Machine Learning for biology

V. Monbet

UFR de Mathématiques
Université de Rennes

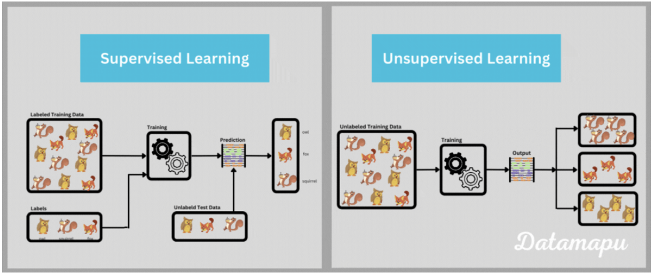
Outline

Introduction

What is "learning"?

- Machine learning gathers tools for describing and understanding complex (big) datasets.
  Machine learning and statistics are closely knit. The reason is that the methods used in most machine learning approaches have origins from statistics.

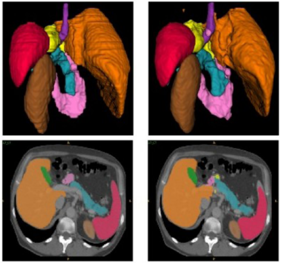
- Supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs.
- Unsupervised statistical learning, there are inputs but no output; nevertheless we can learn relationships and structure from such data.



source : `https://datamapu.com/posts/ml_concepts/supervised_unsupervised/`

Some famous applications

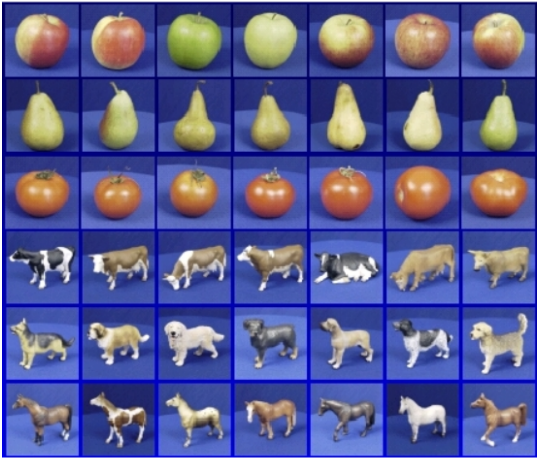
- Image recognition, image segmentation (see figure on the right)
- Medical diagnostic: cancer, alzheimer, diabetes, etc for example from microarray data analysis.
- Some other applications
  - gene prediction: A number of the sequence's genes can be identified by determining which strings of bases within the sequence are homologous to known gene sequences.
  - Functional annotation, systems biology, pathway analysis, ...
  - Animals communication recognition, landscape analysis from satelite images, etc.
- Business analytics, Google ranking, web-data, etc
- Textmining, knowledge extraction is the creation of knowledge from structured (relational databases, XML) and unstructured (text, documents, images) sources.
- Image, sound, text generation

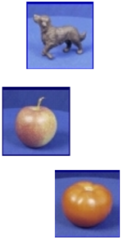


Reference standard NiftyNet segmentation

Some examples we will focus on in this course ▸ here

Image recognition example

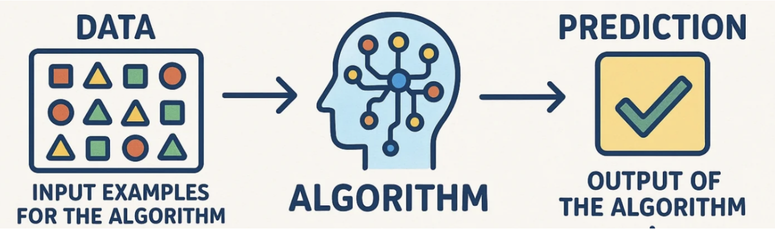




- Build the function : $f($ $) =$ Tomato

$$f : \mathbb{R}^{128\times 128} \rightarrow \{\text{Apple, Pear, Tomato, Cow, Dog, Horse}\}$$

A common process: learning from data



- Given examples (training data), make a machine learn how to predict on new samples, or discover patterns in data
- Statistics + optimization + computer science
- Gets better with more training examples and bigger computers

Schedule/Challenges

Schedule

- 12 × 1h30 lectures.
  Teachers : Valérie Monbet & Valérie Gares
  Slides available on Moodle
  Course name is "Machine Learning for Biology",
  Self registration (auto-inscription) with password MLB2019
- 3 labs per group + individual help
- Homework's material to practice is available ▸ here
  Applications on various datasets : composition of old potteries, NIR spectrometry data of cookies, Microarray data & cancer relapse, Images of hand-written digits.
- Final evaluation based on a project (work in groups of 2 to 4 students): talk + slides + codes (Rmarkdown/Python notebook).
  Project topics are proposed ▸ here but you can work on another one **after validation by one of the teacher!**

Objectives

At the end of the course, you should

- have an overview of the main machine learning methods and be able to choose one of them for a specific application ;
- be able to use the methods carrefully and explain shortly how it works to non specialist scientists.

And hopefully, you will

- have understand what an optimization problem is, in the context of fitting/calibrating a model ;
- be able to deal with the problems of high dimension and overfitting.

Which methods?

- Unsupervised learning
  - Dimension reduction

    PCA, MDS, t-SNE, Auto-encoders
  - Clustering

    k-means, Gaussian mixture model (GMM)
  - Density estimation feature learning

- Supervised learning
  - Regression

    kNN, Trees, Ridge/Lasso regression, Neural networks, Deep learning, SVR
  - Classification

    kNN, Trees, Neural networks, SVM
  - Structured output classification

References

- Hastie, T., Tibshirani, R. & Friedman, J. (2008). The Elements of Statistical Learning; Data Mining, Inference and Prediction. Springer. ▸ link
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (Vol. 112). Springer. ▸ link
- Shalev-Shwartz, S., and Ben-David S., Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. ▸ link
- Tuffery, S., (2017). Modelisation Predictive et Apprentissage Statistique avec R. Technip.