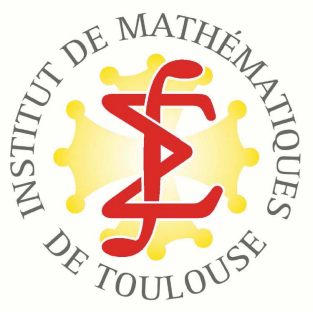


Institut de Mathématiques de Toulouse, INSA Toulouse

Introduction

Machine Learning

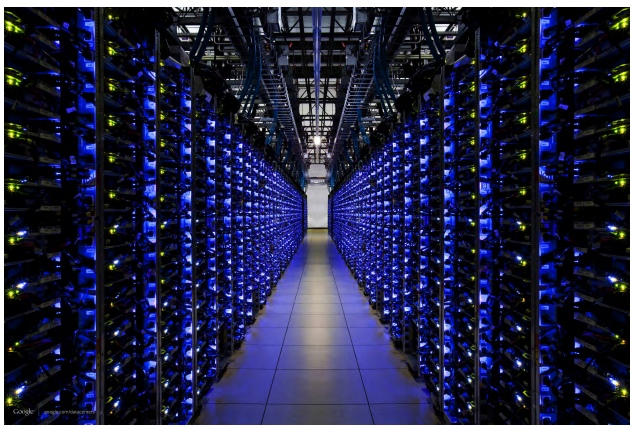
Béatrice Laurent - Philippe Besse - Mélisande Albert

Outline

- From statistic to IA
- Tutorials and datasets
- Introduction to supervised learning
- Strategy for statistical learning

Introduction

- Statistical learning plays a key role in many fields of sciences, medicine, industry, marketing, finance ...
- The development of data storage and computing resources leads to the storage of a huge amount of data from which the data scientist will try to learn crucial informations to better understand the underlying phenomena or to provide predictions.



- Many fields are impacted, here are some examples of learning problems:
  - **Signals:** Aerospace industry produces a huge amont of signal measurements obtained from thousand of on-board sensors. It is particularly important to detect possible anomalies before launching the satellite. Similarly, many sensors are involved in planes and it is important to detect a abnormal behavior on a sensor.
    $\hookrightarrow$ curve clustering or classification and anomaly detections in a set of curves for predictive maintenance purposes.

  - **Images:** Convolutional neural networks and deep learning lead to important progresses for image classification. Many fields are concerned: medical images (e.g. tumor detection), earth observation satellite images, photos, video surveillance images, handwritten text images ...

  - **Geolocalisation data:** Machine learning based on geolocalisation data has also many potential applications: targeted advertising, road traffic forecasting, monotoring the behavior of fishing vessels ...

- **Consumers preferences data:** Websites and supermarkets collect a huge amount of data on the behavior of consumers. Machine learning algorithms are used to valorize these data (gathered sometimes with personal data such as age, sex, job, adress ..) for recommandation systems ..

- **High-throughput sequencing data:** RNA-seq data allow to measure the expression of thousands of genes simultaneously on a single individual. It is, for example, a challenge to try to infer from those kind of data which genes are involved in a certain type of cancer. The number p of genes measured on a microarray is generally much larger than the number n of individuals in the study.
  $\hookrightarrow$ Variable selection in high dimension.

From Statistic to AI through *Data Science*

1930-70 h-Bytes Statistical inference

1950 Beginnings of Artificial Intelligence: Allan Turing

1970s kB Data analysis and *exploratory data analysis*

1980s MB Neural networks, functional data analysis

1990s GB *Data mining*: pre-acquired data

2000s TB Bioinformatics: $p \gg n$, *Machine Learning*

2008 Data Science

2010s PB Big Data p and n very large

2012 *Deep Learning*

2016 Artificial Intelligence (AI): AlphaGo

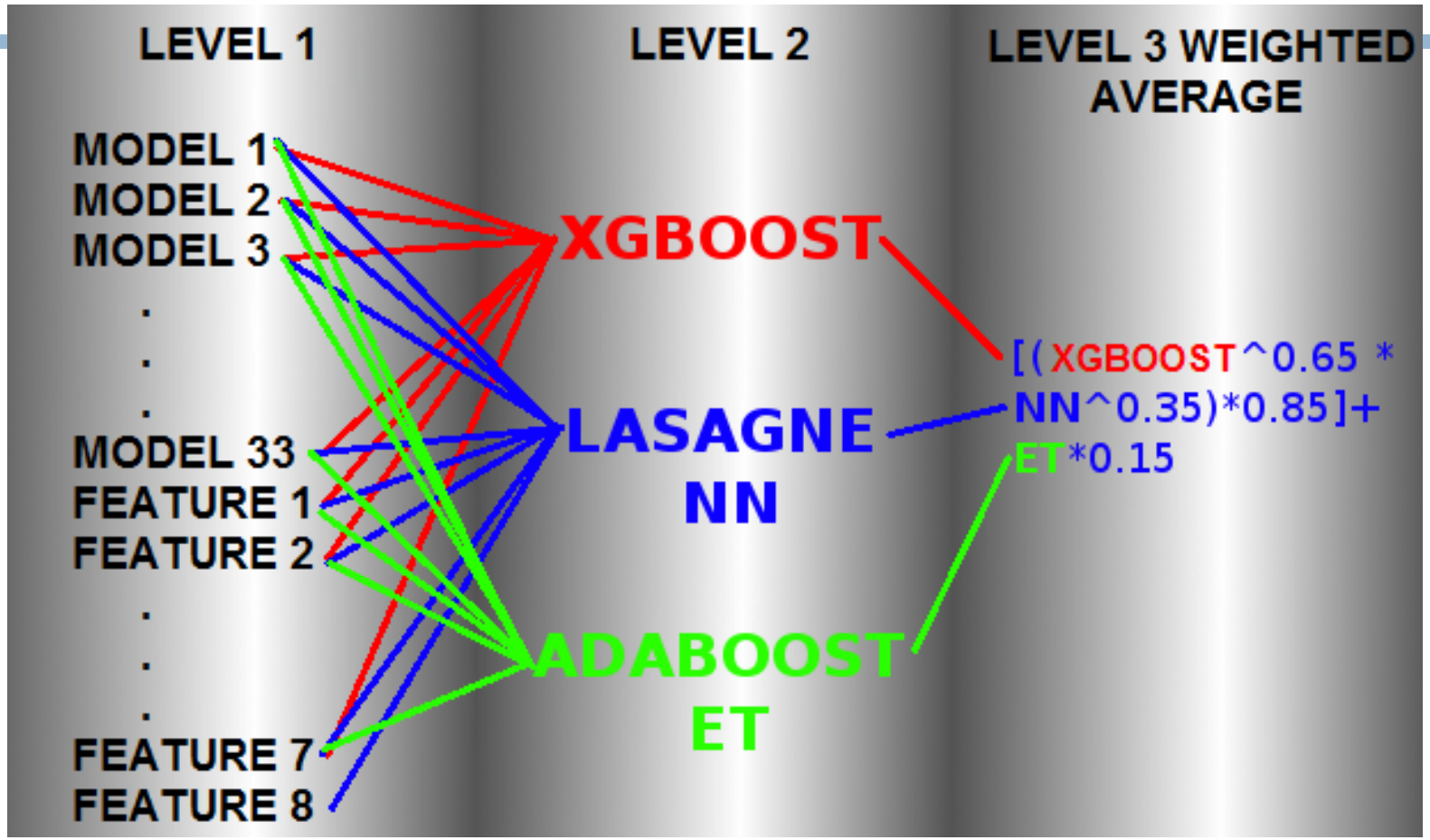
VVV...VV : Volume, Variety, Velocity... Veracity, Valorization

Objectives

- Exploration: description, visualisation, clustering (taxonomy)
- Explanation $Y = f(\boldsymbol{X})$ (supervised learning)
- Prediction and selection, explainable and interpretable models
- "Raw" forecasting (black box models)
- Anomaly Detection

Aim

- Academic publication (*Benchmarks — UCI repository*)
- Valorisation, Industrial solutions
- Kaggle type competition.



Concours Kaggle: Identify people who have a high degree of Psychopathy based on Twitter usage.

Outline

- From statistic to IA
- Tutorials and datasets
- Introduction to supervised learning
- Strategy for statistical learning

Tutorials and data sets

- The analysis of these different usecases is presented in tutorials contained in **jupyter notebooks** in R or Python.
- They are available in the repository `https://github.com/wikistat/Apprentissage`
- We present here some of the datasets that will be considered.

Usecase Ozone

Aim: Prediction of the ozone concentration for the next day at 5 PM (max. of the day) from a learning sample composed of the explanatory variables $X^1, \ldots X^p$:

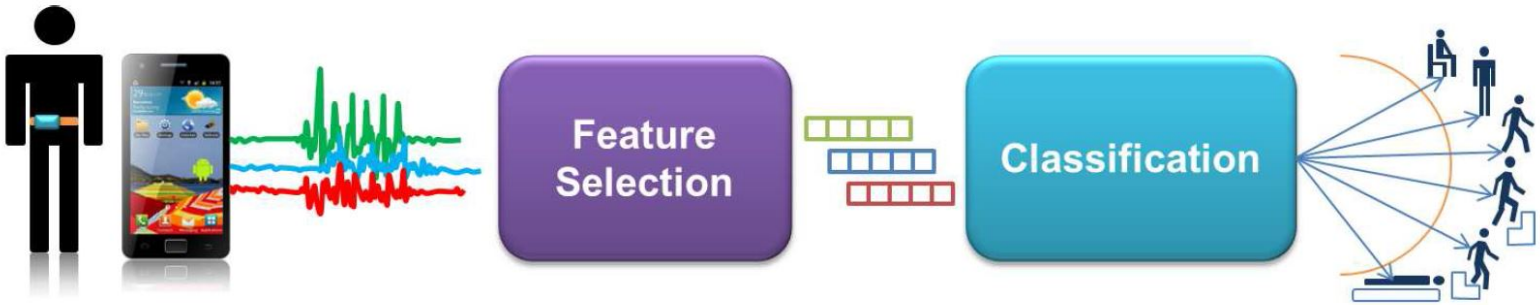
- MOCAGE (deterministic model of Meteo France)
- NO2, NO3
- H2O
- Temperature
- Wind speed and orientation
- Station
- Type of day (holiday or not)

and the variable to explain:

- Y:Ozone concentration

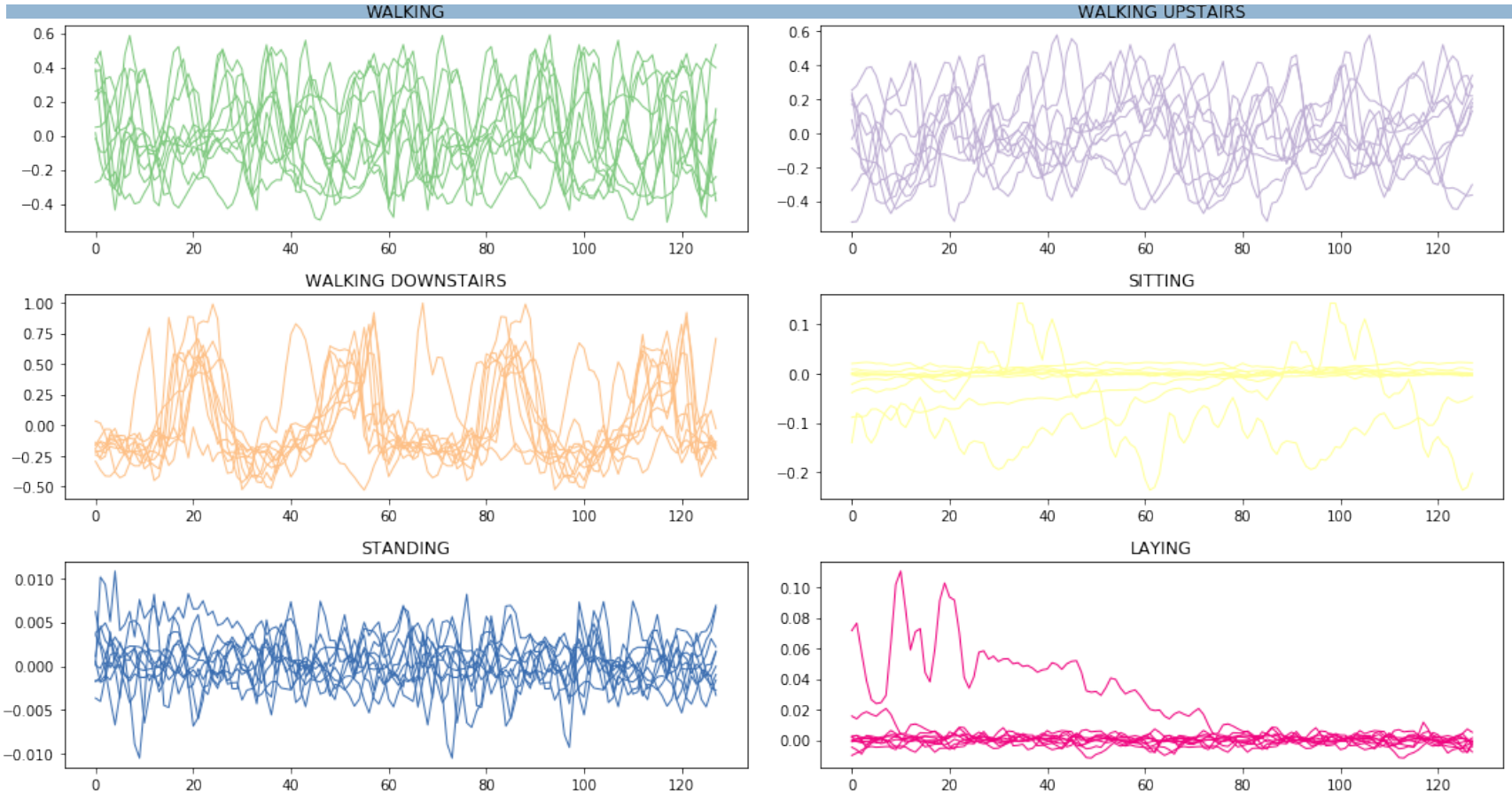
$\hookrightarrow$: **Statistical adaptation**

Usecase HAR



Human activity recognition HAR

- Public data available on *UCI repository*
- 9 signals per individual: The accelerations in $x, y, andz$, those by subtracting the natural gravity and the angular accelerations in x, y, and z obtained from the gyroscope.
- Each signal contains $p = 128$ measures sampled at 64 htz during 2s.
- 7352 samples for learning and 2947 for testing.
- Objectives: Activity recognition (6 classes) standing, sitting, lying, walking, walking upstairs or walking downstairs.



Human activity recognition: acceleration in y by class

HAR First step: "features" variables obtained from signal processing

- $p = 561$ new variables (*features*)
  - Time domain: min, max, means, variances, correlations...
  - Frequency domain: largest, mean, energy per frequency band...

HAR ... to be continued

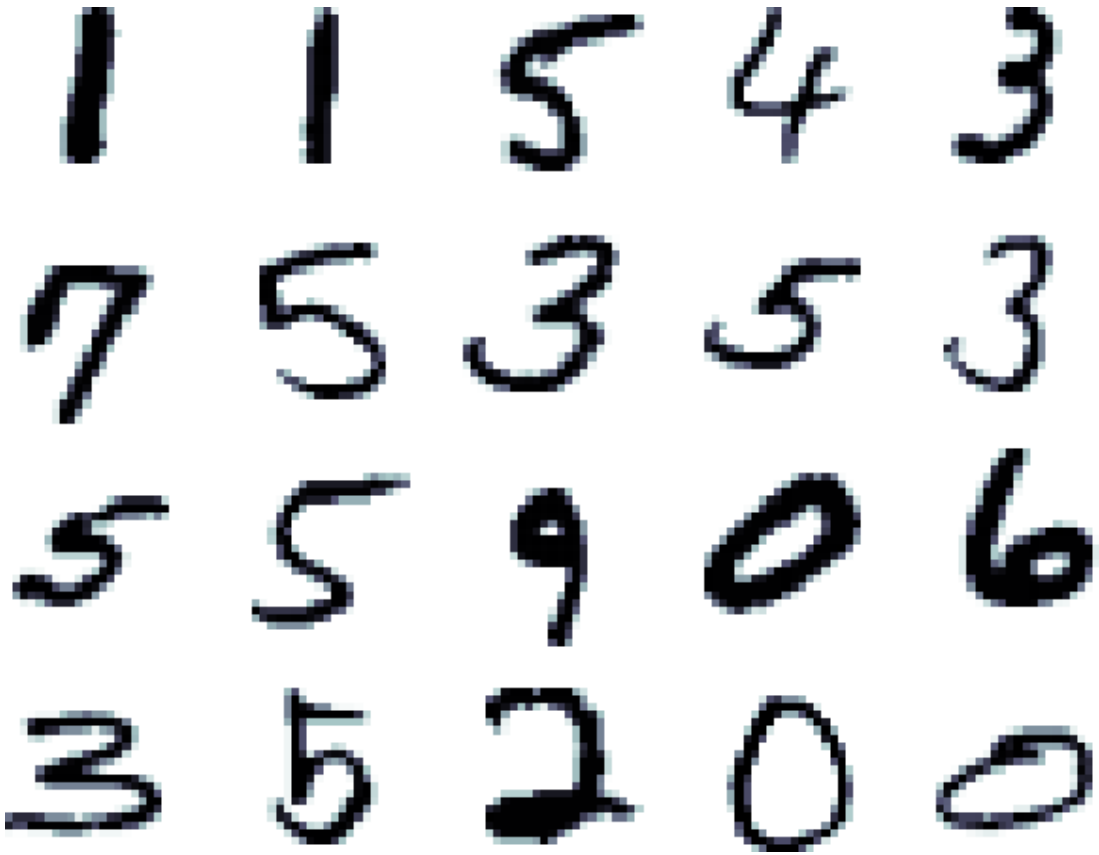
- raw signals and *deep learning*

Usecase MNIST

MNIST dataset

- Yann le Cun website
- 60 000 handwritten digits, $28 \times 28 = 784$ pixels
- Test: 10 000 images
- Classical methods (k-nn, Random Forests)
- Preprocessing: normalisation of the images
- Specific Distance with invariance properties
- Deep learning: *TensorFlow, Keras*

Usecase MNIST



MNIST: some examples of handwritten digits

Outline

- From statistic to IA
- Tutorials and datasets
- Introduction to supervised learning
- Strategy for statistical learning

Non supervised vs Supervised learning

- In the framework of **Supervised learning**, we have a **Learning sample** composed with observation data of the type **input/output**:

$$d_1^n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$$

with $\boldsymbol{x}_i = (x_i^1, x_i^2, \ldots, x_i^p)$ is a p-dimensional variable (quantitative or qualitative or both), $y_i \in \mathcal{Y}$ for $i = 1 \ldots n$ is the variable to explain (label).

Objectives : From the learning sample, we want to

- **Estimate** the link between the input vector $\boldsymbol{x}$ (explanary variables) and the output y:

$$y = f(x^1, x^2, \ldots, x^p)$$

- **Predict** the output y associated to a new entry $\boldsymbol{x}$,
- **Select** the important explanatory variables among $x^1, \ldots, x^p$.

Non supervised vs Supervised learning

- **Supervised learning**,

quantitative output	qualitative output
$\mathcal{Y} \subset \mathbb{R}^p$	$\mathcal{Y}$ finite
$\downarrow$	$\downarrow$
regression	**classification**

- In the framework of **non supervised (or unsupervised) learning**, we only observe $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

  Objectives :

  - Find underlying structures in these unlabeled data.
  - Clustering.

Supervised learning

- We consider **supervised regression or classification problems**.
- We have a training data set with n observation points (or objects) $\boldsymbol{X}_i$ and their associated output Y_i (real value in regression, class or label in classification).
- $\boldsymbol{d}^n$ corresponds to the observation of a random n-sample $\boldsymbol{D}^n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ with joint unknown distribution P on $\mathcal{X} \times \mathcal{Y}$.
- A **prediction rule** is a measurable function $\hat{f} : \mathcal{X} \to \mathcal{Y}$ that associates the output $\hat{f}(\boldsymbol{x})$ to the input $\boldsymbol{x} \in \mathcal{X}$.

Loss function

- In order to quantify the quality of the prevision, we introduce a loss function.

Definition

A measurable function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a *loss function* if $\ell(y, y) = 0$ and $\ell(y, y') > 0$ for $y \neq y'$.

- **In real regression,** it is natural to consider $\mathbb{L}^p$ $(p \geq 1)$ losses

$$\ell(y, y') = |y - y'|^p.$$

- If $p = 2$, the $\mathbb{L}^2$ loss is called "quadratic loss".
- **In classification,** one can consider the 0-1 loss defined, for all $y, y' \in \mathcal{Y}$ by

$$\ell(y, y') = \mathbb{1}_{y \neq y'}.$$

Since the 0-1 loss is not smooth, it may be useful to consider other losses that we will see in the classification courses.

Generalization error

Definition

Let f be a prediction rule defined from the learning sample $\boldsymbol{D}^n$. Given a loss function ℓ, the **risk** - or **generalisation error** - of the prediction rule f is defined by

$$R_P(f) = \mathbb{E}_{(\boldsymbol{X},Y)\sim P}[\ell(Y, f(\boldsymbol{X}))],$$

where, in the above expression, $(\boldsymbol{X}, Y)$ is independent from the learning sample $\boldsymbol{D}^n$.

An accurate evaluation of the generalization error has two objectives:

- **Model selection:** selecting, among a collection of models (or prediction rules), the one with the smallest risk, realizing the best bias/variance trade-off.
- **Model assessment:** once the final model has been chosen, evaluating its generalization error on a new data set.

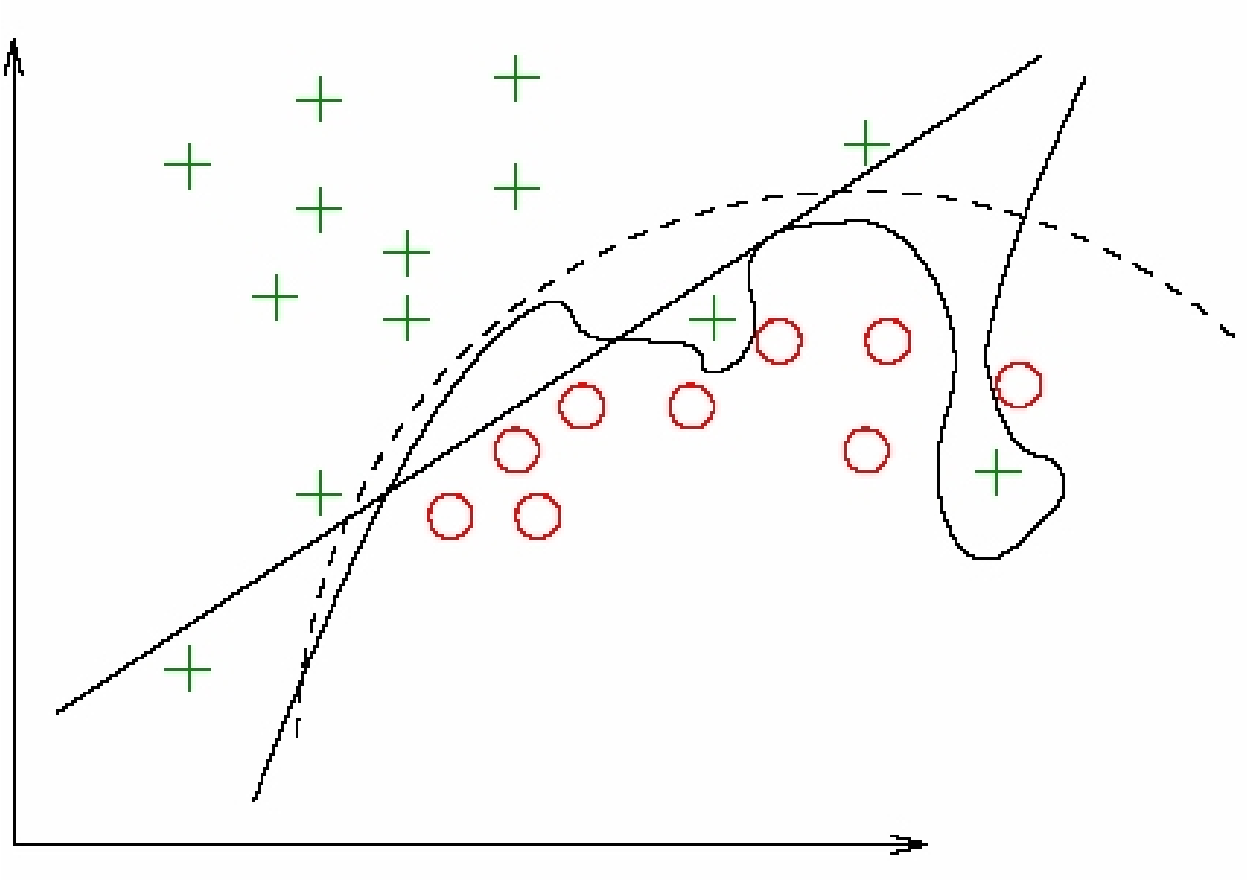
Training error

- In practice, we have a training sample $\boldsymbol{D}^n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ with unknown joint distribution P, from which we construct a regression or classification rule.
- The aim is to find a "good" classification rule, in the sense that its risk is as small as possible.
- A first idea to estimate the risk $R_P(f) = \mathbb{E}_{(\boldsymbol{X},Y)\sim P}[\ell(Y, f(\boldsymbol{X}))]$ is to consider its empirical estimator, called **empirical risk**, or **training error**:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(\boldsymbol{X}_i)).$$

- This is not a good idea: this estimator is optimistic and will under estimate the risk (or generalization error) as illustrated in the following binary classification example.

Supervised learning: risk of overfitting



Model complexity in supervised classification

Estimation of the risk

- The **generalization** performance of a learning procedure is related to its prediction capacity on a **new data set**, independent of the learning sample that was used to build the learning algorithm.
- If we have enough data, the recommended approach is to divide randomly the dataset in two parts: the train sample and the test sample, the train sample being itself divided into a learning sample and a validation sample.
  - **The learning sample** is used to train the models (generally by minimizing the training error).
  - **The validation sample** is used for model selection: we estimate the generalization error of each model with the validation sample and we select the model with the smallest generalization error.
  - **The test sample** is used for model assessment, to evaluate the risk of the final selected model.

  It is generally recommended to take 50% of the data for the learning sample, 25% of the data for the validation sample and 25% of the data for the test sample.

Outline

- From statistic to IA
- Tutorials and datasets
- Introduction to supervised learning
- Strategy for statistical learning

Preparation of the data

The analysis, also called the **Data science** follows the steps described below for most fields of application.
The three first steps correspond to initial preparation of the data, **data munging**, preliminary essential before the modeling or learning phase.

- Data extraction with or without sampling applied to structured databases (SQL) or not (NoSQL)
- Visualization, exploration of the data for the detection of atypical values, errors or anomalies; study of distributions and correlation structures and search for transformations of variables, construction of new variables and / or representation in adapted bases (Fourier, spline, wavelets ...).
- Taking into account missing data, by simple deletion or by imputation.

The steps of a statistical analysis

- Random partition of the sample into a **train set** and a **test set** according to its size and choice of a loss function that will be used to estimate the prediction error.
- **The train set** is separated into a **learning sample** and a **validation sample**. For each method considered: generalized linear model (Gaussian, binomial or Poisson), parametric (linear or quadratic) or nonparametric (k nearest neighbors), neural network (perceptron), binary decision tree, support vectors machine, aggregation (*bagging, boosting, random forest*. . .)
  - Estimate the model with the **learning set** for given values of a parameter of *complexity*: number of variables, neighbors, leaves, neurons, penalization or regularization . . .
  - optimization of this parameter (or these parameters) by minimizing the empirical loss on the **validation set**, or by cross-validation on the **train set** or the training error plus a penalty term.
- Comparison of the previous optimal models (one per method) by estimating the prediction error on the **test set**.

The steps of a statistical analysis

- Possible iteration of the previous approach or **Monte Carlo** cross-validation: if the test sample is too small, the prediction error can be very dependent on this test sample.
  The Monte Carlo cross-validation approach consists in successive random partitions of the sample (train and test) to study the distribution of the test error for each model or at least take the mean of the prediction errors obtained from several Monte-Carlo iterations to ensure the robustness of the final selected model.
- Choice of the "best" method according to its prediction error, its robustness but also its interpretability if necessary.
- Re-estimation of the selected model on all the data.
- Industrialization: implementation of the model on the complete data base.

- The end of this process can be modified by building a combination of the different methods rather than selecting the best one.
- This is often the case with winning "gas factory" solutions in *Kaggle* competitions.

The methods or algorithms

We will see during this course the most widespread learning methods.

- Estimation of the prediction error of an algorithm: this is a crucial step to choose the "best" prediction rule among a collection and to evaluate the performances of the selected procedure
- Reminders on linear models and logistic regression
- Model selection for linear models via penalized criterion: Mallows CP, BIC, Ridge, Lasso...
- Linear methods for classification: the Linear (and quadratic) Discriminant Analysis and the Linear Support Vector Machine (SVM)

The methods or algorithms

- Kernel methods: Support Vector Machine (SVM) and Support Vector Regression
- Classification And Regression Trees (CART algorithm), aggregating methods (bagging) and Random Forests
- Aggregation by boosting algorithms
- Neural networks: multilayer perceptron, backpropagation algorithms, optimization algorithms, introduction to deep learning
- Imputation of missing data
- Ethical aspects of statistical decisions and legal and societal impacts of AI.

References

- Hastie, T. and Tibshirani, R. and Friedman, J, (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer.