Statistical Learning for DoE ?

B. Ghattas

Université d'Aix-Marseille

badihghattas@gmail.com

Outline

Sommaire

Airquality dataset

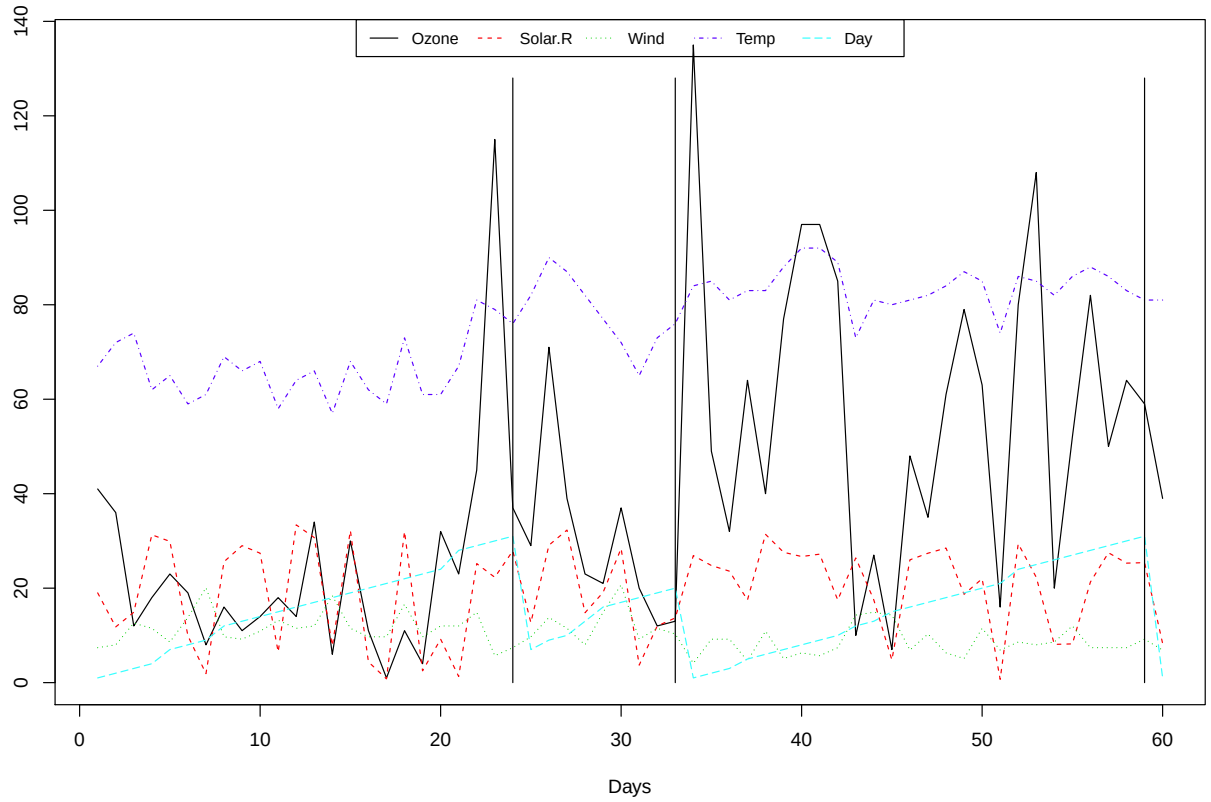
A data frame with 154 observations on 6 variables. Daily readings of the following air quality values for May 1, 1973 to September 30, 1973.

airquality.tex	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5			14.3	56	5	5
6	28		14.9	66	5	6

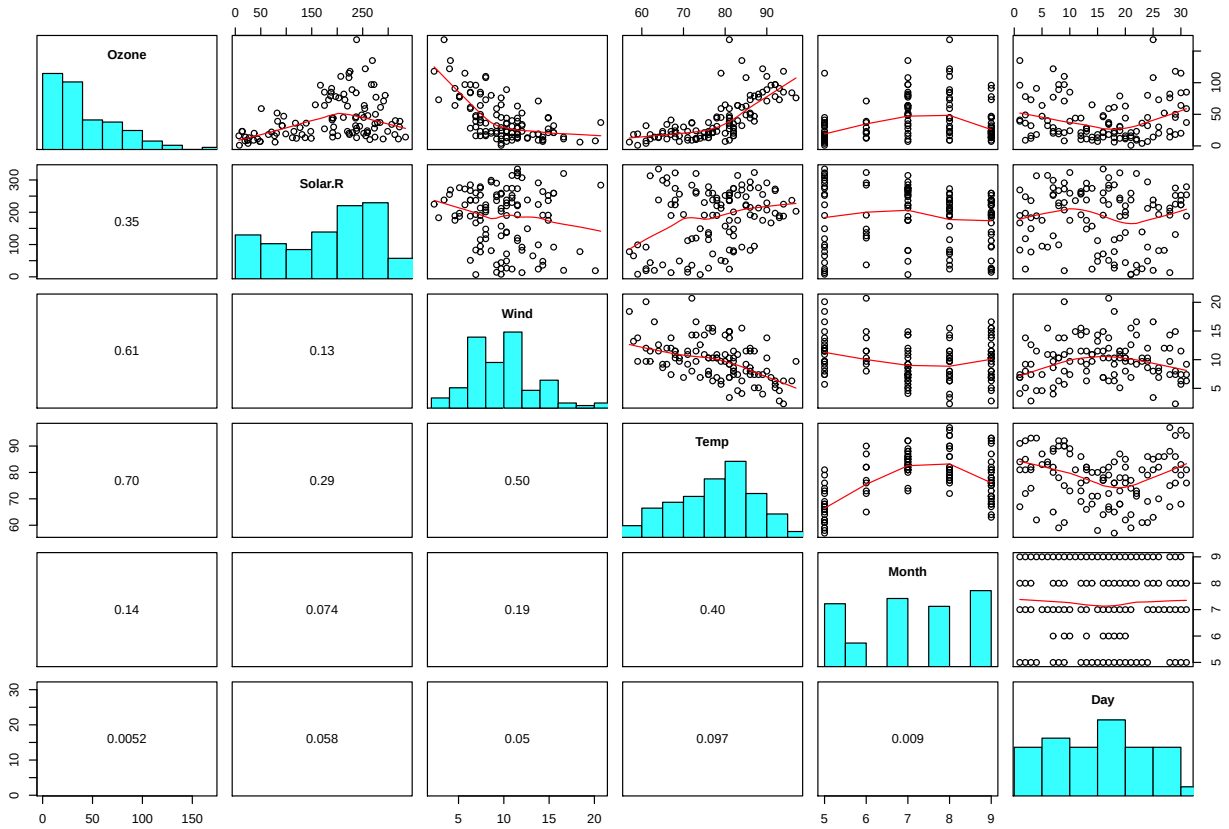
- *Ozone* Mean Ozone (ppb) from 1300 to 1500 hours at Roosevelt Island.
- *Solar.R* Solar radiation in Langleys from 0800 to 1200 hours at Central Park.
- *Wind* Average wind speed in miles per hour at 0700 and 1000 hours at La Guardia Airport.
- *Temp* Maximum daily temperature (degrees F) at La Guardia Airport.
- *Month* Month (1–12)
- *Day* Day of month (1–31)

Objective : Predict Ozone concentration (target variable) using the other variables.

Airquality Series



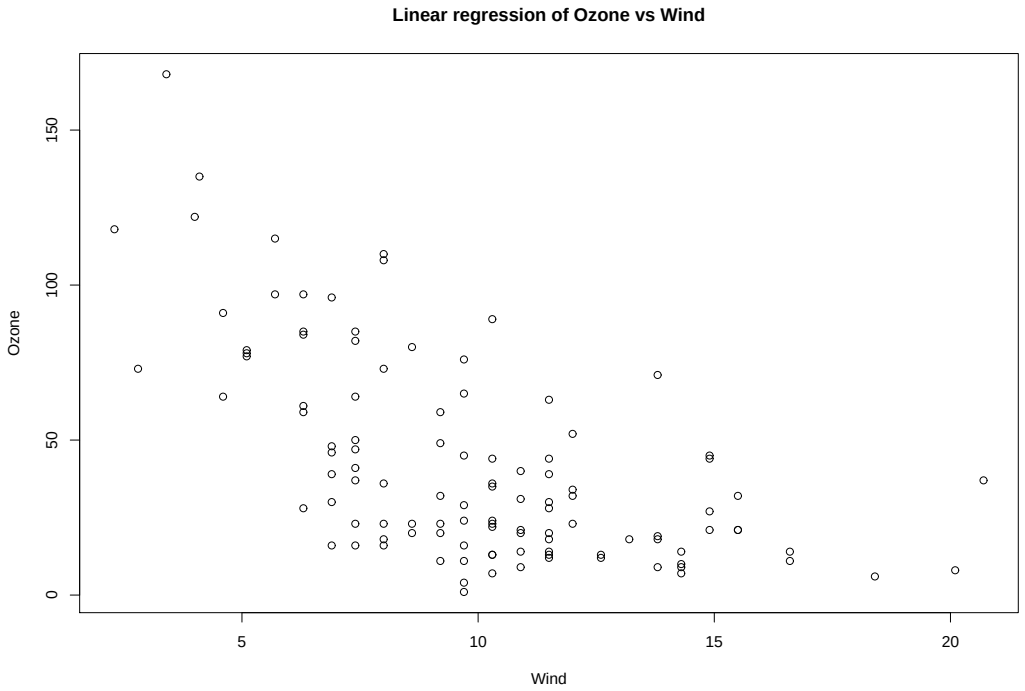
Airquality pairs



Airquality - summary statistics

Ozone	Solar.R	Wind	Temp	Month	Day
Min. : 1.00	Min. : 7.0	Min. : 1.700	Min. :56.00	Min. :5.000	Min. : 1.0
1st Qu. : 18.00	1st Qu. :115.8	1st Qu. : 7.400	1st Qu. :72.00	1st Qu. :6.000	1st Qu. : 8.0
Median : 31.50	Median :205.0	Median : 9.700	Median :79.00	Median :7.000	Median :16.0
Mean : 42.13	Mean :185.9	Mean : 9.958	Mean :77.88	Mean :6.993	Mean :15.8
3rd Qu. : 63.25	3rd Qu. :258.8	3rd Qu. :11.500	3rd Qu. :85.00	3rd Qu. :8.000	3rd Qu. :23.0
Max. :168.00	Max. :334.0	Max. :20.700	Max. :97.00	Max. :9.000	Max. :31.0
NA's :37	NA's :7				

Airquality : Linear Regression



We assume that the relation between *Wind* and *Ozone* has the following form :

$$Ozone_i = \beta_0 + \beta_1 * Wind_i + \epsilon_i$$

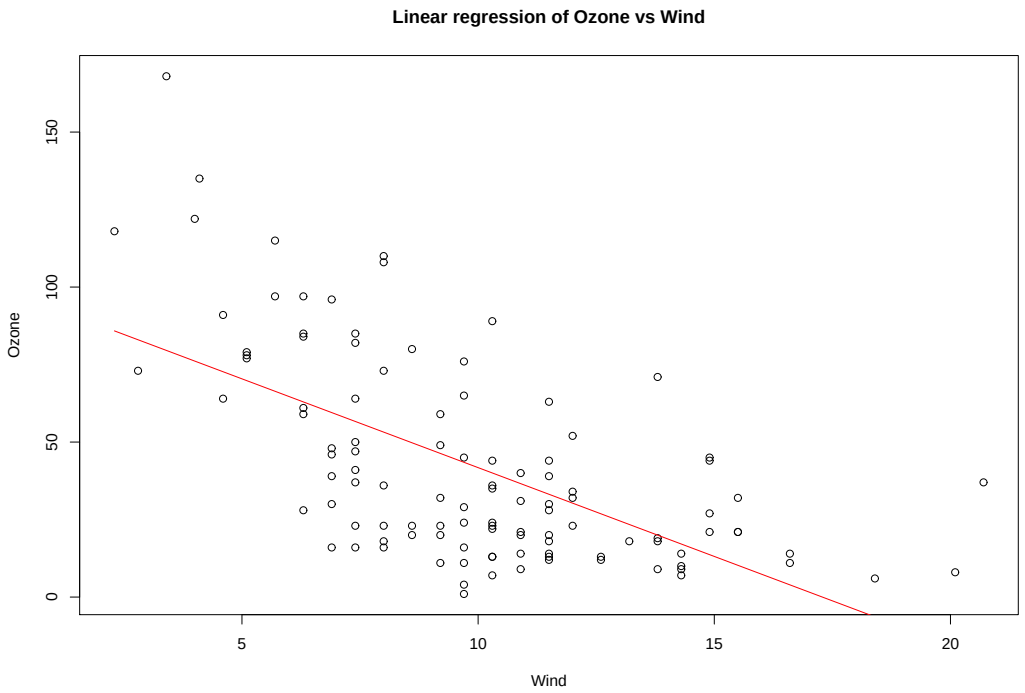
The values of the coefficients may be obtained by minimizing the Mean Squared Error (MSE) :

$$\sum_{i=1}^{n} (Ozone_i - \beta_0 - \beta_1 * Wind_i)^2$$

We get :

$$Ozone_i = 99.041 - 5.729 * Wind_i + \epsilon_i$$

Airquality : Linear Regression



We assume that the relation between *Wind* and *Ozone* has the following form :

$$Ozone_i = \beta_0 + \beta_1 * Wind_i + \epsilon_i$$

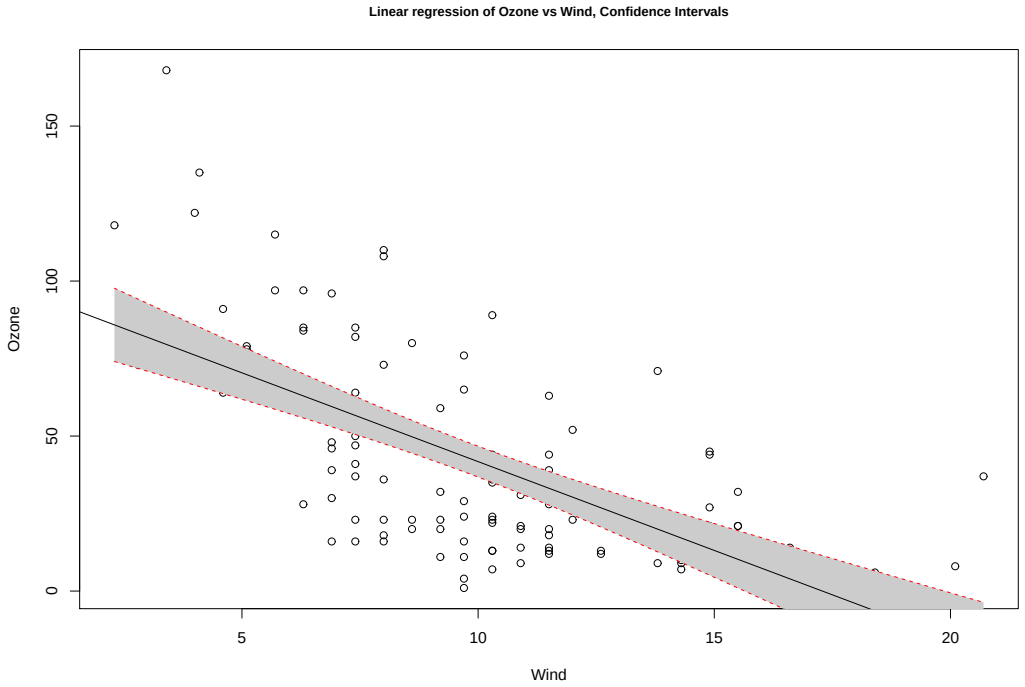
The values of the coefficients may be obtained by minimizing the Mean Squared Error (MSE) :

$$\sum_{i=1}^{n} (Ozone_i - \beta_0 - \beta_1 * Wind_i)^2$$

We get :

$$Ozone_i = 99.041 - 5.729 * Wind_i + \epsilon_i$$

Airquality : Linear Regression



We assume that the relation between *Wind* and *Ozone* has the following form :

$$Ozone_i = \beta_0 + \beta_1 * Wind_i + \epsilon_i$$

The values of the coefficients may be obtained by minimizing the Mean Squared Error (MSE) :

$$\sum_{i=1}^{n} (Ozone_i - (\beta_0 + \beta_1 * Wind_i))^2$$

We get :

$$Ozone_i = 99.041 - 5.729 * Wind_i + \epsilon_i$$

Linear Model, multiple input

$$\widehat{Y} = \widehat{\beta}_0 + \sum_{j=1}^{p} X_j \widehat{\beta}_j$$

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i' \beta)^2$$

The minimum is achieved with :

$$\widehat{\beta} = (X'X)^{-1} X'y$$

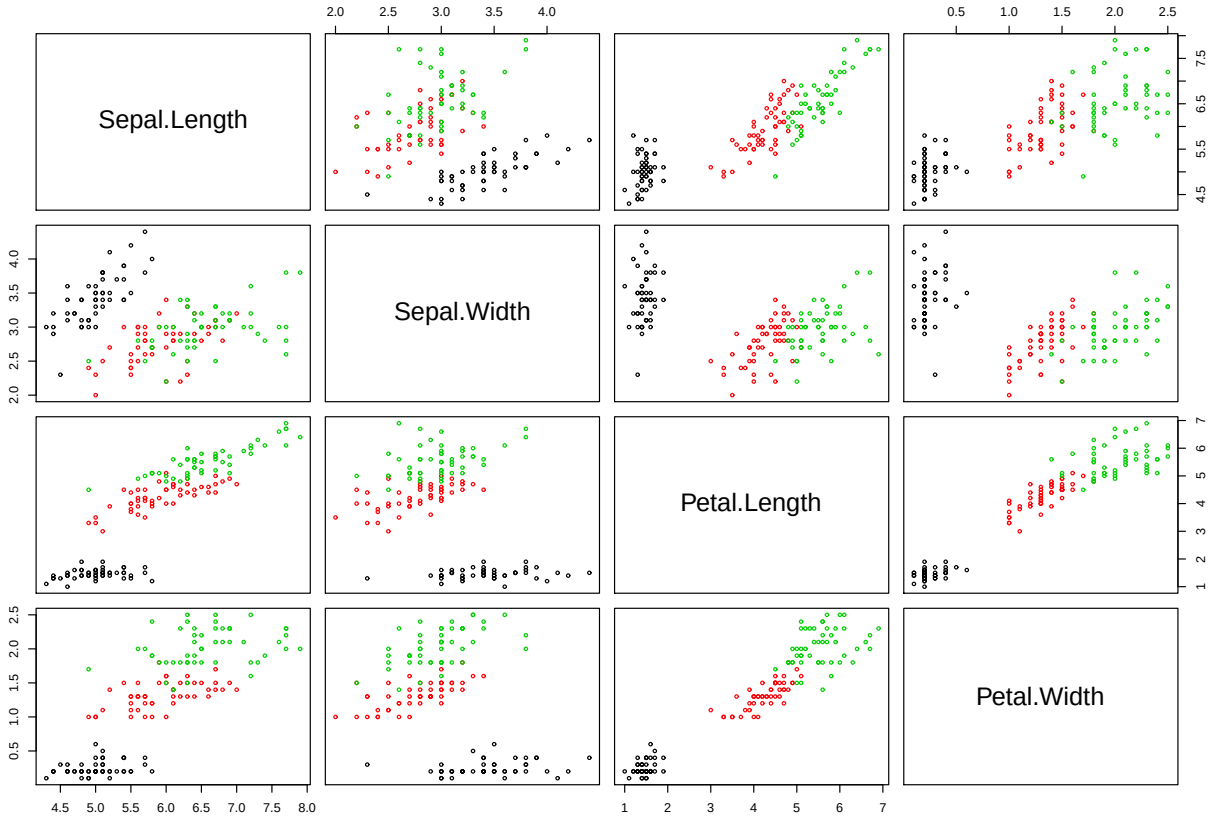
Advantages : stable, does not need a lot of data, nice asymptotic properties for estimators... Drawbacks ; linear, parametric, sensitive to extremes and to correlations, problems at the boundaries, ...

A classification example ; Iris

The Iris data set contains 4 explanatory variables "*Sepal.Length*", "*Sepal.Width*", "*Petal.Length*", "*Petal.Width*", "*Species*", and one target variable *Species* taking one of the three values :*Setosa*, *Virginica*, *Versicolor*.

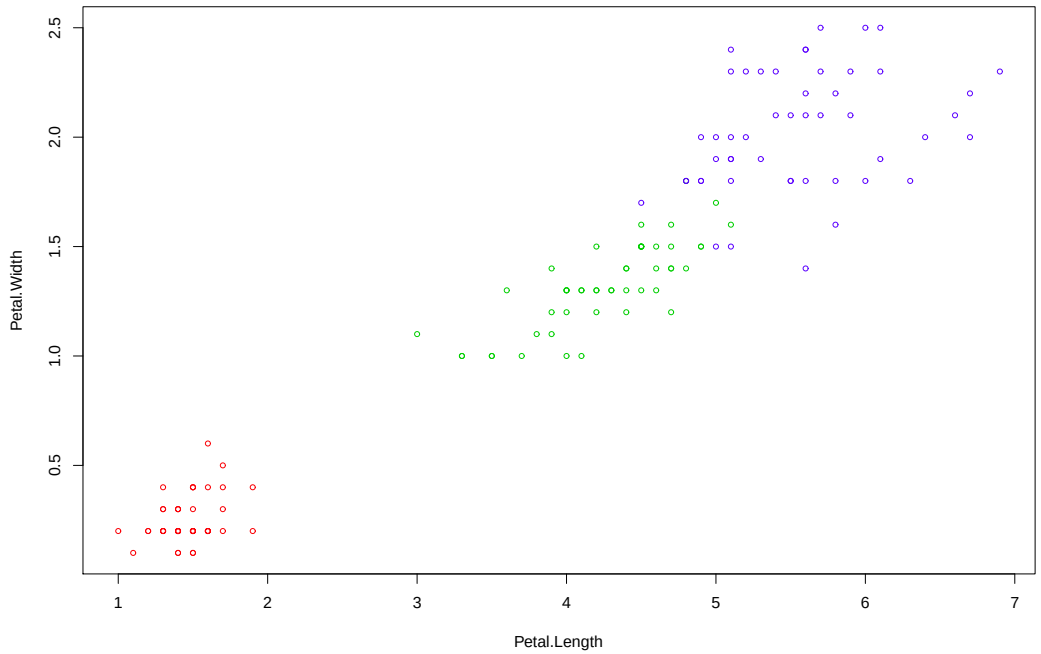
iris	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Iris, Scatterplots



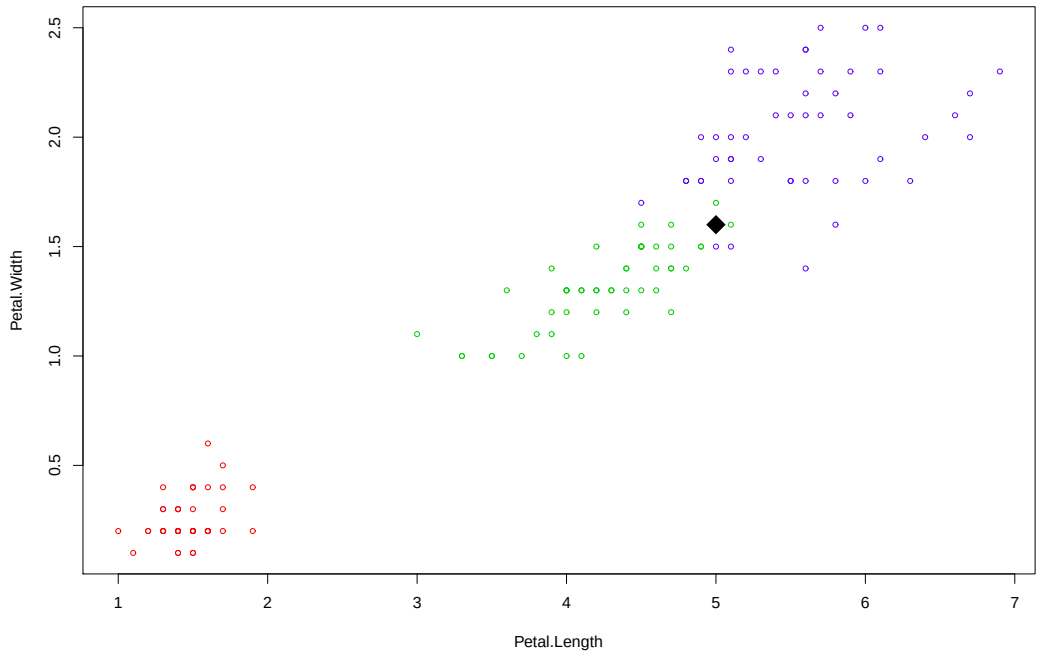
K nearest neighbors

Considering only two of the explanatory variables together with the target we may look at the data in two dimensions. Each flower is a point in R^2 having a label Y, each label presented by a color.



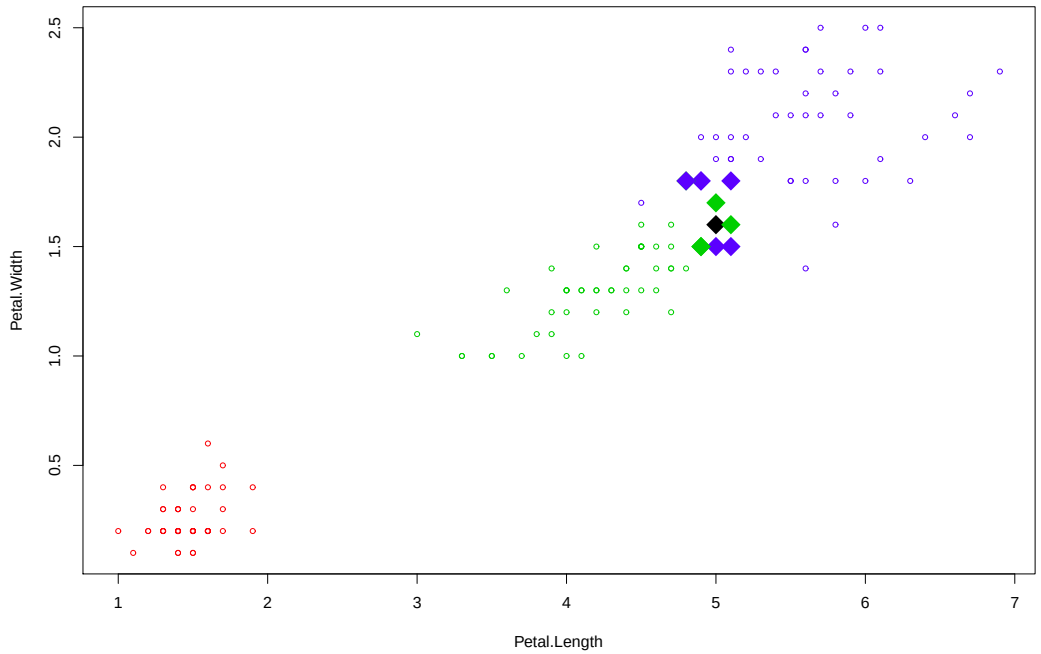
K nearest neighbors

Considering only two of the explanatory variables together with the target we may look at the data in two dimensions. Each flower is a point in R^2 having a label Y, each label presented by a color.



K nearest neighbors

Considering only two of the explanatory variables together with the target we may look at the data in two dimensions. Each flower is a point in R^2 having a label Y, each label presented by a color.



Extensions

- Classification : $Y \in \{0, 1\}$, $Y \in \{1, .., M\}$.
- Regression : $Y \in R$.
- $Y \in R^q$, example : spatial modeling
- $Y \in L^2(R)$, example : curves or signals.
- $X \in R^p$, when $p >> n$ like in genomics.
- $X \in L^2(R)$ as for signals.

Sommaire

Statistical Learning

Supervised

Classification, Regression.

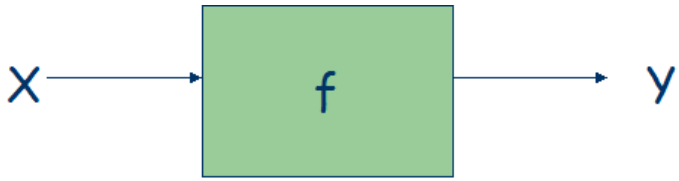
Unsupervised

Clustering, Density estimation.

Other types of Learning

Semi-supervised, Reenforcement, Active Learning...

Notations



We wish to estimate f using the dataset at hand

$$D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

We must choose f within a class of functions, with unknown parameters.

For example : $y = f(x) = a_0 + a_1 x_1 + a_2 x_2 + a_p x_p$

$$D_n \Longrightarrow f_n(X, D_n)$$

Terminology

- X : **Input**, explanatory, "predictor"
- Y : **Output**, "target", "outcome"
- Both variables may be quantitative, discrete (factor), ordered or not, binary or multi-class, and multidimensionalm

Regression if Y is continuous.

Classification if not.

Examples

Several scientific domains, biology environment finance, industry,...

- Predict whether a patient will repeat a heart attack
- Predict stock prices within 6 months using economic parameters and the performance of an enterprise
- Identify handwritten digits of postal codes on envelops
- Estimate the glucose level of a diabetic using the infrared spectra of blood absorption.
- Identify the prostate cancer risk using clinical and demographical parameters.

Learning is essential in statistics, Data Mining and Artificial Intelligence.
We Learn from data ! !

Regression criterion

$$X \in R^p, \quad Y \in R$$

We look for a function f to predict Y, using the entry X. We must define a risk function, $L(Y, f(X))$, and choose the function f which minimizes the risk.
For regression the quadratic risk is often used the criterion minimized using the sample at hand is :

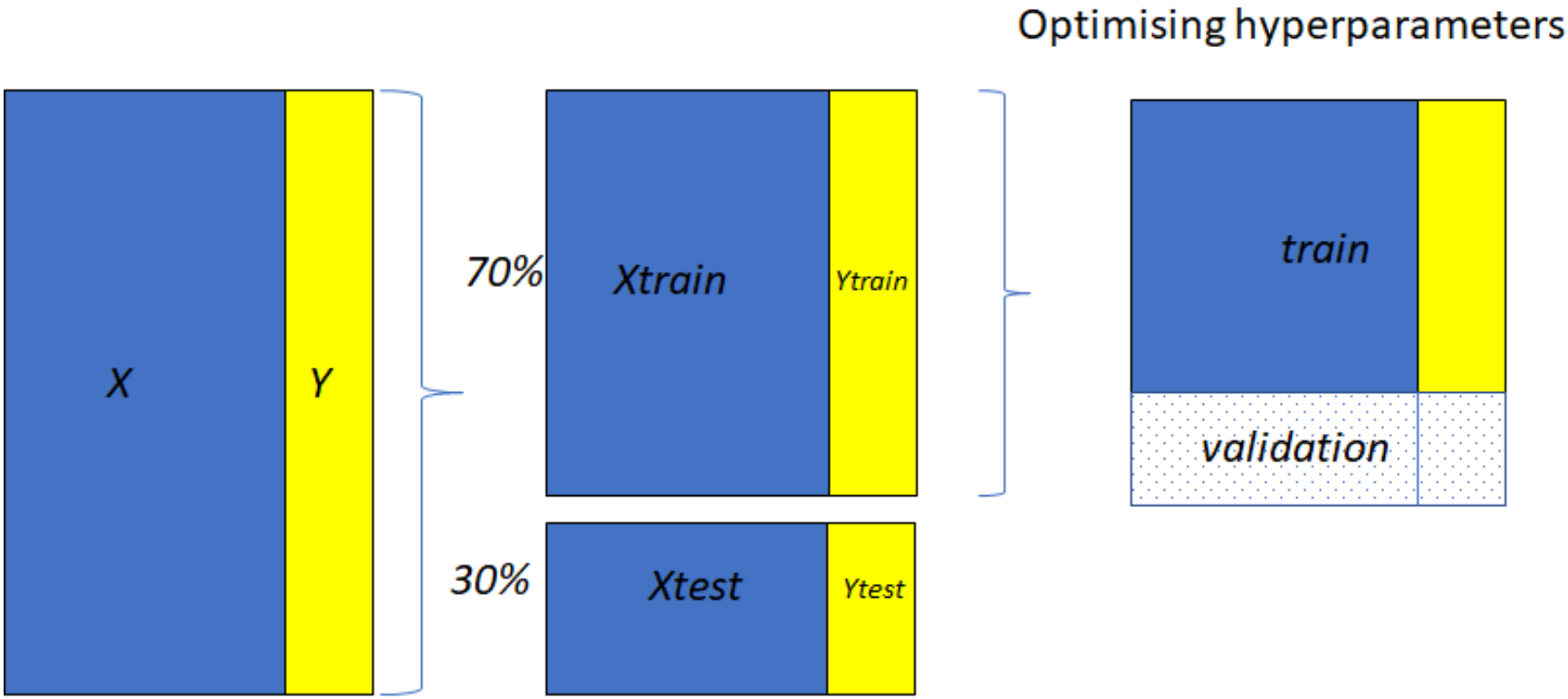
$$MSE(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

Choosing the shape or class of f

Choose f, within a class C, which minimizes $\widehat{L}_n$.
The performance of the classifier are guaranteed to be close to the best classifier of the class if the complexity of C is controlled.
the Choice of C depends on :

- The nature of the problem we model
- Hypotheses and experiences issuing the data
- Experts opinion
- A question of fashion ?

Train/test/validation



Pre-processing

- Filling missing values
- Detecting and suppressing outliers
- Transforming some or all variables (scaling, standardizing, discretizing,...)
- Encoding some variables (one hot)

Some Methods

- K Nearest Neighbours
- CART
- Aggregating classifiers
- Neural Networks
- Support Vector Machines
- Bayesian Networks

Sommaire

The problem

data : $(\mathbf{X}, Y) \in R^p \times C$

$\mathbf{X}$ predictor, attributes, features

$Y \in C$ output to predict. $C = R$ or $C = \{1, ..., J\}$.

Objective :

Using the observations (X_i, Y_i) from D, construct a classifier $\widehat{f}(\mathbf{X})$ having a low generalization error :

$$R(\widehat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(L(y_i, \widehat{f}(\mathbf{x_i}) \right)$$

where L is a loss function.

L is the quadratic error in regression, or the misclassification rate in classification.

The model

Search for a partition of the space X and assign a value of Y to each class of the partition.
In regression :

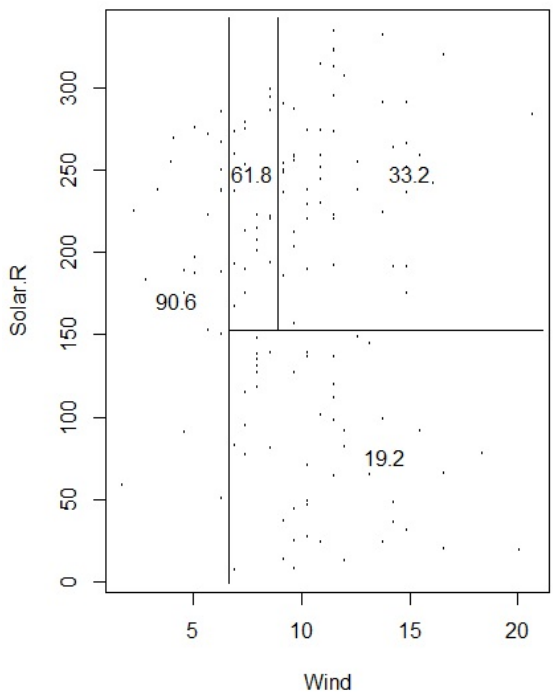
$$f(\mathbf{x}) = \sum_{j=1}^{q} c_j \mathbf{1}_{N_j}(\mathbf{x})$$

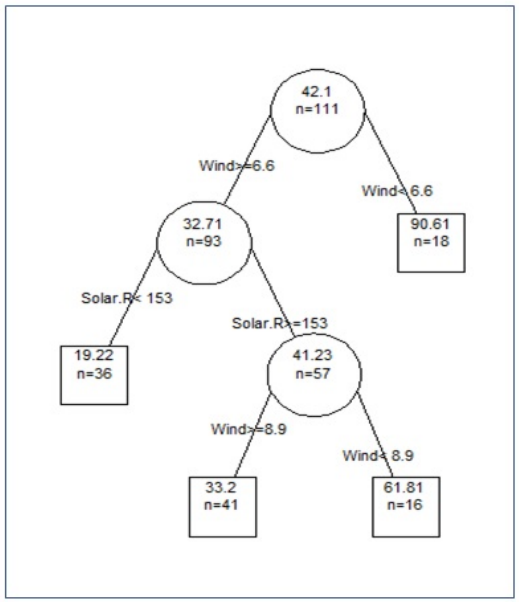
$$\widehat{c_j} = \frac{1}{Card\{i; \mathbf{x}_i \in N_j\}} \sum_{i;\mathbf{x}_i \in N_j} Y_i$$

In Classification : Y discrete having J levels

$$\widehat{c_j} = \text{The most frequent class in } N_j(\mathbf{x})$$

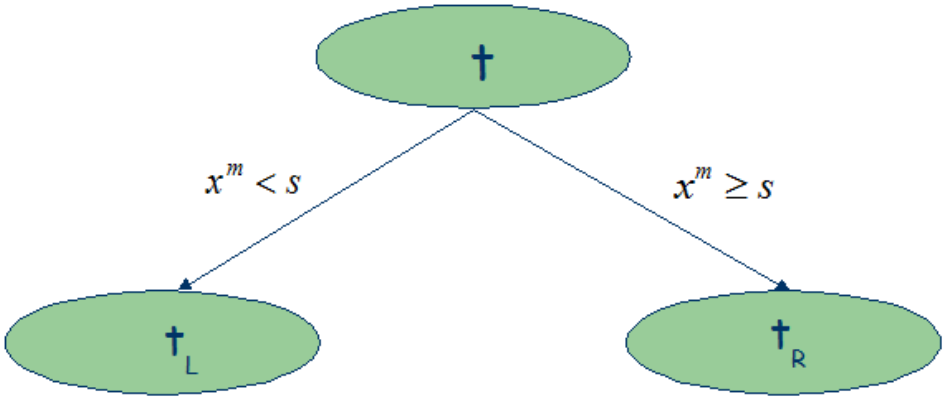
Example, Airquality





2 stages : Maximal Tree and Pruning

All the observations are in the root node.



Splitting rule : one variable and a threshold. How to do ?
Use the deviance to measure the heterogeneity of a node :

$$R(t) = \sum_{x_n \in t} (y_n - \bar{y}(t))^2$$

Optimal Splits : minimize the children's deviance

Minimize total new nodes Heterogeneity. Let s be a split of the form : $x^m < a$,

$$\Delta R(s,t) = R(t) - (R(t_L) + R(t_R)) \geq 0$$

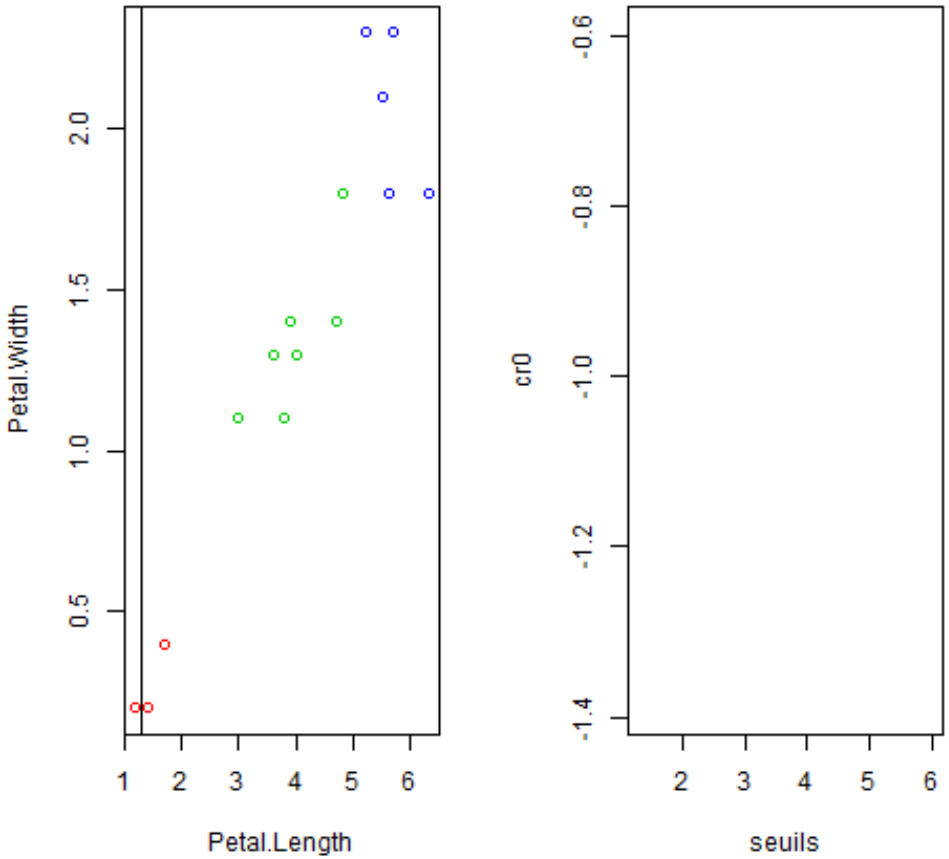
$$\Delta R(s,t) = max_{s \in \Sigma} \Delta R(s,t)$$

In classification,

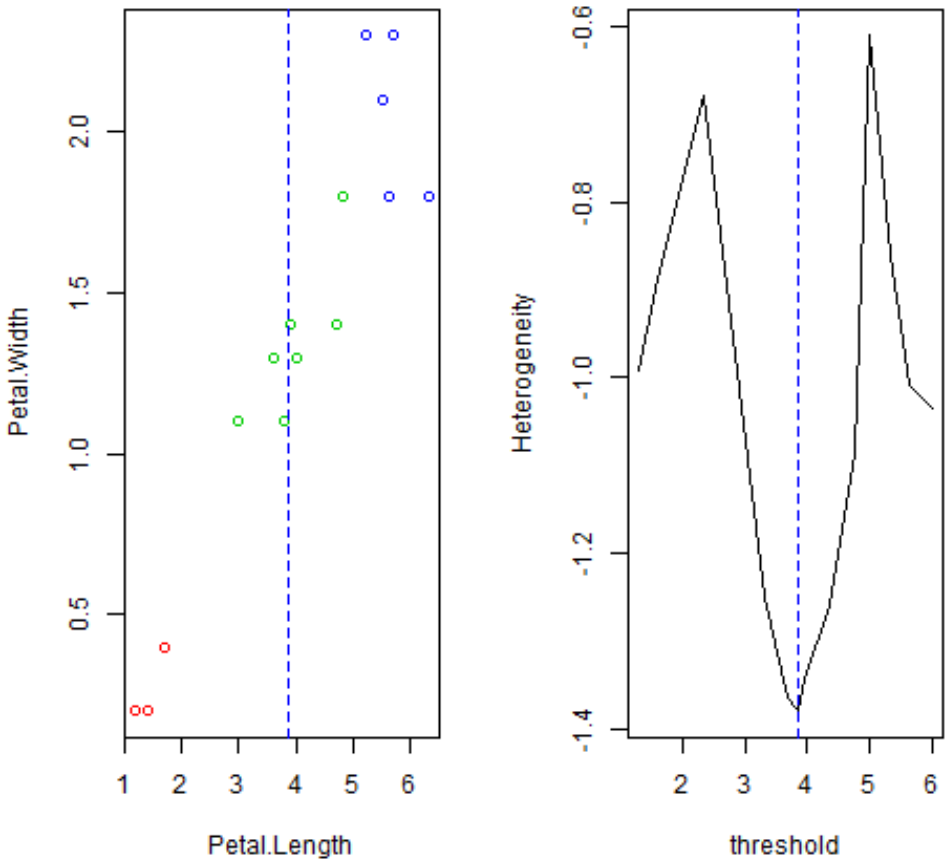
$$R(t) = -\sum_{j \in J} p_j(t) log(p_j(t))$$

where $p_j(t)$ prior probability for each class j in t.

Iris data set : search in first direction



Iris data set : search in second direction



The model

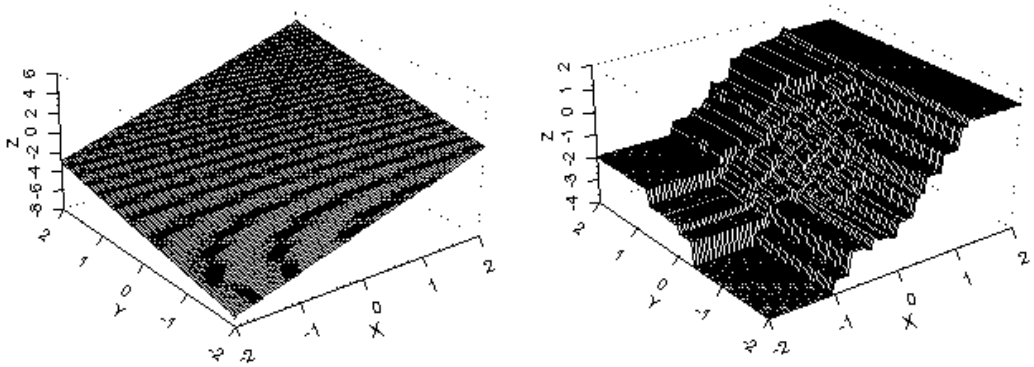
Split the root t into two children t_L et t_R Do the same recursively.
Stop when at least one of the following conditions is satified :

- very few observations in a node, *minsize*
- $\Delta R(s,t)$ is lower than a fixed threshold, *mindev*

The maximal tree :

- has low errors over learning sample
- is poor over test samples
- is too big, thus unreadable

Simulation p=2



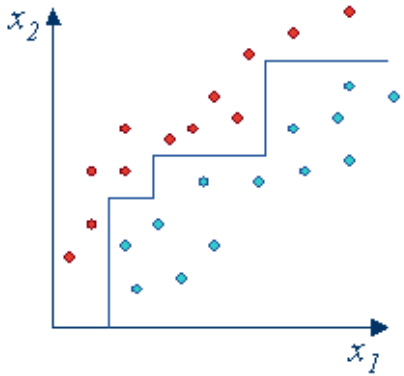
Advantages and drawbacks

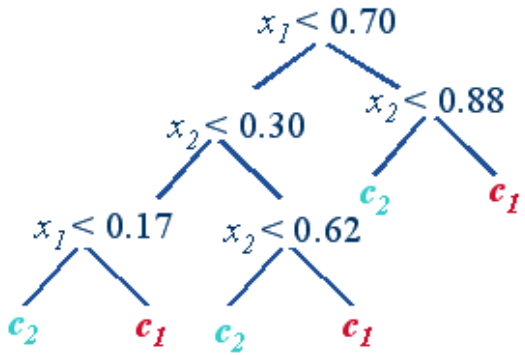
- Working in high dimension
- Variables of different natures
- Regression - Classification
- Model easy to interpret
- Interactions between variables used
- Dealing with missing data
- Variables importance
- Many extensions possible

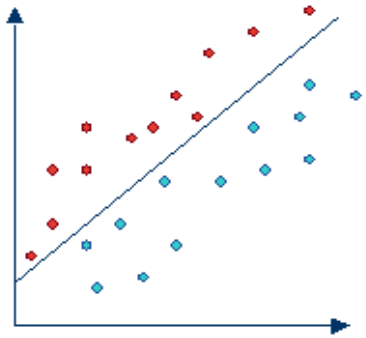
Drawback : Instability

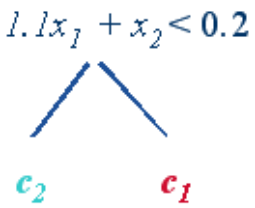
Extensions

- Oblique CART, OC1.
- Multivariate CART in regression and classification
- Multiple Regression within each node.
- Bayesian Cart.
- Bootstrap within nodes (Direct approach of instability).

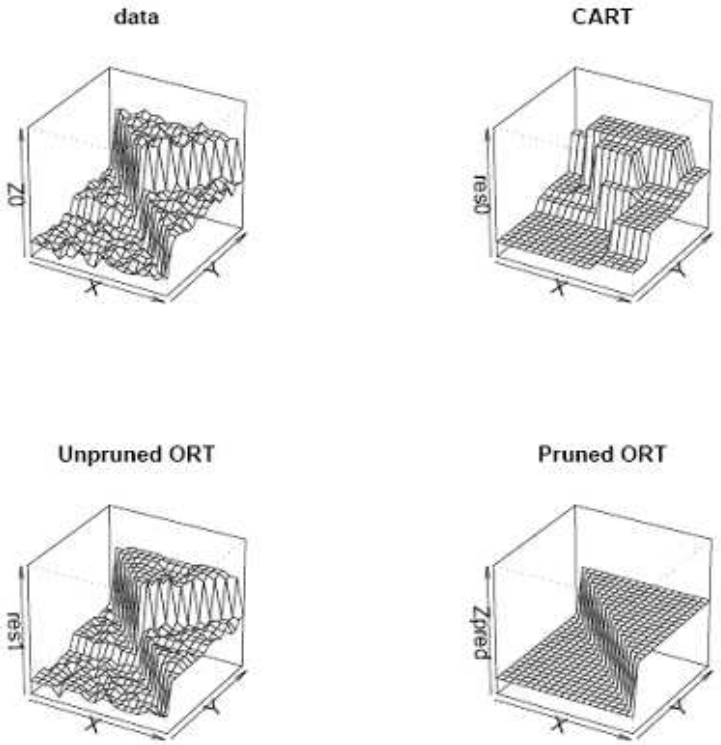

x_2
x_1


$x_1 < 0.70$
$x_2 < 0.30$
$x_2 < 0.88$
$x_1 < 0.17$
$x_2 < 0.62$
c_2
c_1
c_2
c_1
c_2
c_1
c_2
c_1




$1.1x_1 + x_2 < 0.2$
c_2
c_1

Oblique Regression Trees



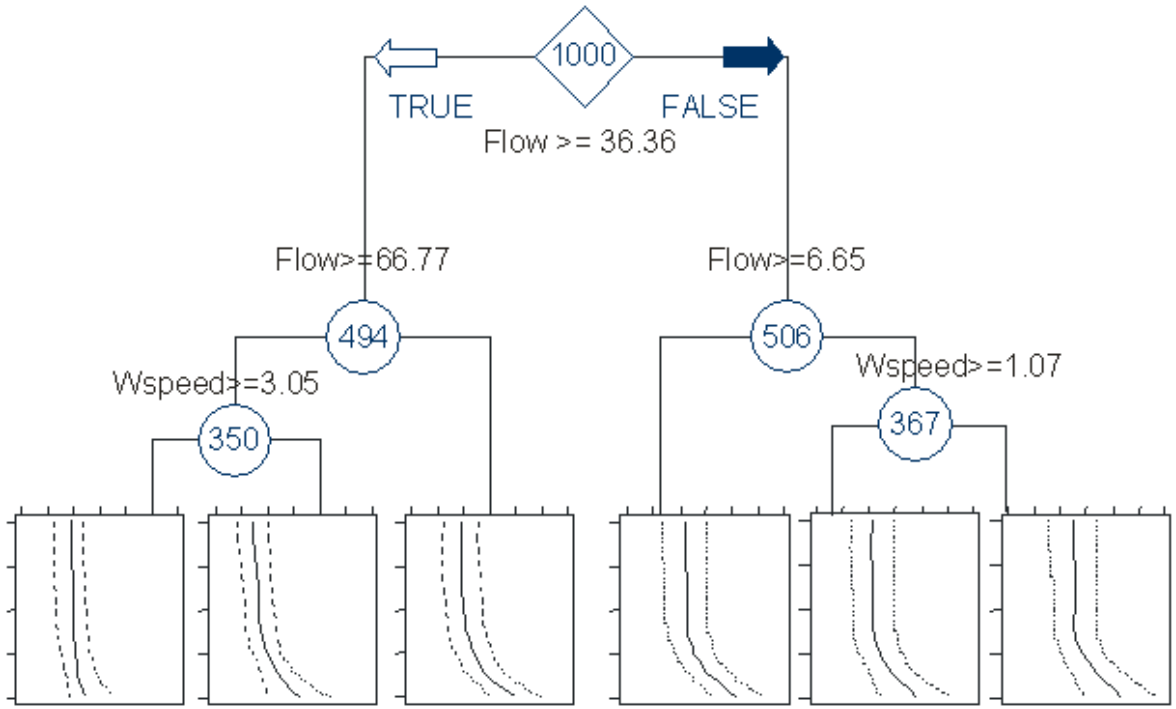
Multidimensional or functional output

- Predict a vector and/or a functional. $Y \in R^d$, or $Y \in L^2(R)$
- Predict the daily ozone profile
- Predict the size distribution of zooplanktons (indicator of changes in climate)
- Predict the profiles of sea salinity

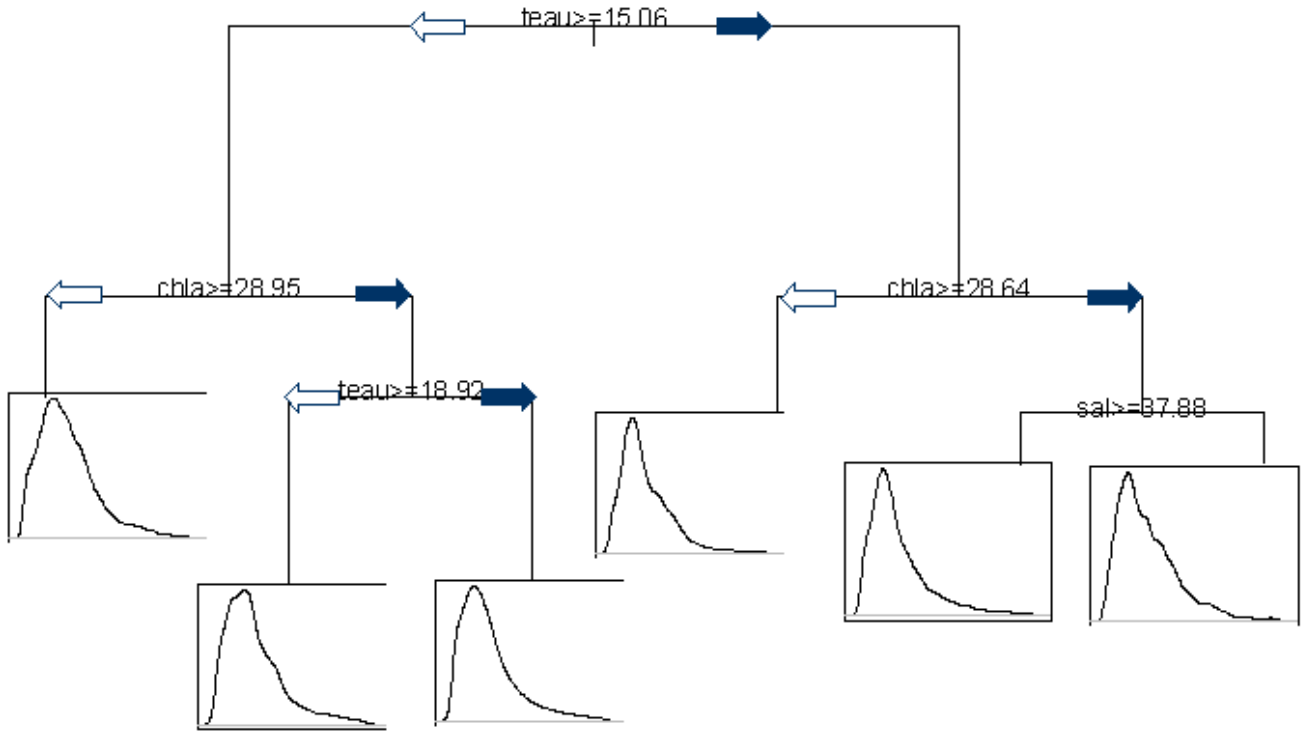
The regression function has the form :

$$f(x) = E[Y|X = x] = \sum_{j=1}^{q} f_j I(X \in N_j)$$

Predicting Salinity profiles



Modeling zooplankton sizes' densities



Sommaire

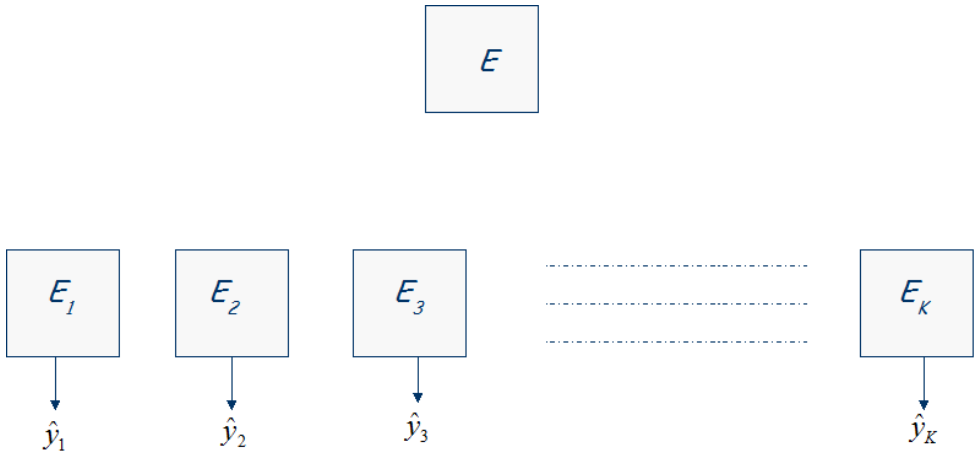
Why might we agregate

- Instabilty ?
- Multiple models ?
- Boost ?

Bagging, Boosting, ...

- Freund : Weak Learner $\Rightarrow$ Strong learner
  (vote within several "learners"),
  "Boosting" (1995)
- Breiman : Unstable "Classifier" $\Rightarrow$ Stable (by bootstrap aggregation)
  "Bagging" (1996), "Arcing" (1999)

Aggregation



In Regression,

$$f^{(a)}(x) = \frac{1}{K}\sum_{k=1}^{K} \widehat{f}_k(x)$$

In Classification,

$$f^{(a)}(x) = Argmax_j \sum_{k=1}^{K} \mathbf{1}_{\widehat{f}_k(x)=j}$$

Example-Breast Cancer

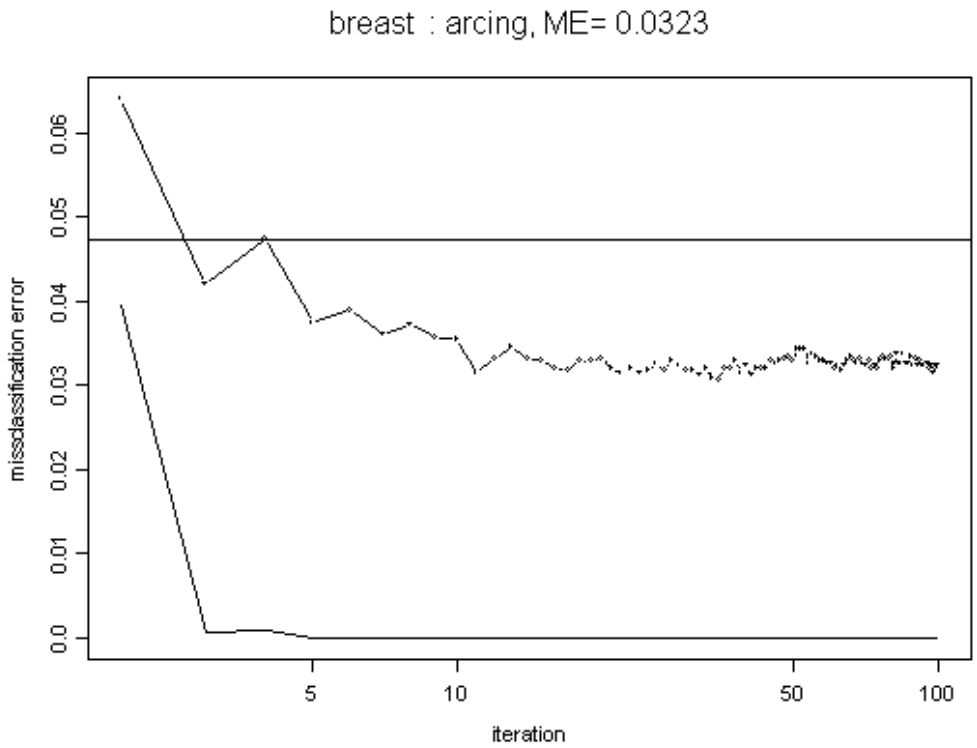


FIGURE – Learning and Test errors of the boosted classifier

Random Forests, Breiman L., 2001

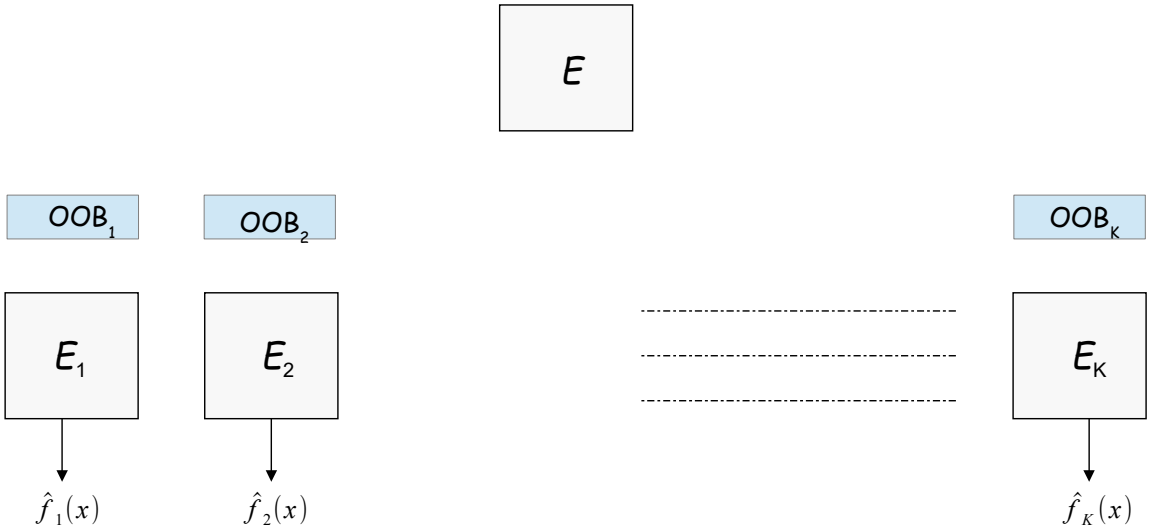
- Construct bootstrap samples of the data
- For each node of the tree, select the optimal split searching over only $p' <<< p$ variables among the p ones, selected randomly.
  ($p' = log(p)$ or $p' = \sqrt{p}$)
- Don't prune the tree
- Aggregate the trees like in bagging

- Leave the OOB sample aside
- Random Features : random linear combination of the selected variables at each node

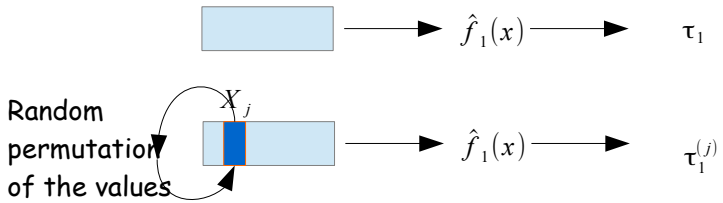
RF Properties

- Each tree has a low bias (but high variance)
- Trees are not correlated
- The correlation is defined to be the one computed between trees' predictions over OOB samples.
- Very high performences, "Best of the chelf classifier"
- Computational complexity reduced
- Possible parallelization

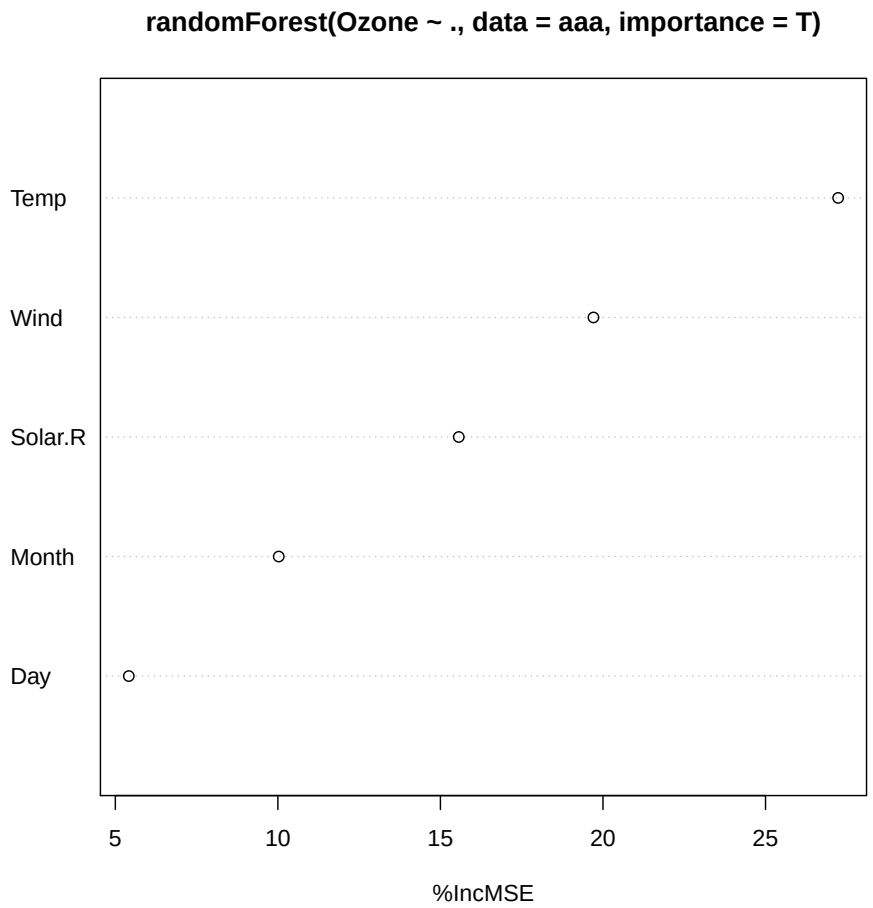
VarImp with RF



For each variable j, for each OOB :



VarImp with RF, Airquality example



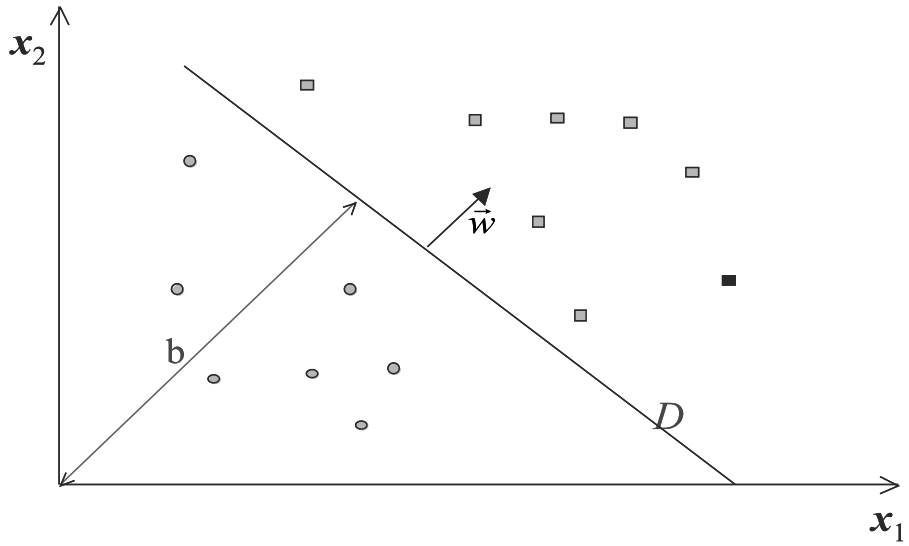
Sommaire

Linear Separation

$\mathcal{S}=$ *n i.i.d.* sample of $(\mathcal{X}, \mathcal{Y}) \subseteq (\mathbb{R}^p, \{-1, +1\})$

$$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subseteq (\mathcal{X} \times \mathcal{Y})^n .$$

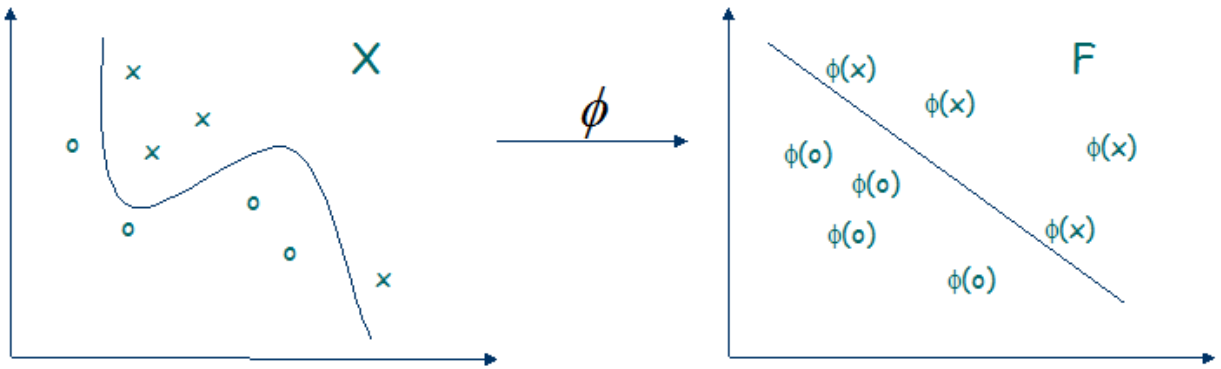
We look for a function : $f(x) = sign(\langle w.x \rangle + b)$



Non linear separation

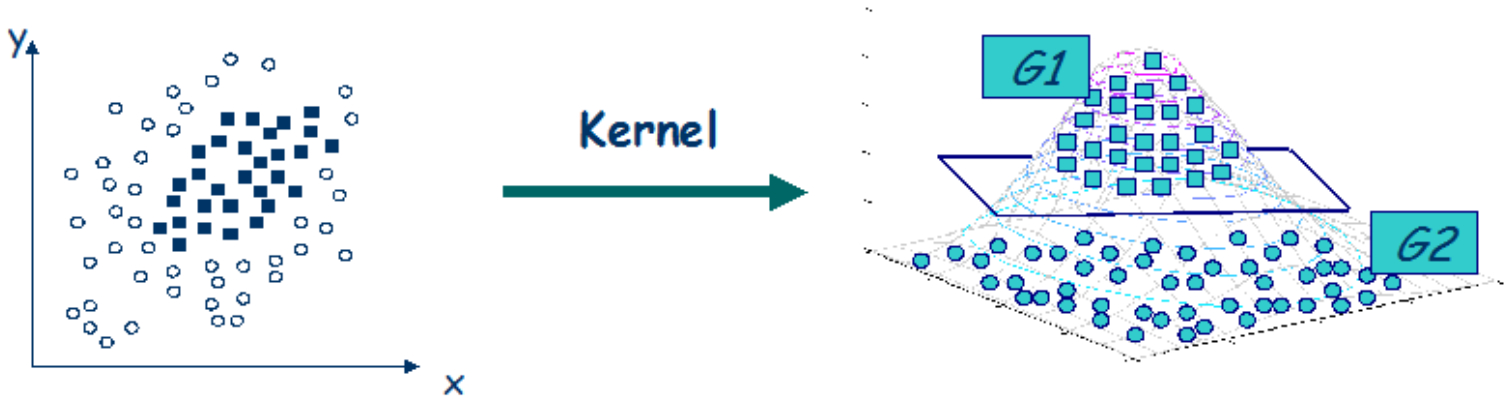
$$\begin{aligned} \phi: \quad & \mathcal{X} \to \mathcal{F} \\ & \mathbf{x} \to \phi(\mathbf{x}) \end{aligned}$$

$\mathcal{X}$ is the "attribute space" and $\mathcal{F}$ "the *feature space*"



If $Q < q$, dimension reduction, example PCA.
ϕ is non linear in general. A linear separator is learned in $\mathcal{F}$.

Non linear separation 2



Sommaire

DoE approach

- Collect measurements of $Y \in R^q$ for fixed values of $X \in R^p$
- Adjust a 2 dimensional polynomial regression $Y = f(X) + \epsilon$.
- Use the adjusted model f to predict Y for any new value of X.
- Find the region within X satisfying some constraints over Y.

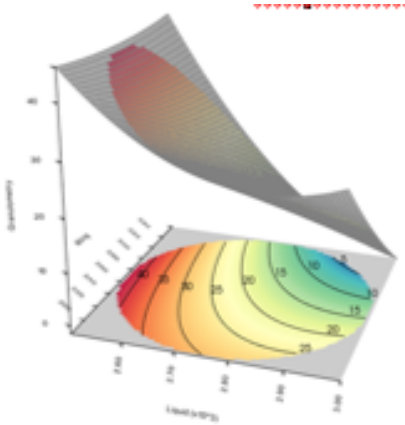
- Polynomial models are simple and easy to estimate
- The quality of the adjustment depends on the regularity of the response within the space.
- Degree of polynomials limited by sample size, as well as interactions...
- Confidence Intervals ?

The dataset, Manzon et. al. 2020

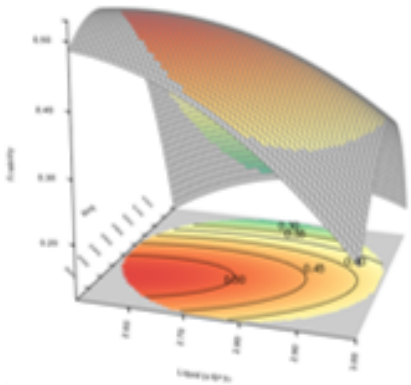
x_1 = Liq de mouillage (g), 2500-3000.
x_2 = Tps de malax (s), 180-360.

	X1	X2	Granulometrie	Friabilite	Durete	Rapport.trans	Ind.cohesion
1	1.000	0.000	12.8	0.38	11.20	89.5	705
2	−1.000	0.000	44.3	0.47	15.50	71.3	916
3	0.500	0.866	5.4	0.23	8.65	91.3	570
4	−0.500	−0.866	34.5	0.50	13.25	75.6	769
5	0.500	−0.866	27.2	0.35	10.95	88.9	706
6	−0.500	0.866	25.8	0.24	10.90	81.4	748
7	0.000	0.000	20.2	0.46	12.10	89.5	742
8	0.000	0.000	21.1	0.48	11.90	89.8	734
9	0.000	0.000	19.3	0.49	11.85	90.6	747
10	−0.433	−0.250	29.3	0.51	13.30	83.4	808
11	0.433	−0.250	19.1	0.46	11.40	91.1	715
12	0.000	0.500	15.9	0.39	10.90	90.7	704

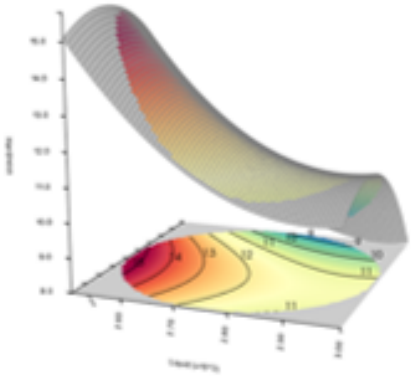
$20\% < Gran < 45\%, Friab < 0.5, Durete > 11, Ratio.transm > 80\%, Ind.cohesion > 600$.



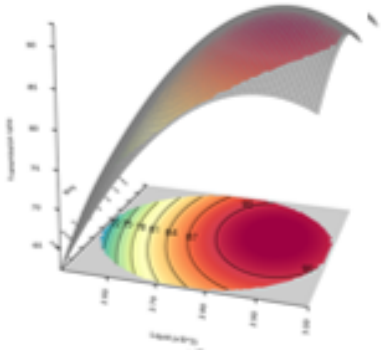
$$Y1 = 20{,}22 - 14{,}6X_1 - 8{,}86X_2 + 8{,}36X_1^2 + 1{,}21X_2^2 - 7{,}72X_1X_2$$



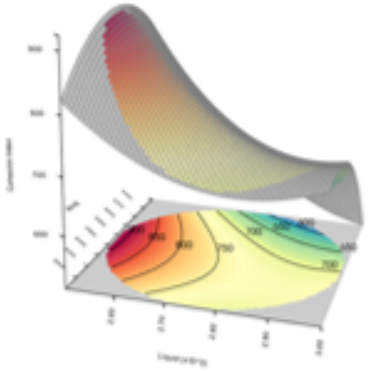
$$Y2 = 0{,}48 - 0{,}05X_1 - 0{,}11X_2 - 0{,}05X_1^2 - 0{,}18X_2^2 + 0{,}079X_1X_2$$



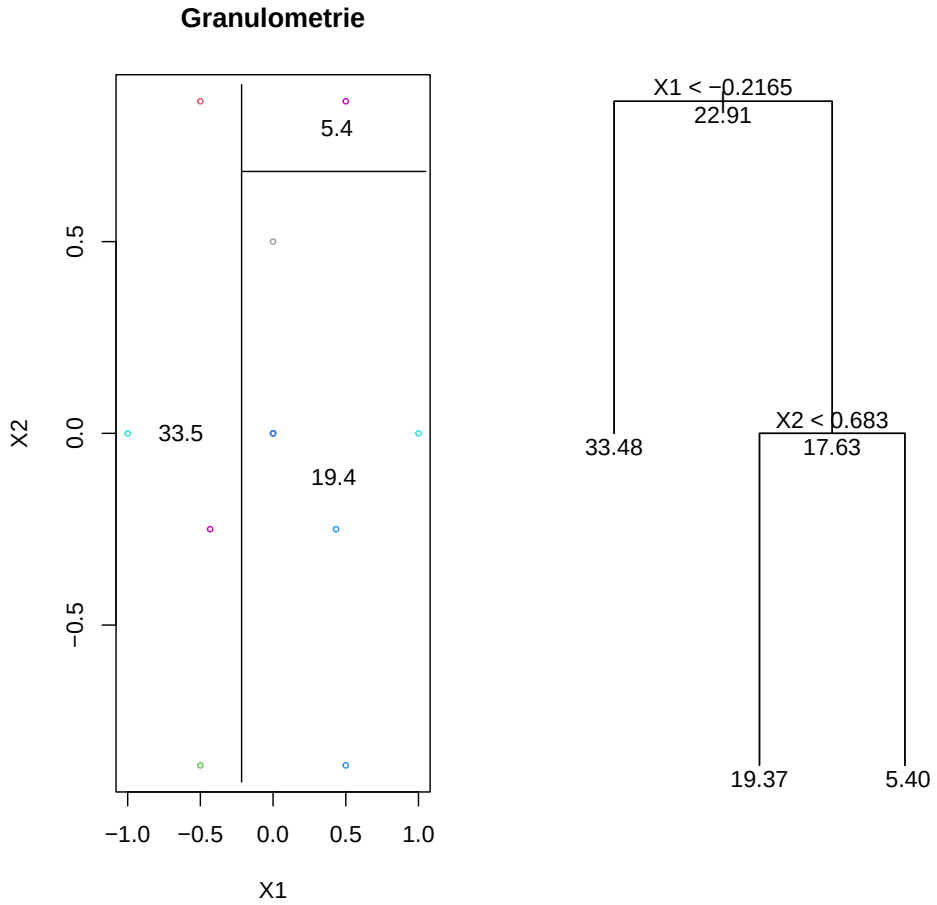
$$Y3 = 11{,}94 - 2{,}19X_1 - 1{,}32X_2 + 1{,}4X_1^2 - 1{,}8X_2^2 + 0{,}03X_1X_2$$

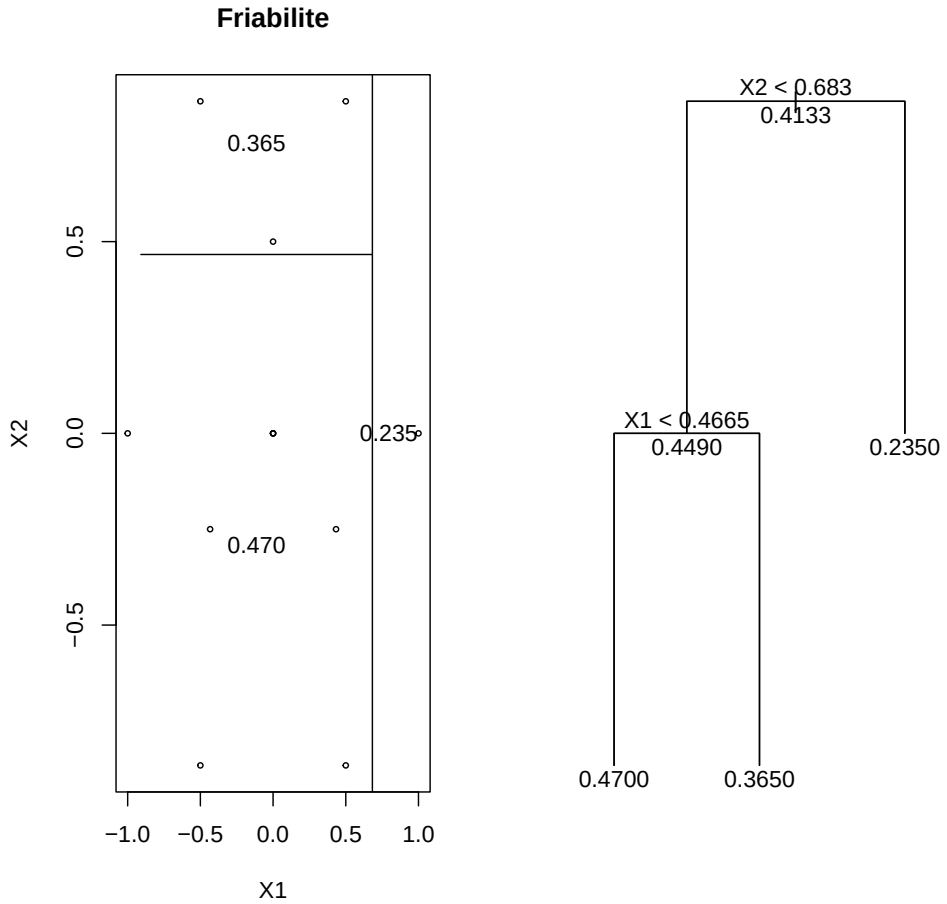


$$Y4 = 90{,}05 + 9{,}77X_1 + 2{,}47X_2 - 9{,}69X_1^2 - 4{,}39X_2^2 - 1{,}8X_1X_2$$



$$Y5 = 742{,}26 - 11{,}96X_1 - 44{,}67X_2 + 68{,}13X_1^2 - 80{,}89X_2^2 - 64{,}94X_1X_2$$


Granulometrie
5.4
33.5
19.4
0.5
0.0
−0.5
X2
−1.0
−0.5
0.0
0.5
1.0
X1
X1 < −0.2165
22.91
33.48
X2 < 0.683
17.63
19.37
5.40


Friabilite
0.365
0.235
0.470
X2
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
X1
X2 < 0.683
0.4133
X1 < 0.4665
0.4490
0.2350
0.4700
0.3650

ML for 2D case sudy

$lm = linear\ model, rf = randomForests,$
$svm = Support\ vector\ machines,\ knn = k - nearest\ neighbors.$

	Granu	Friab	Durete	Rap.trans	Ind.coh
lm	0.6315	0.01128	0.07188	0.7021	5.167
svm	3.5160	0.02707	0.65920	2.0360	35.630
rf	10.2300	0.07714	1.73000	4.2530	78.680
knn	3.8260	0.04062	0.53830	2.7410	28.530

TABLE – RMSE for each response over learning samples.

Models are not optimized !

Simulation case

- $x_1 \sim U[0,1]$, $x_2 \sim U[0,1]$
- $y_1 = 20 - x_1 - x_2 + 18x_1^2 + x_2^2 + 5x_1x_2 + \epsilon$
  $y_2 = x1^3 * \exp(\sin(x_2)) + \epsilon$
  $y_3 = abs(x_1 - 0.5) + abs(x_2 - 0.5) + \epsilon$
- Different sample sizes, $n = 20, 50, 200$.
- Different regression models.
- A learning sample generated used to estimate the RS.
- An independent test sample used to assess its performance.

Comparing ML models over simulated data set

		Learning sample			Test sample		
model	n	y1	y2	y3	y1	y2	y3
lm	20	0.8011	0.8184	0.8452	1.171	1.203	1.221
lm	50	0.9226	0.9362	0.9380	1.056	1.070	1.071
lm	200	0.9920	0.9844	0.9878	1.021	1.020	1.014
svm	20	1.1030	0.8161	0.8390	2.202	1.103	1.092
svm	50	1.0020	0.9043	0.9125	1.494	1.078	1.086
svm	200	0.9914	0.9677	0.9696	1.106	1.044	1.032
rf	20	1.8490	1.1330	1.1660	1.731	1.119	1.130
rf	50	1.3810	1.1070	1.1290	1.378	1.116	1.111
rf	200	1.1780	1.0930	1.0980	1.162	1.102	1.086
knn	20	1.5340	0.8079	0.8478	2.489	1.099	1.094
knn	50	1.1380	0.8098	0.8245	1.751	1.096	1.094
knn	200	0.9039	0.8136	0.8182	1.239	1.100	1.095

Using only two variables in the ML models

		Learning sample			Test sample		
model	n	y1	y2	y3	y1	y2	y3
lm	20	0.8393	0.8133	0.8133	1.153	1.225	1.145
lm	50	0.9440	0.9377	0.9346	1.063	1.080	1.079
lm	200	0.9888	0.9864	0.9861	1.012	1.014	1.017
svm	20	1.1200	0.8214	0.8071	2.171	1.159	1.065
svm	50	1.0140	0.9115	0.9071	1.479	1.097	1.079
svm	200	0.9871	0.9686	0.9678	1.111	1.036	1.036
rf	20	1.8920	1.1460	1.1140	1.748	1.157	1.088
rf	50	1.4080	1.1170	1.1190	1.375	1.127	1.112
rf	200	1.1710	1.0920	1.0900	1.166	1.092	1.088
knn	20	1.5810	0.8097	0.8003	2.565	1.137	1.072
knn	50	1.1460	0.8125	0.8167	1.749	1.120	1.094
knn	200	0.8996	0.8115	0.8160	1.242	1.097	1.097

Going further...

- Other criteria ?
- Other models ; xgboost, neural networks, ...
- Variables or features selection.
- Design space ? how to compare different approaches ?

References

- Freund, Y. *Boosting a weak learning algorithm by majority.*. Inf. Comput. 121, No.2, 256-285 (1995)
- Trevor, H. Tibshirani, R. and Friedman, J. *The elements of statistical learning. Data mining, inference and prediction. (English)*. Springer Series in Statistics. New York, NY : Springer.
- Breiman L., Friedman J., Olshen J., Stone C., *Classification And Regression Trees*, 1984.
- Breiman L. Random Forests, *Journal of Machine Learning Research*, 2001.
- Cristianini, N. & Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, 2000.