Machine Learning (MAL)

Mathilde Mougeot

ENSIIE

MAL 2019

The "Data" phenomena

1. Data tsunami... Todays, data are everywhere.
   - Finance. Transactions data
   - Digital revolution in the Industry. Production data (Supply chain). physical data (Temperature, IR sensors)
   - Marketing/ consumption data. "Click" data
   - On your phone (GPS, mail, musique ...)

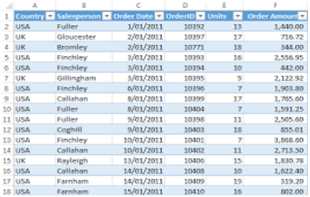
The "Data" phenomena

1. Data tsunami... Todays, data are everywhere.
   - Finance. Transactions data
   - Digital revolution in the Industry. Production data (Supply chain). physical data (Temperature, IR sensors)
   - Marketing/ consumption data. "Click" data
   - On your phone (GPS, mail, musique ...)
2. Data zoology ... A large variety of data, well or no structured.
   - quantitative, qualitative, binary
   - synchronous, asynchronous, event data
   - ... image data, text data, speech data

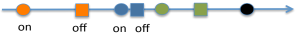
The "Data" phenomena

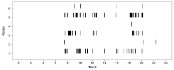
1. Data tsunami... Todays, data are everywhere.
   - Finance. Transactions data
   - Digital revolution in the Industry. Production data (Supply chain). physical data (Temperature, IR sensors)
   - Marketing/ consumption data. "Click" data
   - On your phone (GPS, mail, musique ...)
2. Data zoology ... A large variety of data, well or no structured.
   - quantitative, qualitative, binary
   - synchronous, asynchronous, event data
   - ... image data, text data, speech data
3. Data base, data lakes available.
   From small data set to Big data set

Several data sources



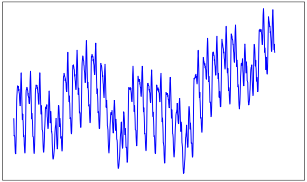
	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amoun
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00



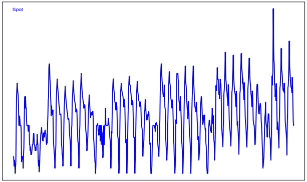


automated data mining survey
responses co ter transcripts
qualatativ oot cause
classificati text analysis insights
ad-hoc an s product
reviews se t vo the
customer dashboards consume
trends ad-hoc analysis early warning

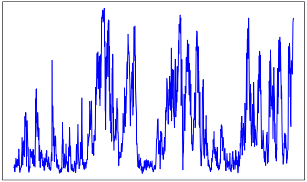
Illustration. Several energy data



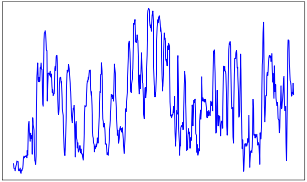
French electrical consumption



Energy Spot prices

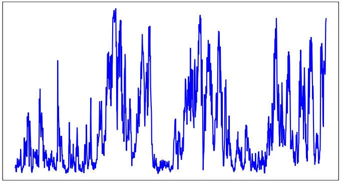


Wind turbine power

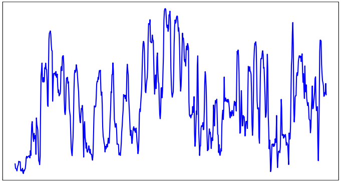


Industrial equipment

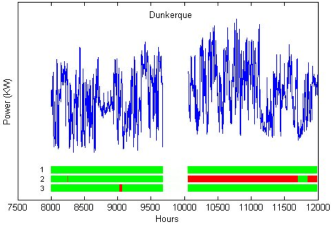
Potential questions on energy data



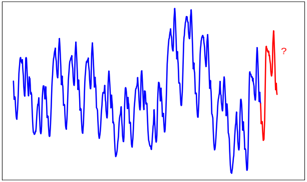
Exploratory analysis



Virtual sensor



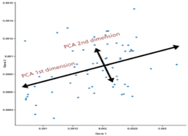
Monitoring



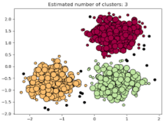
Forecasting electrical consumption

Machine Learning. Statistical settings (1/3) :

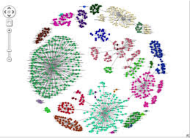
Unsupervised learning. (Inputs X) : $X \in \mathcal{X}$ $(\mathbb{R}^p)$



Exploratory Analysis.



Clustering



Graph Analysis

Machine Learning settings (2/3)

- Supervised regression learning (Y, X) :

$$Y \in \mathbb{R},\ X \in \mathcal{X}\ (\mathbb{R}^p) \qquad \boxed{Y = \mathcal{M}_{\text{data set}}(X) + \epsilon}$$

Machine Learning settings (2/3)

- Supervised regression learning (Y, X) :

  $Y \in \mathbb{R}$, $X \in \mathcal{X}$ $(\mathbb{R}^p)$ $\boxed{Y = \mathcal{M}_{\text{data set}}(X) + \epsilon}$

  - Example. Parametric models.
    $\mathcal{M}(X) : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$
    → Probabilistic or not, depending on the law assumption on ϵ.

    Data Set : n observations : (y_i, x_i)
    → to estimate (to compute) the parameters $\hat{\beta}$ of the model

Machine Learning settings (2/3)

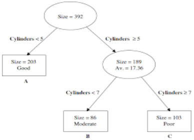
- Supervised regression learning (Y, X) :

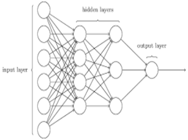
  $Y \in \mathbb{R}$, $X \in \mathcal{X}$ $(\mathbb{R}^p)$ $\boxed{Y = \mathcal{M}_{\text{data set}}(X) + \epsilon}$

  - Example. Parametric models.
    $\mathcal{M}(X) : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$
    → Probabilistic or not, depending on the law assumption on ϵ.

    Data Set : n observations : (y_i, x_i)
    → to estimate (to compute) the parameters $\hat{\beta}$ of the model

  - Non Parametric models No analytical (complex) expression, infinity of parameters





→ Models : Kernels, decision trees, neural networks,...

Crucial question : How to find the correct model ?

Machine Learning settings (3/3)

- Supervised Classification learning (Y, X) :

  $Y \in G_1, ..., G_K,\ X \in \mathcal{X}\ (\mathbb{R}^p)$ $\qquad$ $\boxed{Y = \mathcal{M}_{\text{data set}}(X)}$

Machine Learning settings (3/3)

- Supervised Classification learning (Y, X) :

  $Y \in G_1, ..., G_K$, $X \in \mathcal{X}$ $(\mathbb{R}^p)$ $\boxed{Y = \mathcal{M}_{\text{data set}}(X)}$

  - Example. Parametric model. Logistic regression

    $\mathcal{M}(X) : P(Y = G_k/X = x) = \frac{1+e^{x\beta}}{1-e^{x\beta}}$

    Data Set : n observations : (y_i, x_i)

    → to estimate (to compute) the parameters $\hat{\beta}$ of the model

Machine Learning settings (3/3)

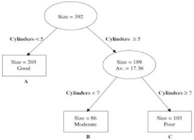
- Supervised Classification learning (Y, X) :

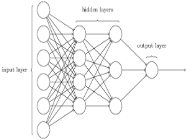
$$Y \in G_1, \ldots, G_K,\ X \in \mathcal{X}\ (\mathbb{R}^p) \qquad \boxed{Y = \mathcal{M}_{\text{data set}}(X)}$$

- Example. Parametric model. Logistic regression
$\mathcal{M}(X) : P(Y = G_k/X = x) = \frac{1+e^{x\beta}}{1-e^{x\beta}}$

  Data Set : n observations : (y_i, x_i)
  $\rightarrow$ to estimate (to compute) the parameters $\hat{\beta}$ of the model

- Non Parametric models No analytical (complex) expression, infinity of parameters



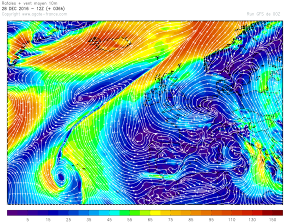


$\rightarrow$ Models : Kernels, decision trees, neural networks,...

Crucial question : How to find the correct classification model ?

Machine Learning or Statistical Modeling vs Physical, simulation models

- Physical Modelling mosly based on physical equations :
  - Explicit equation
    $PV = nRT...$
  - Simulation models. Partial differential equations,...
    $\rightarrow$ Need of a numerical model to study the dynamic and the evolution of a model. Ex : Navier-Stokes equation



Outline

Machine learning. The 2019 choice...

1. Supervised setting. Classification.
   - Parametric models. Bayes model. LDA. QDA
   - Performance criteria.
   - Non Parametric models. Classification trees.
   - Ensemble methods.
     Bagging. Random Forest. Boosting. Stacking.
2. Supervised setting. Regression.
   - Non Parametric models. Regression trees.
   - Ensemble methods.
     Bagging. Random Forest. Boosting. Stacking.
3. Unsupervised setting. Clustering.
   - K-means.
   - Spectral clustering.