Introduction to machine learning

Performance criteria

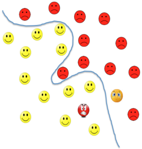
Mathilde Mougeot

ENSIIE

2018

Performances criteria

confusion matrix & co.



Machine learning for binary classification

A Classification machine, M, has been calibrated with historical data base

For one new observation, x_{new}, the machine computes and attributes a binary label $\hat{y}_{new}\{0,1\}$

- Supervized learning :
  If the target has been previously observed y_{new},
  If the (x_{new}, y_{new}) couple is available.
- Evaluation of the answer of the machine M :
  - Correct answers : $\hat{y}_{new} = y$
  - Error : $\hat{y}_{new} \neq y$

Confusion matrix

- Performances for a set of labeled data

$g(x) = \hat{y}$	$y = 0$	$y = 1$
$g(x) = 0$	OK/0	False Negative
$g(x) = 1$	**False Positive**	OK/1
	n_0	n_1

- Criteria : Global Performance (accuracy) , Global Error
- False Positive : wrong diagnose $\hat{Y} = 1$ instead of $Y = 1$.
- False Negative : wrong diagnose $\hat{Y} = 0$ instead of $Y = 0$.
- Sensitivity : Ability to diagnose $\hat{Y} = 1$ for $Y = 1$
- Specificity : Ability to diagnose les $\hat{Y} = 0$ for $Y = 0$

Challenge : To find a trade-off between Sensibility and Specificity

Standard Error for binary classification

Reality / Decision	$y = 0$	$y = 1$	
$\hat{y} = 0$	TN	FN	
$\hat{y} = 1$	FP	TP	#(predicted P)
		#(real P)	

- Accuracy $= \frac{TP + TN}{TP+TN+FP+FN}$ What is the global performance ?
- Recall $= \frac{TP}{\#(\text{real P})} = \frac{TP}{FN+TP}$ How may relevant items are selected ?
- Precision $= \frac{TP}{\#(\text{predicted P})} = \frac{TP}{FP+TP}$ How may selected items are relevant ?
- F-score$= 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Rem. : Recall= sensitivity.
False-Discovery Rate (FDR)= 1-Precision.

Confusion matrix

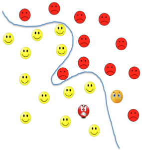
Computed on a test data set (2 classes)
Credit risk (1 : pb of credit).
Data set : $n = 200$, $n_0 = 120$ $\{0\}$, $n_1 = 80$ $\{1\}$ (pb of credit)

$g(x) = \hat{y}$	$\{0\}$	$\{1\}$	TOTAL
prediction $\{0\}$	110	**10**	120
prediction $\{1\}$	**10**	70	80
TOTAL	120	80	200

- Performance : $\frac{110+70}{200} = \frac{180}{200}$. Taux d' Erreur $= \frac{10+10}{200} = \frac{20}{200} = 10\%$
- Sensitivity (Recall) $= 70/80$
- Specificity $= 110/120$
- False Positive rate $= \frac{10}{120} = 8,33\%$
- False Negative rate $= \frac{10}{80} = 12,5\%$

Performance Criteria

ROC curve



Classifier performance : ROC curve

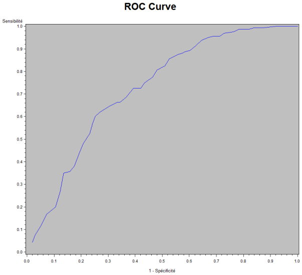
The machine M computes a score, a probability of obtaining $Y = 1$.
An Observation i is affected to class $Y = 1$

$$\text{if } \hat{\eta}(x_i) > S, \text{ (S : Threshod, MAP Threshold =0.5)}$$

ROC : acronyme de Receiver Operating System

- For 2 classes classification problem
- used for comparison of several models
- used to adjust the threshold (for sensibility and False positive rate)
- Sensitivity (recall)
  - If $score(x) > S$ then $\hat{Y} = 1$, "Event detected"
  - $\alpha(s) = P(score(x) > S/x = evenement)$
  - $\alpha(s) = P(\hat{Y} = 1/Y = 1)$
- False Positive rate = 1 - Specificity
  - $P(\hat{Y} = 1/Y = 0) = 1 - \beta(s)$ with $\beta = P(\hat{Y} = 0/Y = 0)$
  - False Positive rate $1 - \beta(s) = P(score(s) > s/x = non - evenement)$

ROC curve. Model performances

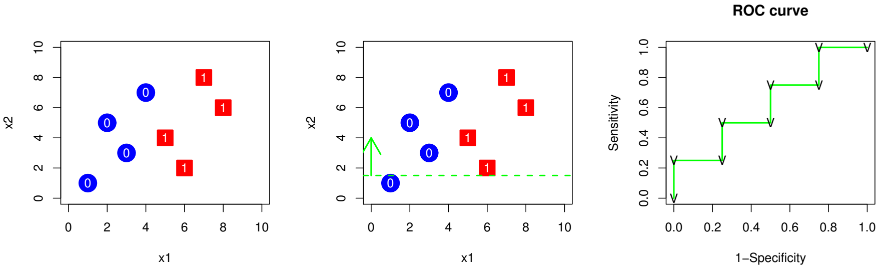


- Sensitivity (True Positive Rate) : to be able to well detect an Event
- Specificity : to be able to well detect a non-Event
- **Graphique Roc : y : Sensibility(c) ; x : 1-Specificity(c)**

The Area under the ROC curve (AUC) is a measure of the "Predictive Power" of the model.

ROC curve, illustration

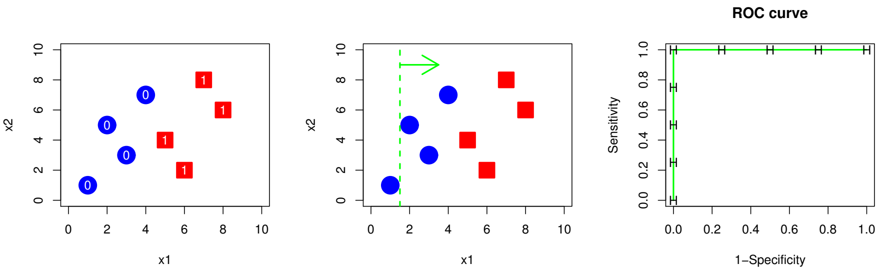
We suppose here that $score(x) = g(x) = x_2$ (very bad choice)



seuilH	α	β	$1-\beta$
0.5	1.00	0.00	1.00
1.5	1.00	0.25	0.75
2.5	0.75	0.25	0.75
3.5	0.75	0.50	0.50
4.5	0.50	0.50	0.50
5.5	0.50	0.75	0.25
6.5	0.25	0.75	0.25
7.5	0.25	1.00	0.00
8.5	0.00	1.00	0.00

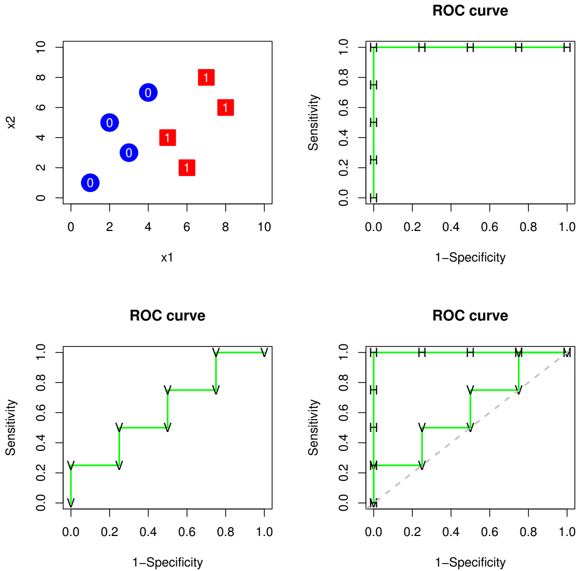
ROC curve, illustration

We suppose here that $score(x) = g(x) = x_1$ (very smart choice)

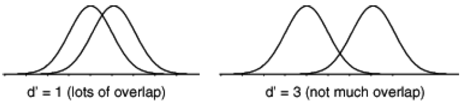


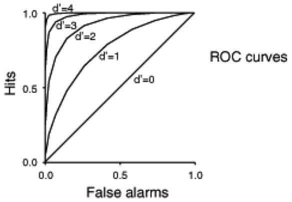
seuilV	α	β	$1-\beta$
0.5	1.00	0.00	1.00
1.5	1.00	0.25	0.75
2.5	1.00	0.50	0.50
3.5	1.00	0.75	0.25
4.5	1.00	1.00	0.00
5.5	0.75	1.00	0.00
6.5	0.50	1.00	0.00
7.5	0.25	1.00	0.00
8.5	0.00	1.00	0.00

ROC curve, illustration

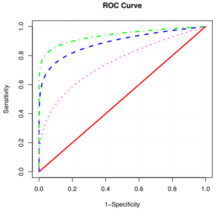


The Gold Standard for Scoring : the ROC curve (k=2)





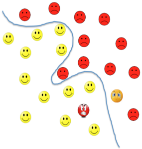
ROC Curve



- Diagonal ROC curve : the performance of the model is like a "random model".
- The more the curve is upper on the left, the better is the model
- The ROC curves let to compare different models (globally with AUC) et locally (around a threshold)
- This curve is independent of $Y = 0$ and $Y = 1$.

Machine learning

Predictive Power



The goal of machine learning is to built machines with capacities of Generalization . The predictive power is compute on Test data *independant* of the Train data.

Cross-validation

- Generalization is the goal of supervised learning
- A trained classifier has to be generalizable. It must be able to work on other data than the training dataset
- Generalizable means "works without over fitting"
- This can be achieved using cross-validation
- There is no machine learning without cross-validation atsome point !
- In the case of penalization, we need to choose a penalizationparameter C that generalizes

Cross-validation

- Cross-validation :$\mathcal{D}_n = \mathcal{D}_{\text{Train}} + \mathcal{D}_{\text{Test}}$
  - $\mathcal{D}_{\text{Train}}$ calibration of the parameters of model (model selection)
  - $\mathcal{D}_{\text{Test}}$ Performance evaluation

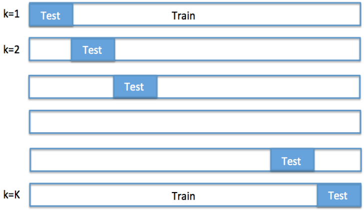
  Problem : Possible impact of the Train or the Test set on the performances, depending on the chosen data.
  The data are often chosen at random.
- K-Fold cross validation
- Leave One Out

Figure on the black board

Notations : $\mathcal{D}_n = \{(x_i, y_i)\ i = 1, \ldots, n\ , y_i \in \mathcal{X},\ x_i \in \mathbb{R}^p\}$

K-Fold cross validation

- K validations with K different data sets for Train and Test.
- Kfold :
  - $\mathcal{D}_n = \mathcal{D}_{Train_k} + \mathcal{D}_{Test_k}$
  - $\mathcal{D}_{Test_k} = \mathcal{D}_n - \mathcal{D}_{Train_k}$



Generalization vs Model Complexity