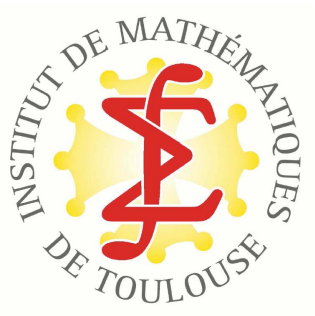


Institut de Mathématiques de Toulouse, INSA Toulouse

High dimensional and Deep Learning Introduction

High Dimensional and Deep Learning
INSA- Applied Mathematics

Béatrice Laurent

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. For example, before the launch of a satellite, many tests are provided to observe the behavior of the satellite in various conditions. 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. Those examples concern curve clustering or anomaly detections in a set of curves.

- **Images :** More and more images are collected and stored, for example medical images, earth observation satellite images, photos, video surveillance images, handwritten text images ...Each image is made from a huge number of pixels. Examples of learning problems are handwritten digit recognition, tumor detection, image classification ..
- **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 ..
- **Microarray data :** DNA microarrays 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.

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 $\boldsymbol{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$ for $i = 1 \ldots n$.

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

Objectives : From the learning sample, we want to

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

$$y = f(\boldsymbol{x})$$

- **Predict** the output y associated to a new entry $\boldsymbol{x}$,
- **Select** the important explanatory variables among the components of the input $\boldsymbol{x}$.
- **Predict abnormal behaviors.** (Anomaly detection is most of the time in a non supervised framework)

- In this course, we focus on learning from high dimensional and complex data such as signals or images. Learning from high dimensional data is challenging for several reasons.
  - The curse of dimensionality : In high dimension, data points are isolated.
    We consider a supervised classification problem where the predictor X is p-dimensional. Assume that P_X be the uniform distribution on the unit hypercube of $\mathbb{R}^p$. We want to use a local method such as nearest neighbors, with a proportion q of neighbors (with a majority vote) . In order to select a proportion q of the observation points, we have to select an hypercube of side length $q^{1/p}$.
    In the case where $p = 10$:
    $q = 1\% \Rightarrow q^{1/p} = 0.63$;
    $q = 10\% \Rightarrow q^{1/p} = 0.80$ (close to 1 !).
    When $p = 100$:
    $q = 10\% \Rightarrow q^{1/p} = 0.98$.

- **High dimension leads to high variance** Let us consider a linear model with n observations.

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \cdots + \beta_p X_i^p + \varepsilon_i \quad i = 1, 2, \ldots, n$$

where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$. The variables ε_i are assumed to be i.i.d., centered, with variance σ^2.
Let $\mathbf{X}(n \times (p+1))$ the matrix with general term X_i^j, where the first column contains the vector $\mathbf{1}$ ($X_0^i = 1$), and let $\mathbf{Y}$ the vector containing the responses Y_i. We set $\boldsymbol{\varepsilon} = [\varepsilon_1 \cdots \varepsilon_p]'$ and $\boldsymbol{\beta} = [\beta_0 \beta_1 \cdots \beta_p]'$. We have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

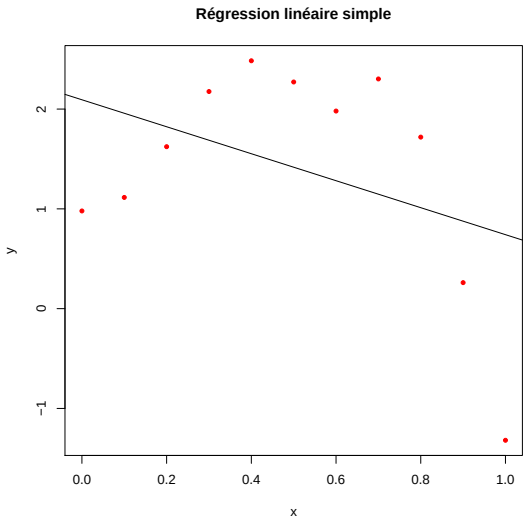
The least square estimator of $\boldsymbol{\beta}$ minimizes $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$.
If the matrix $\mathbf{X}$ is of full rank $p+1$ (or equivalently if $\mathbf{X}'X$ is invertible), the solution is unique and equals :

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

- This estimator is unbiased, with quadratic risk, :

$$\mathbb{E}\left(\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2\right)=p\sigma^2.$$

- Hence, in high dimension, it might be preferable to have a biased estimator, with lower variance, and hence possibly lower quadratic risk (which is the sum of the squared of the bias and the variance).
- Let us illustrate this bias-variance trade-off on a pedagogical example for polynomial regression.



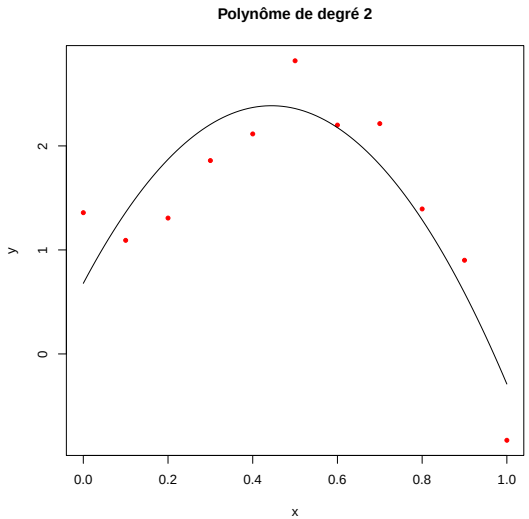


FIGURE – Polynomial regression : adjusted model, on the left : $y = \beta_0 + \beta_1 x + \epsilon$, $R^2 = 0.03$, on the right : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, $R^2 = 0.73$.

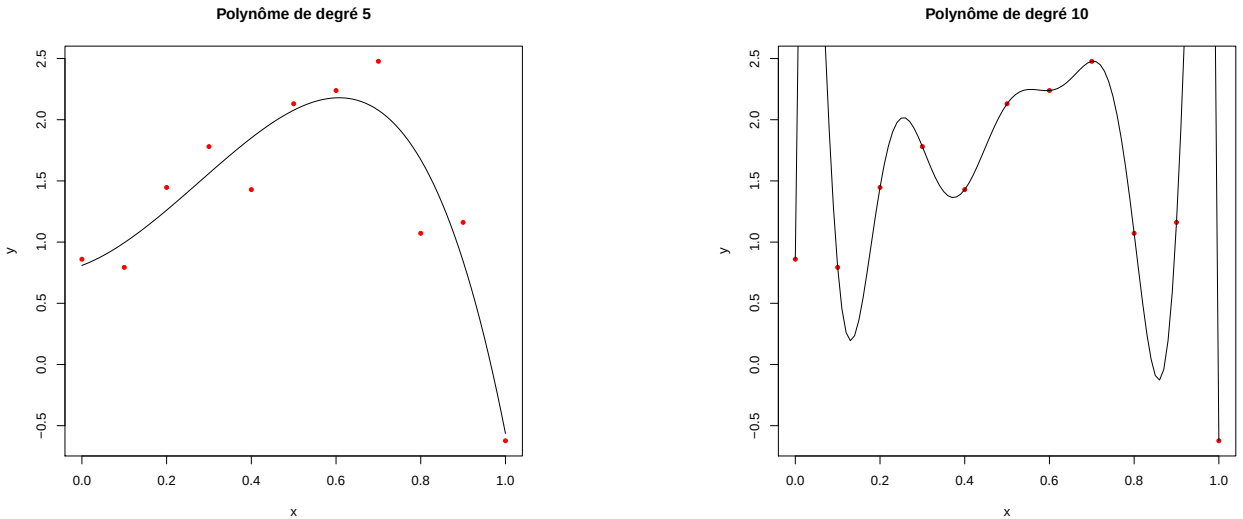


FIGURE – Polynomial regression : adjusted model, on the left : $y = \beta_0 + \beta_1 x + \ldots + \beta_5 x^5 + \epsilon$, $R^2 = 0.874$, on the right : $y = \beta_0 + \beta_1 x + \ldots + \beta_{10} x^{10} + \epsilon$, $R^2 = 1$.

The determination coefficient is equal to 1 for the polynomial of degree $n-1$ (which has n coefficients) and passes through all the training points (overfitting !).

- We have presented some issues of statistical learning in high dimension.
- Fortunately, the data lying in high dimensional spaces are generally much simpler than they appear and can be represented in much smaller dimension spaces. The data can generally be well approximated by low dimensional structures.
- For example, handwritten digits do not correspond to p totally random pixels, they have a geometrical structure.
- Signals with a huge number of observation times can be represented with a few coefficients onto functional bases,
- Features can be extracted from images for classification purposes.

- The purpose of this course is to propose statistical methods to analyse high dimensional and complex data.
- Concerning the Machine Learning algorithms, we focus on Neural networks and Deep Learning.
- Many other algorithms will be studied in the course of Machine Learning (linear models with Lasso and Ridge penalization, SVM, random forests, Gradient Boosting, XGBoost ..).

Outline of the course

Teachers : Béatrice Laurent and Brendan Guillouet
Slides and tutorials available on Wikistat 2.0 Github

https ://github.com/wikistat

- Course 1 : Neural Networks and Introduction to Deep Learning
- Course 2 : Convolutional Neural Networks
- Course 3 : Functional data analysis
- Course 4 : Anomaly detection
- Course 5 : Autoencoders and Generative Adversarial Networks

References

- Ian Goodfellow, Yoshua Bengio and Aaron Courville : *Deep Learning*, MIT Press, http ://www.deeplearningbook.org/
- Giraud, C. *Introduction to high dimensional statistics*, CRC Press.
- Hastie T. et al. *The elements of Statistical Learning*
- Ramsay, J. O. and Silverman B.W. *Functional data analysis*, Springer
- Charles Ollion et Olivier Grisel *Deep learning course* https ://github.com/m2dsupsdlclass/lectures-labs