

Introduction to Machine Learning

Formation ENSTA ParisTech

Conférence IA

Florent Chatelain* Olivier Michel*

* GIPSA-lab, Univ. Grenoble Alpes,

10-13 février 2020
ENSTA, France

The presenters

Florent Chatelain

- ▶ Ph.D. degree in signal processing from the National Polytechnic Institute, Toulouse, France, in 2007
- ▶ Post-doc position at INRIA - ARIANA Team, 2007-2008
- ▶ Since 2008, Associate Professor at GIPSA-Lab, University of Grenoble, France.
- ▶ Research interests are centered around estimation, detection and large scale inference

Olivier J.J. Michel

- ▶ Former student ENS-Cachan (ENS Paris-Saclay), agrégation in applied physics, 1986
- ▶ Ph-D degree in signal processing from Univerisity paris 11-Orsay, 1991, Post-Doc at Univ. of Michigan, USA
- ▶ Associate prof at Lab. de Physique, ENS-Lyon 1991-1999
- ▶ Prof. at Univ. Nice Sophia-Antipolis, Astrophysics lab, 1999-2008
- ▶ Prof at GIPSA-Lab, University of Grenoble, France.
- ▶ Research interest : random processes, estimation, detection, AI for astro and geo-sciences.

Organization

Volume

- ▶ 4 × 7h lecture and practices sessions

| | Monday | Tuesday | Wednesday | Thursday | |
|--------------|----------------------|-------------|------------|------------------|----|
| 09h00-12h00 | Intro + Clustering | Neural nets | Classif.1 | PCA, KPCA | XX |
| Lunch | | | | | |
| 14h00-17h00 | Hierarchical methods | CNN | Classif. 2 | Model validation | XX |

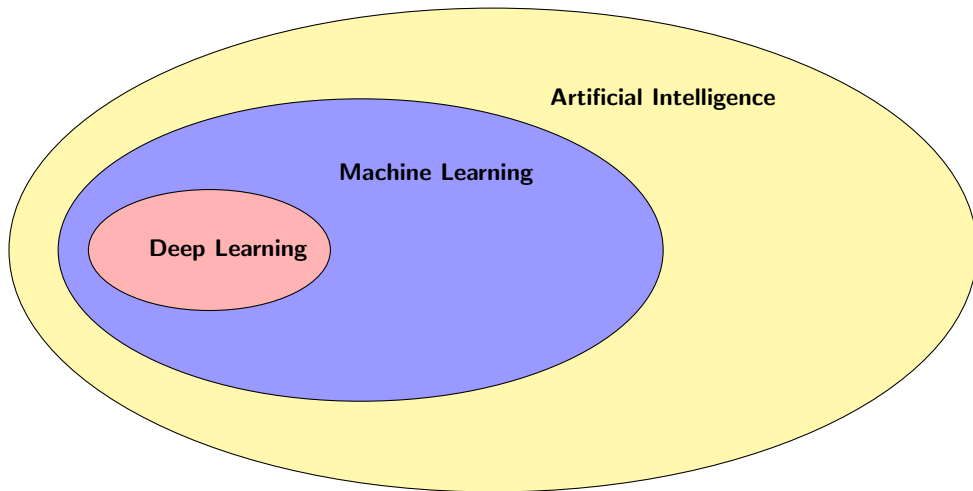
Objectives

- ▶ Understand the theoretical basis of data science/machine learning/AI
- ▶ Assess the quality of predictions and inferences
- ▶ Implement/apply data science algorithms and models using state-of-the-art frameworks

The material

- ▶ Slides (pdf) and notebooks available here :
`https://gricad-gitlab.univ-grenoble-alpes.fr/chatelaf/conference-ia/`
- ▶ Jupyter notebooks are available to illustrate concepts and methods in Python (.ipynb files)
- ▶ Binders are also available to run them remotely and interactively (no need to install Python and its dependencies, see README.md)

Machine Learning \subset Artificial Intelligence



Objective

How to extract knowledge or insights from data ?

Learning problems are at the cross-section of several applied fields and science disciplines

- ▶ *Machine learning* arose as a subfield of
 - ▶ Artificial Intelligence,
 - ▶ Computer Science.

Emphasis on large scale implementations and applications: **algorithm centered**

- ▶ *Statistical learning* arose as a subfield of
 - ▶ Statistics,
 - ▶ Applied Maths,
 - ▶ Signal Processing, ...

Emphasizes models and their interpretability: **model centered**

Definitions of Learning

Machine Learning in Computer Science

Tom Mitchell (The Discipline of Machine Learning, 2006)

A computer program CP 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

Key points

- ▶ Experience E: **data and statistics**
- ▶ Performance measure P: **optimization**
- ▶ tasks T: utility
 - ▶ automatic translation
 - ▶ playing Go
 - ▶ ... doing what human does

Definitions of Learning

Machine Learning in Computer Science

Tom Mitchell (The Discipline of Machine Learning, 2006)

A computer program CP 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

Key points

- ▶ Experience E: **data and statistics**
- ▶ Performance measure P: **optimization**
- ▶ tasks T: utility
 - ▶ automatic translation
 - ▶ playing Go
 - ▶ ... doing what human does

Definitions of Learning

Machine Learning in Computer Science

Tom Mitchell (The Discipline of Machine Learning, 2006)

A computer program CP 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

Key points

- ▶ Experience E: **data and statistics**
- ▶ Performance measure P: **optimization**
- ▶ tasks T: utility
 - ▶ automatic translation
 - ▶ playing Go
 - ▶ ... doing what human does

Definitions of Learning

Machine Learning in Computer Science

Tom Mitchell (The Discipline of Machine Learning, 2006)

A computer program CP 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

Key points

- ▶ Experience E: **data and statistics**
- ▶ Performance measure P: **optimization**
- ▶ tasks T: utility
 - ▶ automatic translation
 - ▶ playing Go
 - ▶ ... doing what human does

Experience E: the data!

Type of data: qualitatives / ordinales / quantitatives variables

- ▶ Text: strings
- ▶ Speech: time series
- ▶ Images/videos: 2/3d dependences
- ▶ Networks: graphs
- ▶ Games: interaction sequences
- ▶ ...

Big data (volume, velocity, variety, veracity)

Data are available without having decided to collect them!

- ▶ importance of preprocessings (cleaning up, normalization, coding,...)
- ▶ importance of a good representation : from raw data to vectors

Objective and performance measures P

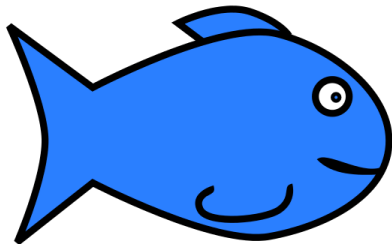
Generalize

- ▶ Perform well (minimize P) on **new data** (fresh data, i.e. unseen during learning)
- 👉 Derive good (P/error rate) prediction functions

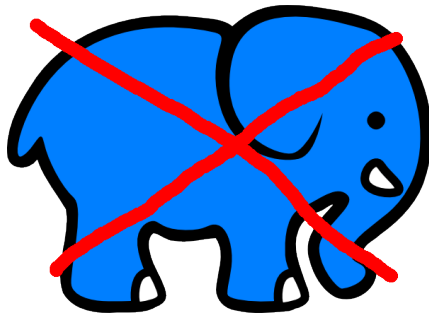
Objective and performance measures P

Generalize

- ▶ Perform well (minimize P) on **new data** (fresh data, i.e. unseen during learning)
- 👉 Derive good (P/error rate) prediction functions






A fish








A fish

References

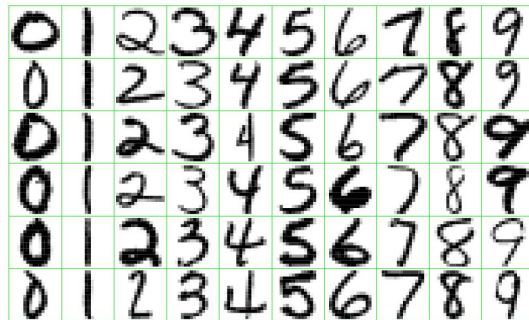
Reference books

-  Trevor Hastie, Robert Tibshirani et Jerome Friedman (2009), The Elements of Statistical Learning (2nd Edition), *Springer Series in Statistics*
-  Christopher M. Bishop (2007), Pattern Recognition and Machine Learning, *Springer*
-  Kevin P. Murphy (2012), Machine Learning: a Probabilistic Perspective, *MIT press*

Supplementary materials, datasets, online courses, ...

-  <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
-  <https://www.cs.ubc.ca/~murphyk/MLbook/>
-  <https://www.coursera.org/course/ml> *very popular MOOC (Andrew Ng)*
-  <https://work.caltech.edu/telecourse.html> *more involved MOOC (Y. Abu-Mostafa)*
-  https://scikit-learn.org/stable/auto_examples/index.html *Examples from the sklearn library*

Recognition of handwritten digits (US postal envelopes)



- ☞ Predict the class (0,...,9) of each sample from an image of 16×16 pixels, with a pixel intensity coded from 0 to 255
- Low error rate to avoid wrong allocations of mails!

Supervised classification

Spams Recognition

Spam

WINNING NOTIFICATION

We are pleased to inform you of the result of the Lottery Winners International programs held on the 30th january 2005.
[...] You have been approved for a lump sum pay out of 175,000.00 euros.
CONGRATULATIONS!!!

No Spam

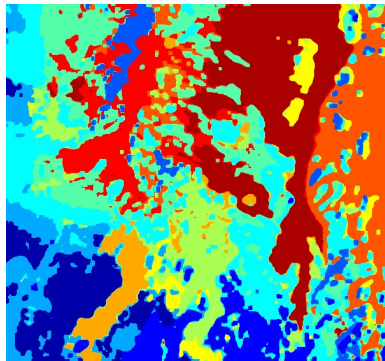
Dear George,
Could you please send me the report #1248 on the project advancement?
Thanks in advance.

Regards,
Cathia

- 👉 Define a model to predict whether an email is spam or not
- ▶ Low error rate to avoid deleting useful messages, or filling the mailbox with useless emails

supervised classification

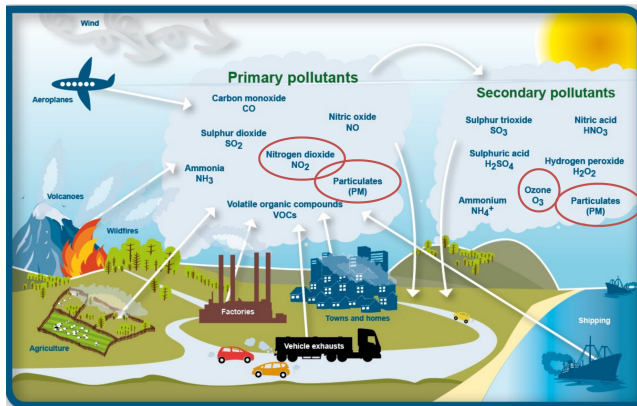
Recognition of Hekla Volcano landscape, Iceland



- ✎ Predict the class of landscape $\in \{ \text{Lava 1970, Lava 1980 I, Lava 1980 II, Lava 1991 I, Lava 1991 II, Lava moss cover, hyaloclastite formation, Tephra lava, Rhyolite, Scoria, Firn-glacier ice, Snow} \}$ from digital remote sensing images

supervised or unsupervised classification

Prediction of pollutant concentrations



- 👉 Predict pollutant concentrations (O_3 , NO_2 , PM_{10} , $PM_{2.5}$) at time $D_0+1, +2, +3$ from hourly measures timeseries + weather data + chemistry based forecasting models

supervised regression/classification (pollution alert or not)

Definitions

Variable terminology

- ▶ Observed data referred to as *input* variables, *predictors* or *features*: X
- ▶ Data to predict referred to as *output* variables, or *responses*: Y

Type of prediction problem: regression vs classification

Depending on the type of the *output* variables

- ▶ When Y are **quantitative** data (e.g. O3 concentration values): **regression**
- ▶ When Y are **categorical** data (e.g. handwritten digits $Y \in \{0, \dots, 9\}$): **classification**

Two very close problems

Prediction problem

Assumptions

- ▶ Input variables X_i are vectors in \mathbb{R}^p :

$$X_i = (X_{i,1}, \dots, X_{i,p})^T \in \mathcal{X} \subset \mathbb{R}^p$$

- ▶ Output variables Y_i take values:
 - ▶ In $\mathcal{Y} \subset \mathbb{R}$ (regression)
 - ▶ In a finite set \mathcal{Y} (classification)
- ▶ $Y = f(X) + \epsilon$

Prediction rule

Function of prediction / rule of classification \equiv function $\hat{f}: \mathcal{X} \rightarrow \mathcal{Y}$ to get predictions of new elements Y given X

$$\hat{Y} = \hat{f}(X)$$

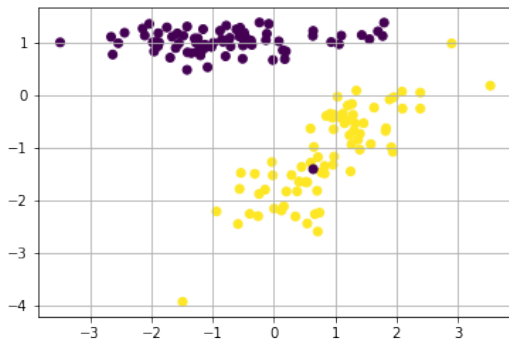
Supervised or unsupervised learning

Training set \equiv available sample \mathcal{T} to learn the prediction rule f

For a sized n training set, different cases:

- ▶ **Supervised learning**: $\mathcal{T} \equiv \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are available
- ▶ **Unsupervised learning**: $\mathcal{T} \equiv (X_1, \dots, X_n)$ are available only
- ▶ **Semi-supervised**: mixed scenario (often encountered in practice, but less information than in the supervised case)

Binary classification



Simple linear model for classification

We seek a prediction model based on the linear regression of the outputs $Y \in \{-1, 1\}$:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where $\beta = (\beta_1, \beta_2)^T$ is a 2D unknown parameter vector

Learning problem \Leftrightarrow Estimation of β

Least Squares Estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T$: minimize the training error rate (quadratic cost sense)

$$RSS(\beta) = \sum_{i=1}^N (Y_i - \beta_1 X_{i,1} - \beta_2 X_{i,2})^2$$

Classification rule based on least squares regression

$$f(X) = \begin{cases} 1 & \text{if } \hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \geq 0, \\ -1 & \text{otherwise} \end{cases}$$

Notebook

Model complexity

Most of methods have a complexity related to their *effective* number of parameters

Linear classification: model order p

E.g. d th degree polynomial regression: $p = d + 1$ parameters a_k s.t.

$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_d x^d + \epsilon, \\ &= \mathbf{X}_d \boldsymbol{\beta}_d + \epsilon, \end{aligned}$$

where

$$\begin{aligned} \mathbf{X}_d &= \begin{bmatrix} 1, & x, & x^2, & \dots, & x^d \end{bmatrix}, \\ \boldsymbol{\beta}_d &= [\beta_0, \beta_1, \beta_2, \dots, \beta_d]^T. \end{aligned}$$

Notebook

Test error vs Train Error

Error rate vs polynomial order d

Notebook

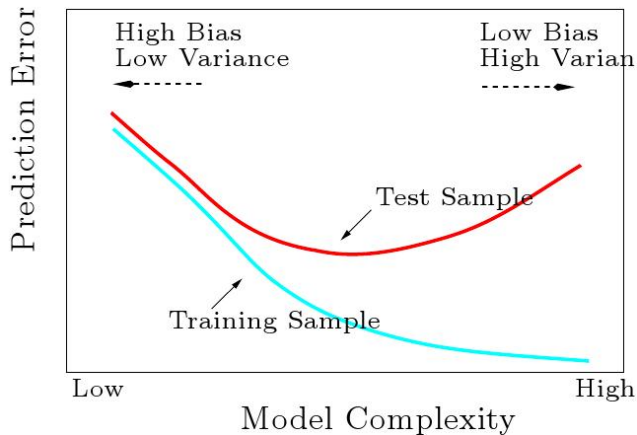
- ▶ Training error rate (i.e. error rate for train data used for learning) minimized when $d = 19$
- ▶ True error rate (i.e. error rate for test data not used for learning) minimized when $d = 5 \dots$

👉 Training error always decrease with the model complexity. **Can't use alone to select the model!**

Model Selection

Fundamental trade-off

- ▶ Too simple model (high bias) → **under-fitting**
- ▶ Too complex model (high variance) → **over-fitting**



Fundamental Bias-Variance trade-off

If the true model is

$$Y = f(X) + \epsilon,$$

then for any prediction rule $\hat{f}(X)$, Mean Squared Error (MSE) expresses as

$$E \left[\left(Y - \hat{f}(x) \right)^2 \right] = \text{Var} \left[\hat{f}(x) \right] + \text{Bias} \left[\hat{f}(x) \right]^2 + \text{Var} [\epsilon]$$

- ▶ $\text{Var} [\epsilon]$ is the *irreducible* part
 - ▶ as the flexibility of $\hat{f} \nearrow$, its variance \nearrow and the bias \searrow
- 👉 overfitting/underfitting trade-off