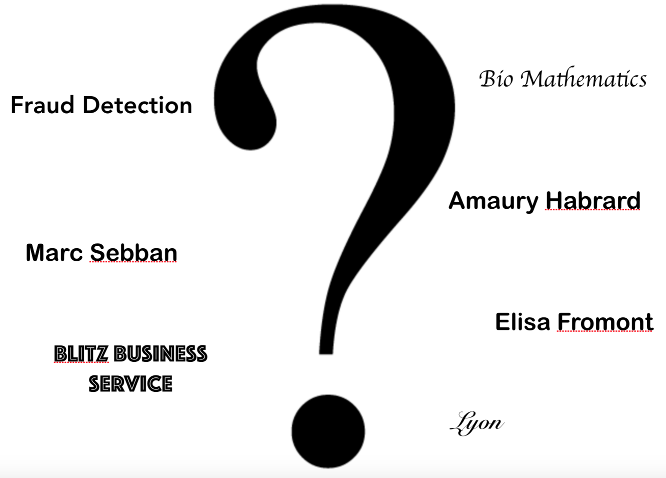
# Cascade Classification

Metzler Guillaume
Under the Supervision of Marc Sebban,
Fromont Elisa & Habrard Amaury

Université Jean Monnet - Laboratoire Hubert Curien

30 Juin 2016

# Let me introduce myself



Fraud Detection

Bio Mathematics

Amaury Habrard

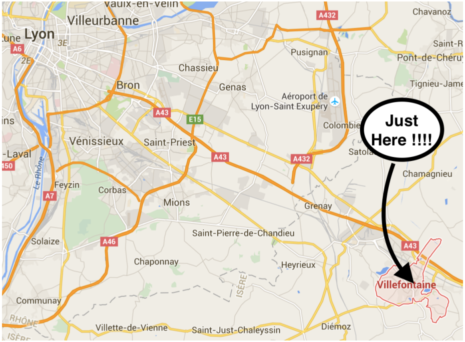
Marc Sebban

Elisa Fromont

BLITZ BUSINESS SERVICE

Lyon

# Blitz Business Service



**Where are we ?**

**Who are we ?**

- Works with large scale distributions (hypermarket) : payement security
- Computer Science engineers, one researcher.
- We mainly work on cheque transaction : false cheque and unpaid transactions

# Blitz Business Service

## Few numbers about cheques

- Represent 8% of the transactions in France,
- but more than 10% of the turnover of a supermarket.
- Less than 0.3% of unpaid transaction.
- Around 0.02% of cheques are false.

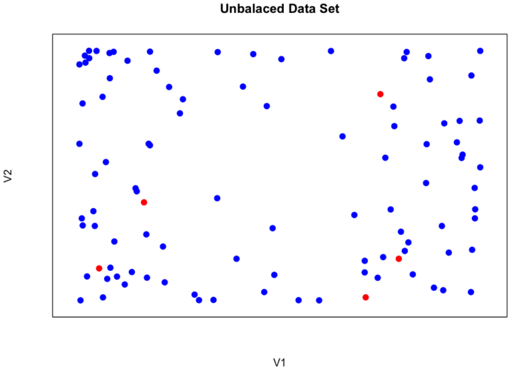
Cheques are currently less used and their period of validity is reduced.

# Context

The context is the following :

- Binary classification (creditworthy or not), extremely unbalanced data.
- A huge number of data.
- Fraudsters' behaviour evolve through time (Concept Drift).
- Common machine learning technics don't work in this context.

# Context

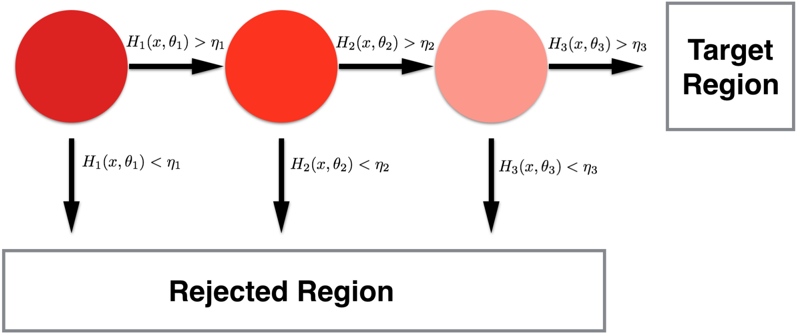


### Measure

We'll avoid to minimize the error rate : not appropriate for the context.

We should prefer optimizing any loss function $\mathcal{L}$ based on F-Measure or AUC for instance.

# Cascade Classification : what is it ?


$H_1(x,\theta_1) > \eta_1$
$H_2(x,\theta_2) > \eta_2$
$H_3(x,\theta_3) > \eta_3$
Target Region
$H_1(x,\theta_1) < \eta_1$
$H_2(x,\theta_2) < \eta_2$
$H_3(x,\theta_3) < \eta_3$
Rejected Region

Z. Xu, Matt J.Kusner, M. Chen, O.Chapelle. *Classifier Cascades and Trees for Minimizing Feature Evaluation Cost. Journal Of Machine Learning* ,pages 2113-2144, 2014.

# How does it work ?

- At each stage : you supress data from the negative class
- To be predicted postive, a data must satisfy :

$$\forall i \in \mathcal{I}, \quad sign(H_i(x, \theta_i) - \eta_i) > 0$$

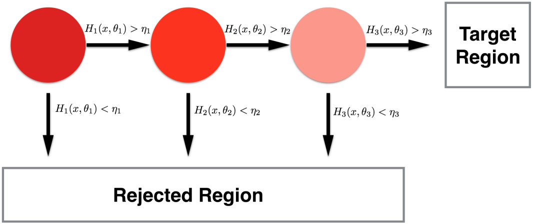
- If it exists an index $j$ such that $sign(H_j(x, \theta_j) - \eta_j) > 0$, $x$ is then predicted negative.

## Advantages

- It works well in the context of unbalanced data.
- Possibility to optimize a global loss function.
- Interaction during the learning process.

## Disadvantages

- The more stages you have, the harder the learning process will be. (Time + Complexity)
- No guarantee to achieve the optimal solution.


$H_1(x,\theta_1) > \eta_1$
$H_2(x,\theta_2) > \eta_2$
$H_3(x,\theta_3) > \eta_3$
Target Region
$H_1(x,\theta_1) < \eta_1$
$H_2(x,\theta_2) < \eta_2$
$H_3(x,\theta_3) < \eta_3$
Rejected Region

# Two versions of the cascade classification

Hard : Number of data decrease along the cascade during the learning process.
Resolution of a discrete optimization problem.

Soft : probabilistic and continuous version of cascade classification.
Prediction of each classifier are probabilities got a sigmoid function (logistic function) :

$$\frac{1}{1+\exp(-<\theta,x>)}.$$

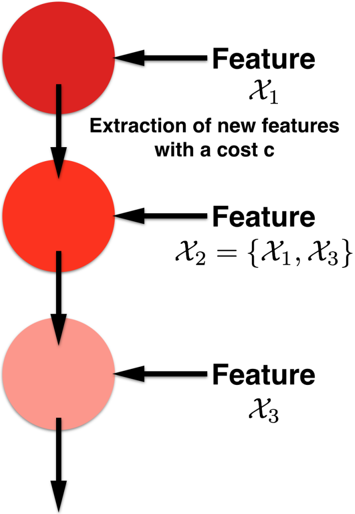
# Tools

- Gradient or Block Coordinate Gradient Descent. Stochastic versions of Gradient Descent.
- Boosting methods and ensemble methods with specific coefficients for the ensemble methods.
- We use a validation data set to learn the different thresholds.

# Context of my PhD

## Hypothesis and objectives

- We want to minimize a loss function $\mathcal{L}(TP, TN, FP, FN)$ which is not convex.
- We only have access to a limited number of features at each stage.
- We want to achieve a high detection rate by minimizing the cost of the procedure $\mathcal{L}$.


Feature
$\mathcal{X}_1$
Extraction of new features with a cost c
Feature
$\mathcal{X}_2 = \{\mathcal{X}_1, \mathcal{X}_3\}$
Feature
$\mathcal{X}_3$

# How do we tackle the problem

- Due to the context we have three stages in our cascade (or more) : each classifier consists of a logistic regression :

$$h(x,\theta) = \mathcal{P}(Y = 1 \mid x, \theta) = \frac{1}{1 + \exp(- < x, \theta >)}.$$

- The problem is non linearly separable → ensemble methods :

$$H_j(x,\theta) = \sum_{k=1}^{K_j} \alpha_k \cdot h_k(x, \theta^{(j)})$$

# How do we tackle the problem

We learn a sort of "logistic regression forest" in order to improve the classification. The experience has shown that random forest are effective to tackle this problem but cannot be inclued in cascade in order to minimize a global loss function.

Coefficients are defined as the value of the F-Measure associated to each weak learners :

$$\frac{2P(Y=1 \mid x,\theta,y=1)}{2P(Y=1 \mid x,\theta,y=1)+P(Y=1 \mid x,\theta,y=0)+P(Y=0 \mid x,\theta,y=1}$$

# In practice

- Blitz cost function uses indicator functions that are non differentiable → surrogate loss which is a logistic loss :

  $$\mathcal{L} = y \cdot \log(H(x,\theta) + (1-y) \cdot \log(1-H(x,\theta)),$$

  where $H(x,\theta) = \prod_{j=1}^{J} H_j(x,\theta_j)$
- We use ensemble methods to solve the problem.
- Parameters of the model are then adjusted to optimize the reel loss function.
- We learn the the different threshold with a validation procedure.

# The End ! !

Thank you for your attention
Any questions ?