf{y}$ is supposed to depend on the other feature vectors (independent variables)

  $\Longrightarrow$ $\mathbf{y}$ is called the output

  The data is the tuple $(\mathbf{X}, \mathbf{y})$

- In machine learning, we assume that there is a unknown function $f(\cdot)$ linking the independent part of the observation vector $\mathbf{x}_i$ to y_i

$$y_i = f(x_i)$$

Data - inputs/outputs

- In some cases, a feature vector $\mathbf{y}$ is supposed to depend on the other feature vectors (independent variables)

  $\Longrightarrow$ $\mathbf{y}$ is called the output

  The data is the tuple $(\mathbf{X}, \mathbf{y})$

- In machine learning, we assume that there is a unknown function $f(\cdot)$ linking the independent part of the observation vector $\mathbf{x}_i$ to y_i

$$y_i = f(x_i)$$

- **One of the objectives of machine learning algorithms:** obtain an approximation $\hat{f}(\cdot)$ of $f(\cdot)$ from the data $(\mathbf{X}, \mathbf{y})$ such that we can obtain a reasonable prediction $\hat{y} = \hat{f}(\mathbf{x})$ for an observation $\mathbf{x}$ which is not in data.

Data - inputs/outputs and machine learning objective

- **One of the objectives of machine learning algorithms:** obtain an approximation $\hat{f}(\cdot)$ of $f(\cdot)$ from the data $(\mathbf{X}, \mathbf{y})$ such that we can obtain a reasonable prediction $\hat{y} = \hat{f}(\mathbf{x})$ for an observation $\mathbf{x}$ which is not in data.

e.g. $\mathbf{X} = [\mathbf{x}_A, \mathbf{x}_W]$ and $\mathbf{y} = \begin{bmatrix} 178 \\ 173 \\ 158 \end{bmatrix}$ is height in centimeters

$\Longrightarrow$ Predict height from age and weight.

Data - types

- **One of the objectives of machine learning algorithms:** obtain an approximation $\hat{f}(\cdot)$ of $f(\cdot)$ from the data $(\mathbf{X}, \mathbf{y})$ such that we can obtain a reasonable prediction $\hat{y} = \hat{f}(\mathbf{x})$ for an observation $\mathbf{x}$ which is not in data.

e.g. $\mathbf{X} = [\mathbf{x}_A, \mathbf{x}_W]$ and $\mathbf{y} = \begin{bmatrix} \text{no} \\ \text{no} \\ \text{yes} \end{bmatrix}$ says if the person is diabetic or not

$\Longrightarrow$ Predict if a person is diabetic from age and weight.

Data - Types: quantitative *vs.* qualitative I

Quantitative

- measurable quantities - numerical
- mathematical functions can be applied (*e.g.* sum, mean)
- comparisons are possible (*e.g.* $=, \neq, >, <$)

Qualitative

- characteristics or qualities (which type/category?)
- mathematical functions cannot be applied
- not all comparisons are possible

Data - Types: quantitative *vs.* qualitative II

Quantitative	Continuous	any value in an interval *e.g.* height
	Discrete	only a finite number of values *e.g.* number of rooms in a house
Qualitative	Nominal	no possible ordering *e.g.* diabetic? Yes/No
	Ordinal	ordering is possible *e.g.* product quality? Bad/Good

Data - Types: structred *vs.* non-structured

Structured	▸ column/row structured data ▸ easier too retrieve (SQL databases) ▸ *e.g.* company databases
Non structured	▸ image, text, video ▸ harder too retrieve (NOSQL databases) ▸ *e.g.* emails, documents
Semi structured	▸ XML, JSON, CSV, logs ▸ easier too retrieve (NOSQL databases) ▸ *e.g.* Twitter API data, Google API data

Algorithms - Types: supervised *vs.* unsupervised

- **One of the objectives of machine learning algorithms:** obtain an approximation $\hat{f}(\cdot)$ of $f(\cdot)$ from the data $(\mathbf{X}, \mathbf{y})$ such that we can obtain a reasonable prediction $\hat{y} = \hat{f}(\mathbf{x})$ for an observation $\mathbf{x}$ which is not in data.

Supervisor/student analogy

- $\mathbf{X}$ = a set of multiple instances of a problem to be solved by the student
- $\mathbf{y}$ = corresponding solutions given by a supervisor of the learning process
- The student “learns" (find $\hat{f}(\cdot)$) using the solutions given by the supervisor

$\Longrightarrow$ **Supervised learning**

Algorithms - Types: supervised *vs.* unsupervised

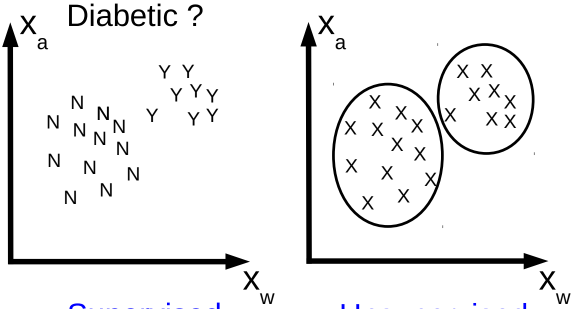
- **One of the objectives of machine learning algorithms:** from the data $(\mathbf{X}, \cdot)$ retrieve some structure underlying it: grouping of observations, retrieving a lower dimensional representation of the features (simplification of data)

Supervisor/student analogy

- $\mathbf{X}$ = a set of multiple instances of a problem to be solved by the student
- $\mathbf{y}$ is not given
- the students try to group or simplify the instances of the problem it observes under some predefined criterion

$\Longrightarrow$**Unsupervised learning**

Algorithms - Types: supervised *vs.* unsupervised



Supervised Unsupervised

Some remarks

- Output **y** is human expert data: doctor, scientist, translator, user, *etc.*
  $\Longrightarrow$ It often has a cost.

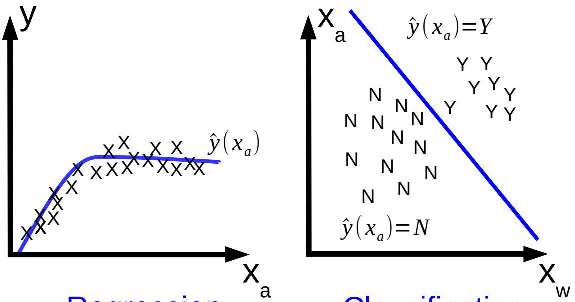
  You can even earn money from it, *e.g.* Amazon Mechanical Turk

- Unsupervised learning may be useful before supervised learning
  $\Longrightarrow$Can $\hat{\mathbf{y}}$ be guessed only from grouping?
  $\Longrightarrow$Can we use less features?

Algorithms - Types: regression *vs.* classification

Supervised algorithms types depending on the nature of $\hat{Y}$ (or Y)

Regression	**Y** can take any value in a continuous interval
Classification	**Y** can take only a finite number of predefined values (quantitative) or labels/classes (qualitative)

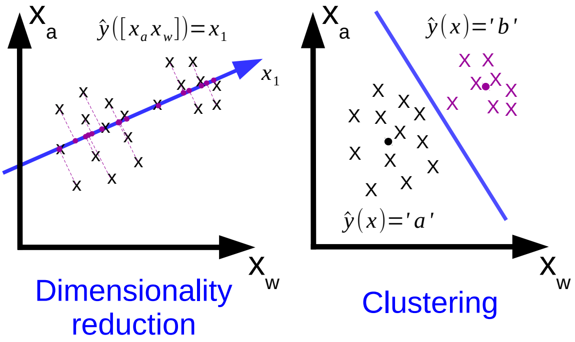


Regression   Classification

Algorithms - Types: dim. reduction *vs.* clustering

Unsupervised algorithms types depending on the nature of $\hat{\mathbf{Y}}$

Dimensionality reduction	▸ $\hat{\mathbf{Y}}$ can take any value in a continuous interval
Clustering	▸ $\hat{\mathbf{Y}}$ can take only a finite number of predefined values (quantitative) or labels/classes (qualitative)



Algorithms - Taxonomy and examples

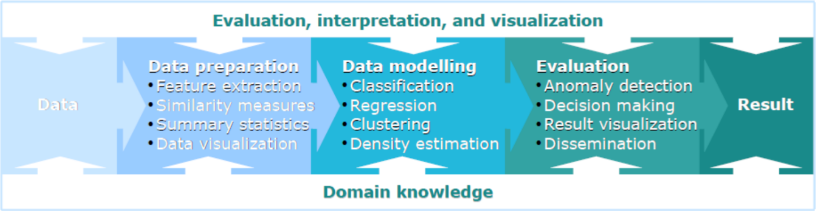
		Nature of $\hat{\mathbf{Y}}$: Continuous	Nature of $\hat{\mathbf{Y}}$: Finite
Presence of $\mathbf{Y}$	Yes	**Regression** ▸ Linear regression ▸ Neural networks ▸ Regression trees and forests ▸ Support vector regression (SVR)	**Classification** ▸ Logistic regression ▸ Naive Bayes ▸ Decision trees ▸ Random forests ▸ Support vector classification (SVC)
	No	**Dimensionality reduction** ▸ Principal component analysis (PCA) ▸ Multidimensional scaling (MDS) ▸ Kernel PCA	**Clustering** ▸ k-Means ▸ Hierarchical clustering ▸ Self-organizing maps (SOM)

Algorithms - Data modeling pipeline

Data may be easy to get, but

- we do not know which features and outputs to use
- we do not know exactly which $f(\cdot)$ to use
- features need to be extracted from it
- some parts of it may be missing
- abnormal/fake data - outliers

Data modeling pipeline



Software platforms and languages

Specialized software platforms - visual interfaces, easy to use but specialized

- Weka, Orange, Knime

Numerical computing environments - harder to use, but more general

- Matlab (proprietary), Scilab, Octave

Programming languages - hardest to use, but most general and professional

- Python, R, Julia, Java, Scala

Sources: https://commons.wikimedia.org/wiki/File:Weka_(software)_logo.png, https://commons.wikimedia.org/wiki/File:Matlab_Logo.png,https://fr.wikipedia.org/wiki/Fichier:Python_logo_and_wordmark.svg

Python for the labworks

- We will use Python 3, which is distributed with Anaconda

  Doc.:
  https://docs.python.org/3.8/tutorial/index.html

- We will also use Jupyter Notebook: web application allowing to create and share documents with code, text, figures and equations.

  Doc.: https://jupyter-notebook.readthedocs.io/en/latest/notebook.html#the-jupyter-notebook

- You can download Anaconda (Python 3 and Jupyter are included) from

  $\Longrightarrow$ https://www.anaconda.com/download/

- If you do not want to install Anaconda but you have a *google* account, you can use *google colab* :

  $\Longrightarrow$ https://colab.research.google.com

Python for the labworks

Useful Python libraries

- Numpy: support for vectors, matrices and multi-dimensional arrays along with high-level mathematical functions to operate on these arrays

  Doc.: `https://numpy.org/doc/stable/user/quickstart.html`

- Scipy: library for scientific and technical computing. It contains modules for optimization, linear algebra, integration, interpolation, signal/image processing and other tasks common in science and engineering

  Doc.: `https://docs.scipy.org/doc/scipy/reference/tutorial/index.html`

Python for the labworks

Useful Python libraries

- Matplotlib: plotting library, it provides a large number of plotting options, 2D line graphs, bar graphs, scatterplots, 3D surfaces, contour plots, images, polar charts and pie charts.

  Doc.: `https://matplotlib.org/stable/tutorials/index.html`

- **Scikit-learn**: machine learning library featuring algorithms for regression, classification, dimensionality reduction and clustering.

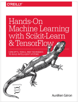
  Doc.: `https://scikit-learn.org/stable/tutorial/index.html`

Bibliography

Some bibliography on machine learning

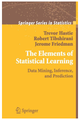
- Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. A. Géron. O'Reilly Media, Inc.

  More practical introduction



- The elements of statistical learning. J. Friedman, T. Hastie and R. Tibshirani. Springer series in statistics

  Theoretical, available on the internet



- Data science: fondamentaux et études de cas: Machine learning avec Python et R. M. Lutz, E. Biernat. Editions Eyrolles

  French, introduction level but require some math background

Evaluation guidelines

- First 45min. of the 4th class - 40% of the final grade
  - 1 A4 sheet handwritten allowed;
  - Bring a simple calculator;
  - Individual;
  - Computer and cellphone strictly forbidden.
- The test is focused on concepts from the classes (no python code), but you may be asked to do some calculations.

Conclusions and summary of main points

- Machine learning algorithms are computer programs that learn from experience (data) to improve its performance (prediction error/retrieving data structure) on a given task (prediction/data structure)

  $\Longrightarrow$ Automation of inductive reasoning

- Automation of some essentially human-labor activities in the third sector may increase efficiency and quality of services with reduced costs

  $\Longrightarrow$ Long-term effect on labor market is a great concern

Conclusions and summary of main points

- Data is the raw material of machine learning and it can be of a variety of types.

$\Longrightarrow$**Data** = $\mathbf{X}$, **rows** = observations and **columns** =features

1. Supervised learning: predict $\hat{\mathbf{y}}$ dependent feature $\mathbf{y}$ from $\mathbf{X}$
   1.1 Regression: $\hat{\mathbf{y}}$ continuous amplitude
   1.2 Classification: $\hat{\mathbf{y}}$ discrete
2. Unsupervised learning: predict $\hat{\mathbf{y}}$ only from $\mathbf{X}$
   2.1 Dim. reduction: $\hat{\mathbf{y}}$ continuous amplitude
   2.2 Clustering: $\hat{\mathbf{y}}$ discrete

Support

Moodle's course name:

UE Intro to Artificial Intelligence : Data Analysis and Machine Learning

Course code:

KMUIAIU

Course URL:

`https://lms.univ-cotedazur.fr/course/view.php?id=15732`

Access password: INTRO_IA_EUR_2021