Model-based Statistical Learning

Pr. Charles BOUVEYRON

Professor of Statistics
Chair of the Institut 3IA Côte d'Azur
Université Côte d'Azur & Inria

charles.bouveyron@univ-cotedazur.fr
@cbouveyron

Preamble

"Ce qui est simple est toujours faux.
Ce qui ne l'est pas est inutilisable."

Paul Valéry

Outline

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The recent and impressive NN results should not hide the remaining issues:

- deep learning has impressive results in a few specific cases and with a high-level supervision,
- use of DL techniques in various fields are promising but not well understood.

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The recent and impressive NN results should not hide the remaining issues:

- deep learning has impressive results in a few specific cases and with a high-level supervision,
- use of DL techniques in various fields are promising but not well understood.

"*Artificial Intelligence: the revolution hasn't happened yet*"
M. Jordan (UC Berkley)

Open problems of AI

Some open problems are critical:

- reliability of models and algorithms,
- handling data heterogeneity (categorical, functional, networks, images, texts, ...),
- unsupervised learning (clustering, dimension reduction),
- learning from HD and small data (n small / p large),

Open problems of AI

Some open problems are critical:

- reliability of models and algorithms,
- handling data heterogeneity (categorical, functional, networks, images, texts, ...),
- unsupervised learning (clustering, dimension reduction),
- learning from HD and small data (n small / p large),

Combination of statistical theory with deep learning techniques is certainly the future of AI!

AI in France

French policy for AI:

- C. Villani presented in March a recommendation report for AI,
- President Macron announced the creation of a network of AI institutes.

The 3IA institutes:

- 12 french research centers applied for the 3IA call in Sept.,
- 4 projects have been selected in the Spring 2019:
  - Paris, Toulouse, Grenoble
  - and Nice!

A few examples: Cervical cancer detection

Cervical cancer detection:

- it is an important public health field which is currently treated mostly manually,
- screening by human experts is complicated by the amount of cells (20 000/smear),
- and by the very small proportion of cancer cells (less than 1%).

frequency
asymmetry

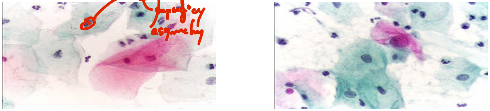


Figure: Normal (left) and abnormal (right) pap smears.

Classification is useful in this context:

- for building supervised classifiers which can select the most likely cancer cells,
- for helping experts in labeling the learning data through weakly-supervised classification,
- for selecting discriminative variables which can be used in a semi-automatic process.

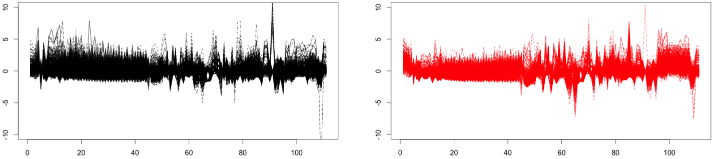


Figure: Control and (cervical) cancer data

A few examples: Sparse models in Medicine (HEGP)

Problem:

- overcome the curse of dimensionality that occurs in Metabolomics,
- for disease diagnostic and early-stage marker identification,
- metabolomic data fall into the "ultra-high dimensional data" case.

Our solution:

- a Bayesian variable selection technique for PCA,
- that identify the relevant variable for each stage of the disease.

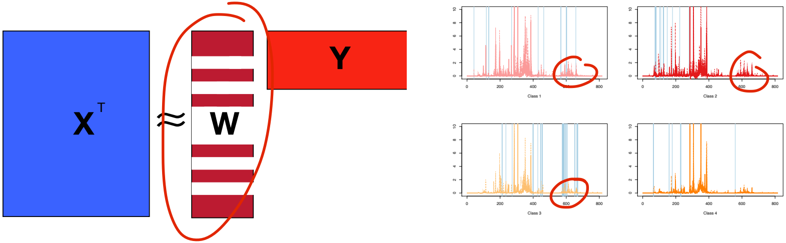


Figure: gsPPCA and variable selection on MNR specrta for CKD diagnosis.

Analysis of massive functional data (Linky / EDF)

Problem:

- Linky meters will allow EDF to have access to 27 million of Linky data,
- data are functional data and are measured every 30 minutes -> 17 520 obs./year,
- necessity to summarize those massive data before exploitation.

Our solution:

- a statistical co-clustering technique for functional data,
- that form homogeneous groups of both individuals and days.

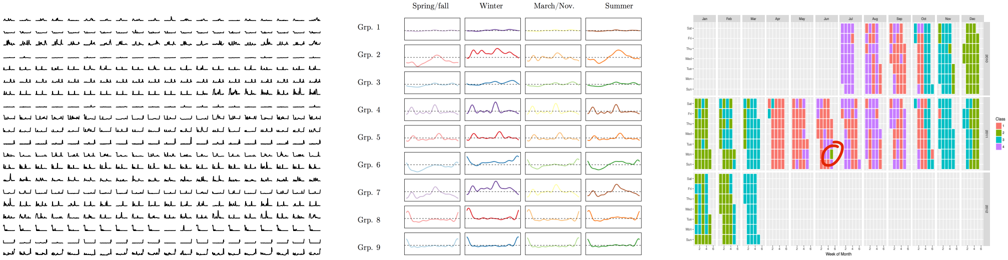


Figure: Functional co-clustering of Linky data (EDF).

Outline of the course

The course will be organized as follows:

1. Introduction to model-based statistical learning (CB)
2. Linear models for classification (PAM)
3. Mixture models and the EM algorithm (CB)
4. Another view on the EM algorithm (PAM)
5. Practical work (1st evaluation, PAM)
6. Between supervised and unsupervised classification (PAM)
7. Practical work (2nd evaluation, CB)
8. Missing values (~~PAM~~ AS)
9. Model-based image analysis (CB)
10. Co-clustering (CB)

Outline

Learning from data...

One task, several families of approaches:

- Statistical learning

$$X \xrightarrow{\text{learn}} \hat{f}_X = f_{\hat{\theta}_X} = f_X(\hat{\theta}) \xrightarrow{\text{predict}} f_X(\hat{\theta}, x^*) = \hat{y}^*$$

- Machine learning

$$X \xrightarrow{\text{learn}} \hat{f}_X \xrightarrow{\text{predict}} \hat{f}(x^*) = \hat{y}^*$$

- Deep learning

$$X \xrightarrow{\text{learn}} \hat{f}_X \text{ where } f \text{ is a very complex function} \xrightarrow{\text{predict}} \hat{f}(x^*) = \hat{y}^*$$

- ...

Learning from data...

Learning is a two-head problem:

Supervised	Unsupervised
(X, Y) ← target variable; ← explanatory variables. In this supervised context, we need examples of both variables to learn the predictor f	In this situation, we only observe X and we would like to infer Y from it.

$f = f_{X,Y}$

$(X, Y) \longrightarrow f_{X,Y}$

$\boxed{\text{semi-supervised learning.}}$

Learning from data...

Methods are specific to each task:

Supervised	Unsupervised
• classification y is categorical	• clustering $X \xrightarrow{\text{predict}} y$ is categorical
• regression: y is continuous.	• dimension reduction / representation learning $X \xrightarrow{\text{predict}} y$ is (multivariate) continuous
• time series forecasting	• image denoising
• ...	

Supervised learning

Supervised learning is also a field with different sub-tasks:

- classification:

  (X, Y) Y is categorical

  - Logistic regression
  - decision trees
  - SVM
  - Naive Bayes
  - LDA
  - k-NN

- regression:

  (X, Y) Y is continuous

  - linear regression
  - d.trees
  - SVM
  - k-NN.

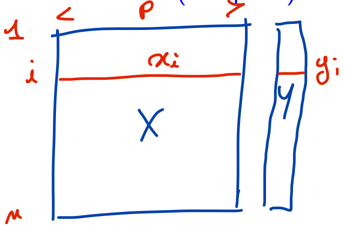
- time series analysis:

  - ARIMA
  - ...

- ...

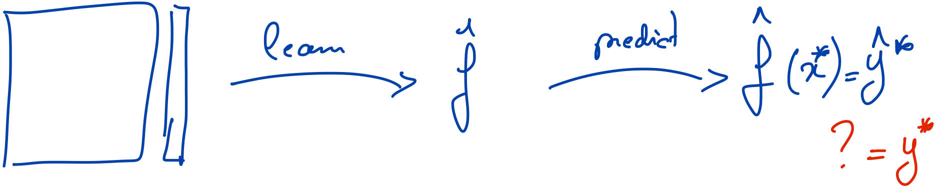
The supervised learning process

The material: a set of (complete) data



(X,Y) is called the learning data set.

The goal: learn a predictor $f(.)$ from the (complete) data



Measuring the learning performance

One comfortable thing of working in the supervised context is:

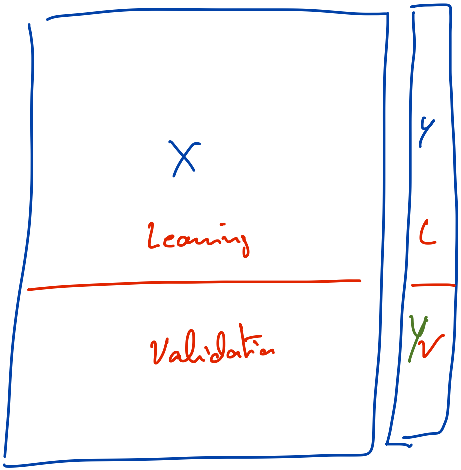
- to be able to measure the performance of the learned predictor,
  - classification error : $e_{\hat{f}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{y}_i \neq y_i\} \in [0,1]$
  - regression error : $MSE = \frac{1}{n} \sum_{i=1}^{n} \| \hat{y}_i - y_i \|^2 \geq 0$
- compare several predictors and pick the most efficient one.

$f_1 \longrightarrow e_{\hat{f}_1} = 0.05$

$f_2 \longrightarrow e_{\hat{f}_2} = 0.04$

A minimal setup for supervised learning

The minimal setup for building a supervised predictor $f()$ from data is as follows:



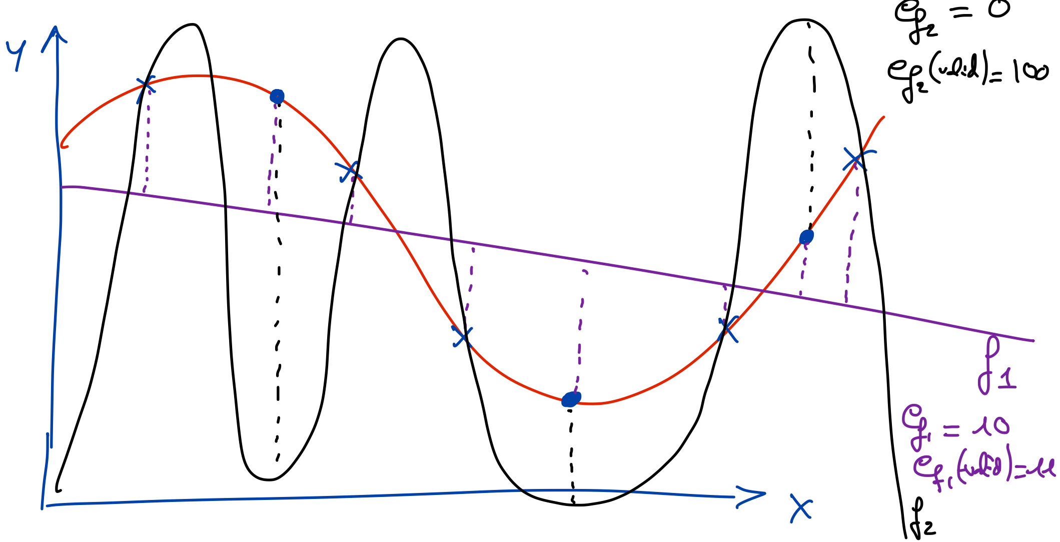
$$\xrightarrow{\text{learn}} \hat{f}$$

$$\hat{f}(x_v) = \hat{y}_v \stackrel{?}{=} y_v$$

$$e_{\hat{f}}$$

Why such a minimal setup?

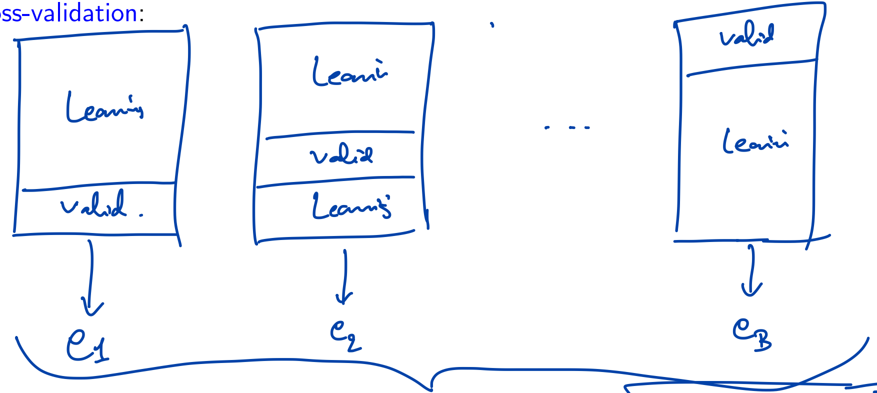
The goal is to avoid over-fitting when choosing the model or the model parameters:



An advanced setup for supervised learning

Resampling techniques:

- there are several methods (leave-one-out, V-fold cross-validation, bootstrap) depending on the context (sample size, computing time, ...),
- V-fold cross-validation:



$$\bar{e} = \frac{1}{B}\sum_b e_b \quad \pm \quad \sigma = \sqrt{\frac{1}{B}\sum (e_b - \bar{e})^2}$$

$$e_{f_1} = 0.05 \pm 0.02$$

$$e_{f_2} = 0.04 \pm 0.08$$

In the case of comparing methods with tuning parameters, we have to use double-CV to evaluate correctly the average performance of the method.