Collective learning strategies

Aurélie Fischer

LPMA, Université Paris Diderot – Paris 7

Août 2014, Journées MAS, Toulouse

Outline

1. Introduction
2. Collective strategy in classification
3. Collective strategy in regression
   - with G. Biau, B. Guedj, J. D. Malley
4. Experimental results
5. Toward “mixed” collective methods
   - with M. Mougeot

Outline

Classification...

- $(\mathbf{X}, Y) \in \mathbb{R}^d \times \{0,1\}$: Predict the label Y using $\mathbf{X}$.
  $\Rightarrow$ Find a function $g : \mathbb{R}^d \to \{0,1\}$, where $g(X)$ is our guess for Y.
- Probability of error

$$L(g) = \mathbb{P}(g(\mathbf{X}) \neq Y)$$

minimal when g is the Bayes classifier :

$$g^\star(\mathrm{x}) = \begin{cases} 1 \text{ if } \mathbb{E}[Y|\mathbf{X} = \mathrm{x}] > \frac{1}{2} \\ 0 \text{ otherwise} \end{cases} .$$

- Distribution of $(\mathbf{X}, Y)$ unknown in practice, access to $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, i.i.d random pairs $\sim (\mathbf{X}, Y)$.
- Aim: Construct the best possible classifier g_n based on $\mathcal{D}_n$, performance measured by
  $L_n = L(g_n) = \mathbb{P}(g_n(\mathbf{X}) \neq Y | (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$.

- $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$: Find a relationship between the observation vector $\mathbf{X}$ and the response Y.
  $\Rightarrow$ Function f such that $f(\mathbf{X})$ is a good approximation of Y.
- Quadratic risk

$$\mathbb{E}\,|f(\mathbf{X}) - Y|^2,$$

  minimal when f is the regression function

$$r^\star(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}].$$

- Distribution of $(\mathbf{X}, Y)$ unknown, sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$.
- Aim: Estimate the regression function $r^\star$ using the data $\mathcal{D}_n$.

Combining several methods

Growing number of different estimation methods + parameters

How to choose the right strategy ?

$\rightarrow$ Combine several estimators $f_1, \ldots, f_M$.

- Model selection
- Linear / convex combination

For instance :

Nemirovski (2000), Juditsky and Nemirovski (2000), Catoni (1999, 2004), Massart (2007), Tsybakov (2003, 2004), Wegkamp (2003), Yang (2000, 2001, 2004), Györfi, Kohler, Krzyżak, and Walk (2002), Audibert (2004), Birgé (2006), Dalalyan and Tsybakov (2008), van de Geer (2008), Koltchinskii (2009), Bunea, Tsybakov, and Wegkamp (2004, 2006, 2007a,b).

Example: Aggregate with exponential weights:

$$f_n^{\mathrm{AEW}} = \sum_{m=1}^{M} w(f_m) f_m, \qquad w(f_m) = \frac{\exp(-\frac{n}{T} R_n(f_m))}{\sum_{m=1}^{M} \exp(-\frac{n}{T} R_n(f_m))}$$

Outline

Mojirsheibani (1999, 2000, 2002a,b)

- For each $\mathbf{x}$, find the $\mathbf{X}_i$'s such that every individual classifier predicts the same label for $\mathbf{X}_i$ and $\mathbf{x}$.
- Label $\mathbf{y}$ estimated by majority vote among corresponding Y_i's.

$\mathbf{x}$	$C_1(\mathbf{x})$	$C_2(\mathbf{x})$	$C_3(\mathbf{x})$	y
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	1	1	1	1
$\mathbf{x}_4$	1	0	1	1
$\mathbf{x}_5$	0	0	0	1
$\mathbf{x}_6$	1	0	1	0
$\mathbf{x}_7$	0	1	0	1
$\mathbf{x}_8$	1	0	1	0
$\mathbf{x}_9$	1	1	0	1
$\mathbf{x}_{10}$	1	0	1	0
$\mathbf{x}_0$	1	0	1	?

Combination of classifiers

Mojirsheibani (1999, 2000, 2002a,b)

- For each $\mathbf{x}$, find the $\mathbf{X}_i$'s such that every individual classifier predicts the same label for $\mathbf{X}_i$ and $\mathbf{x}$.
- Label $\mathbf{y}$ estimated by majority vote among corresponding Y_i's.

$\mathbf{x}$	$C_1(\mathbf{x})$	$C_2(\mathbf{x})$	$C_3(\mathbf{x})$	y
$\mathbf{x}_1$	0	1	1	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	1	1	1	1
$\mathbf{x}_4$	1	0	1	1
$\mathbf{x}_5$	0	0	0	1
$\mathbf{x}_6$	1	0	1	0
$\mathbf{x}_7$	0	1	0	1
$\mathbf{x}_8$	1	0	1	0
$\mathbf{x}_9$	1	1	0	1
$\mathbf{x}_{10}$	1	0	1	0
$\mathbf{x}_0$	1	0	1	0

Outline

The setting

Sample splitting $\mathcal{D}_n = \mathcal{D}_k \cup \mathcal{D}_\ell$.

- $\mathcal{D}_k$: Initial estimators $r_{k,1}, \cdots, r_{k,M}$, parametric, semi-parametric or nonparametric + possible tuning rules
  $\Rightarrow$ linear regression, nearest neighbour, kernel regression, Lasso, random forests...
- $\mathcal{D}_\ell$: Combination step

The general idea

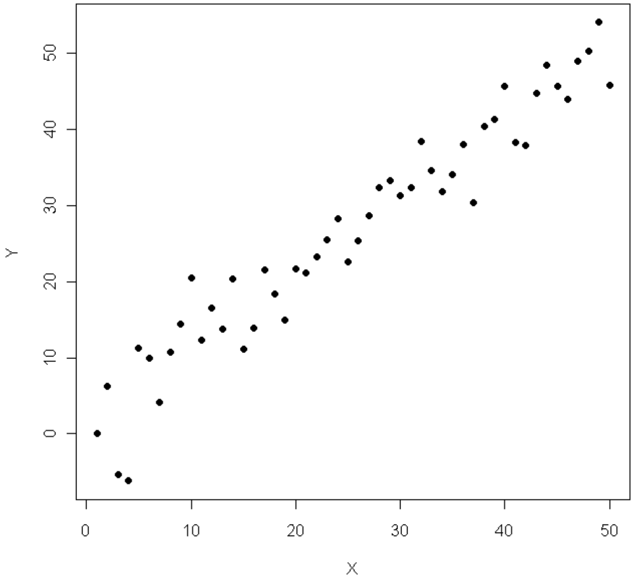
Local average procedure: Neighbourhood ?

Initial estimators as distance indicator between the observations.

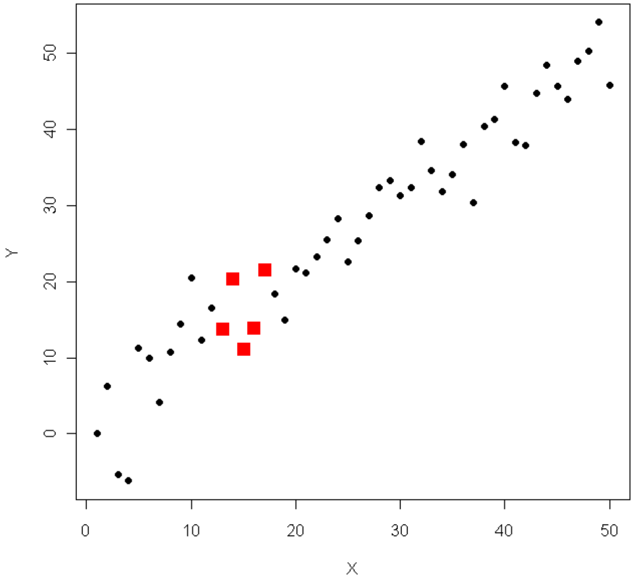
- Observation $\mathbf{X}_i$ considered to be reliable if all initial estimators predict a similar value for this data point and $\mathbf{x}$.
  $\rightarrow$ not more distant than a prespecified threshold ε.
- Average of corresponding Y_i's.

Nonlinear with respect to $r_{k,1}, \cdots, r_{k,M}$.

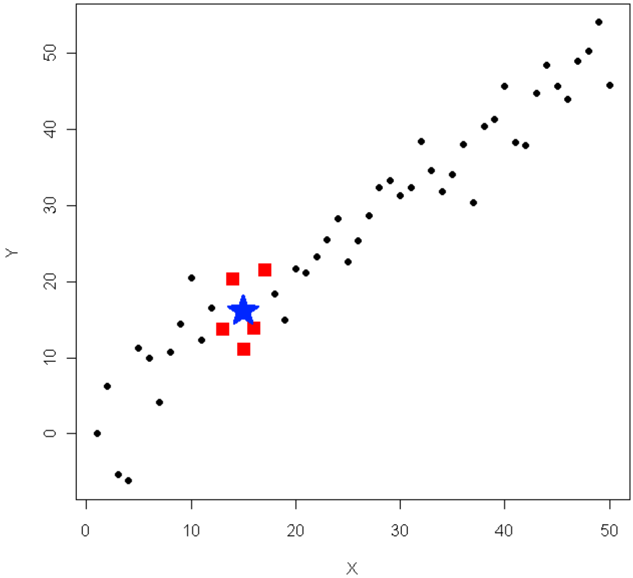
Example: x=15



Example: x=15



Example: x=15



Formal definition

Let $\mathbf{r}_k(\mathbf{x}) = (r_{k,1}(\mathbf{x}), \cdots, r_{k,M}(\mathbf{x}))$.

Collective estimator:

$$T_n(\mathbf{r}_k(\mathbf{x})) = \sum_{i=1}^{\ell} W_{n,i}(\mathbf{x}) Y_i, \quad \mathbf{x} \in \mathbb{R}^d,$$

where

$$W_{n,i}(\mathbf{x}) = \frac{\mathbf{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_i)| \le \varepsilon_\ell\}}}{\sum_{j=1}^{\ell} \mathbf{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{X}_j)| \le \varepsilon_\ell\}}},$$

where $\varepsilon_\ell > 0$ (convention: $0/0 = 0$).

Parameter $\varepsilon_\ell \sim$ kernel bandwidth

- Too large: Average of all the Y_i's.
- Too small: Not enough data retained.

Performance of T_n

Assume :

- $\mathbb{E}|r_{k,m}(\mathbf{X})|^2 < \infty$ for all $m = 1, \ldots, M$.
- For any $m = 1, \ldots, M$,

$$r_{k,m}^{-1}((t, +\infty)) \underset{t \uparrow +\infty}{\searrow} \emptyset \quad \text{and} \quad r_{k,m}^{-1}((-\infty, t)) \underset{t \downarrow -\infty}{\searrow} \emptyset.$$

Performance of T_n assessed by

$$\mathbb{E}\left|T_n\left(\mathbf{r}_k(\mathbf{X})\right) - r^\star(\mathbf{X})\right|^2.$$

Let

$$T\left(\mathbf{r}_k(\mathbf{x})\right) = \mathbb{E}\left[Y|\mathbf{r}_k(\mathbf{x})\right].$$

then

$$\mathbb{E}\left|T(\mathbf{r}_k(\mathbf{X})) - Y\right|^2 \leq \inf_f \mathbb{E}\left|f(\mathbf{r}_k(\mathbf{X})) - Y\right|^2.$$

An upper bound

For all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$,

$$\begin{aligned}&\mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \\ &\quad \leq \mathbb{E}|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))|^2 + \inf_f \mathbb{E}|f(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2,\end{aligned}$$

In particular :

Proposition 1

For all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$,

$$\begin{aligned}&\mathbb{E}\,|T_n\,(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \\ &\quad \leq \min_{m=1,\cdots,M} \mathbb{E}\left|r_{k,m}(\mathbf{X}) - r^\star(\mathbf{X})\right|^2 + \mathbb{E}\,|T_n\,(\mathbf{r}_k(\mathbf{X})) - T\,(\mathbf{r}_k(\mathbf{X}))|^2\,.\end{aligned}$$

Convergence of T_n to T

Proposition 2

Assume that

$$\varepsilon_\ell \to 0 \text{ and } \ell\varepsilon_\ell^M \to \infty \text{ as } \ell \to \infty.$$

Then

$$\mathbb{E}\left|T_n(\mathbf{r}_k(\mathbf{X})) - T(\mathbf{r}_k(\mathbf{X}))\right|^2 \to 0 \text{ as } \ell \to \infty,$$

for all distribution of $(\mathbf{X}, Y)$ *with* $\mathbb{E}Y^2 < \infty$.

An asymptotic result as corollary

Corollary 1

$$\limsup_{\ell \to \infty} \mathbb{E}\, |T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \leq \min_{m=1,\ldots,M} \mathbb{E}\, \left| r_{k,m}(\mathbf{X}) - r^\star(\mathbf{X}) \right|^2.$$

- The combined estimator performs asymptotically at least as well as the best one in the list.
- For all distributions of $(\mathbf{X}, Y)$: Universal result.

Proposition 3

Assume that :

- Y *and the* $r_{k,m}$*'s are* *bounded* *by a constant* R.
- $|T(\mathbf{r}_k(\mathbf{x})) - T(\mathbf{r}_k(\mathbf{y}))| \leq L|\mathbf{r}_k(\mathbf{x}) - \mathbf{r}_k(\mathbf{y})|$ *for* $k \geq 1$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Then, with the choice $\varepsilon_\ell \propto \ell^{-\frac{1}{M+2}}$,

$$\mathbb{E}\,|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})|^2 \leq \min_{m=1,\ldots,M} \mathbb{E}\left|r_{k,m}(\mathbf{X}) - r^\star(\mathbf{X})\right|^2 + C(R,L)\ell^{-\frac{2}{M+2}}.$$

$C(R, L)$ independent of k.

Optimality

If one of the initial estimators is consistent for a given smoothness class $\mathcal{M}$ of distributions, then T_n inherits this property.

Corollary 2

Assume that one of the original estimators, say r_{k,m_0}, satisfies

$$\mathbb{E}\left|r_{k,m_0}(\mathbf{X}) - r^\star(\mathbf{X})\right|^2 \to 0 \quad \text{as } k \to \infty,$$

for all distribution of $(\mathbf{X}, Y)$ in some smoothness class $\mathcal{M}$. Then, under the assumptions of Proposition 3, with the choice $\varepsilon_\ell \propto \ell^{-\frac{1}{M+2}}$,

$$\lim_{k,\ell\to\infty} \mathbb{E}\left|T_n(\mathbf{r}_k(\mathbf{X})) - r^\star(\mathbf{X})\right|^2 = 0.$$

Alternate definition

- In the definition

$$W_{n,i}(\mathbf{x}) = \frac{\mathbf{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\leq\varepsilon_\ell\}}}{\sum_{j=1}^{\ell}\mathbf{1}_{\bigcap_{m=1}^{M}\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_j)|\leq\varepsilon_\ell\}}},$$

  all estimators are asked to satisfy the closeness condition: Unanimity.
- May be relaxed : For example, require only a fraction $\alpha \in (0,1]$ of the estimators:

$$W_{n,i}(\mathbf{x}) = \frac{\mathbf{1}_{\{\sum_{m=1}^{M}\mathbf{1}_{\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_i)|\leq\varepsilon_\ell\}}\geq M\alpha\}}}{\sum_{j=1}^{\ell}\mathbf{1}_{\{\sum_{m=1}^{M}\mathbf{1}_{\{|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{X}_j)|\leq\varepsilon_\ell\}}\geq M\alpha\}}}.$$

- $\alpha \to 1$.
- Measure of "homogeneity" of the estimators.

Outline

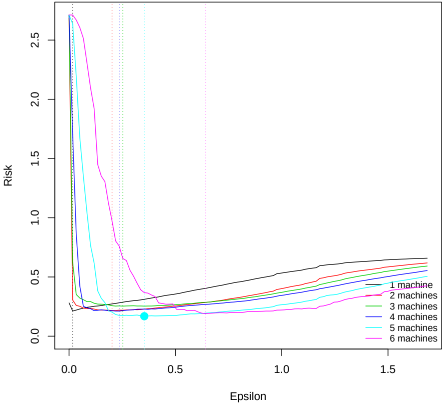
COBRA

- R Package COBRA.
- Computations may be parallelized: Very fast.
- List of default methods to be combined: `lars`, `ridge`, `FNN`, `tree`, `randomForest`, or provide your own predictions.
- Automatic calibration of ε and α : Minimize the empirical risk (data-splitting).

ε and α: Example 1

$$\mathbf{X} \sim \mathcal{U}(-1,1), n = 700, d = 20,$$

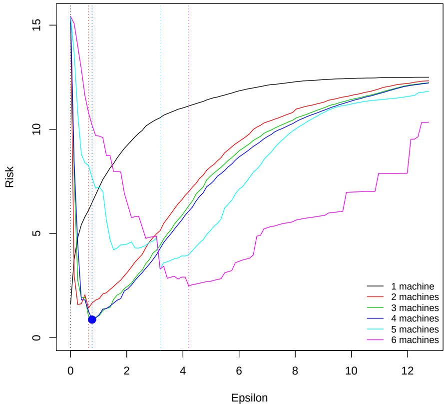
$$Y = \mathbf{1}_{\{X_1>0\}} + X_2^3 + \mathbf{1}_{\{X_4+X_6-X_8-X_9>1+X_{14}\}} + \exp(-X_2^2) + \mathcal{N}(0, 0.5).$$



ε and α: Example 2

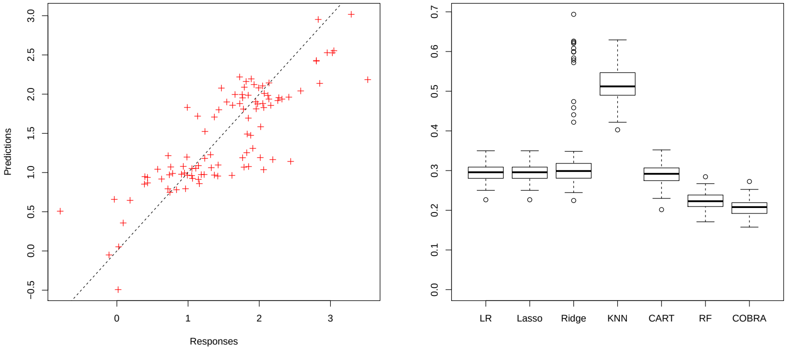
$$\mathbf{X} \sim \mathcal{N}(0, \Sigma), n = 700, d = 20,$$

$$Y = \mathbf{1}_{\{X_1>0\}} + X_2^3 + \mathbf{1}_{\{X_4+X_6-X_8-X_9>1+X_{14}\}} + \exp(-X_2^2) + \mathcal{N}(0, 0.5).$$



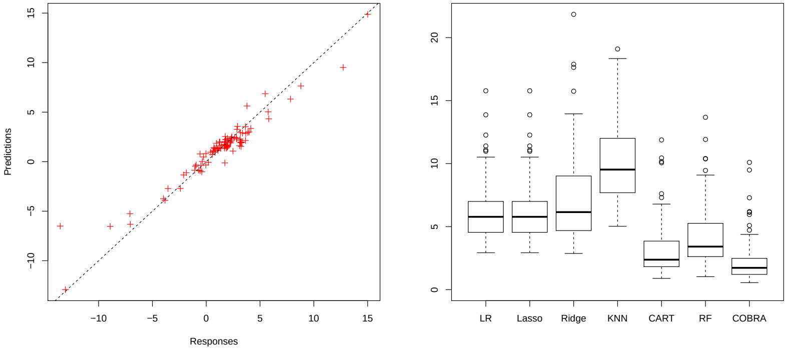
Predictive performance: Example 1

$\mathbf{X} \sim \mathcal{U}(-1, 1)$, $n = 700$, $d = 20$,
$Y = \mathbf{1}_{\{X_1>0\}} + X_2^3 + \mathbf{1}_{\{X_4+X_6-X_8-X_9>1+X_{14}\}} + \exp(-X_2^2) + \mathcal{N}(0, 0.5)$.



Predictive performance: Example 2

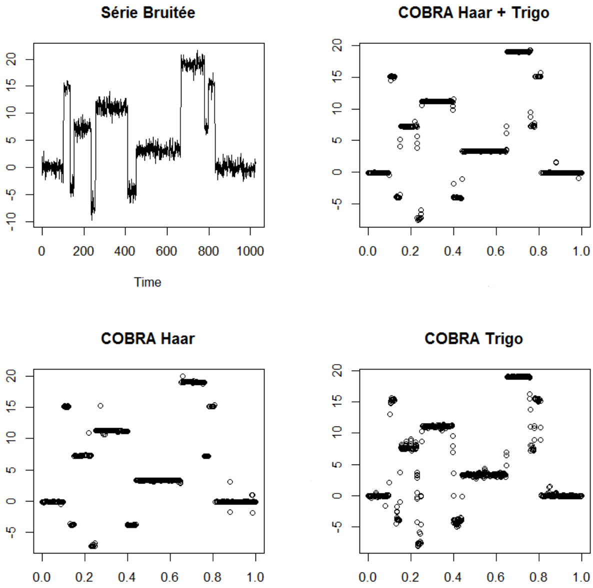
$\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, $n = 700$, $d = 20$,
$Y = \mathbf{1}_{\{X_1>0\}} + X_2^3 + \mathbf{1}_{\{X_4+X_6-X_8-X_9>1+X_{14}\}} + \exp(-X_2^2) + \mathcal{N}(0, 0.5)$.



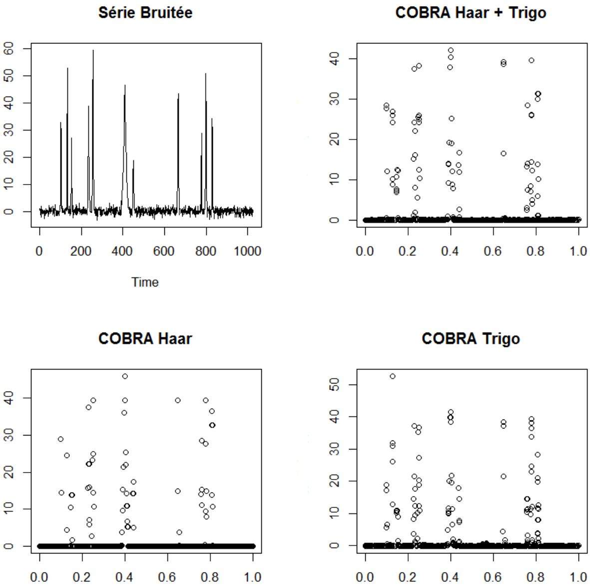
Experimental results : summary

- ⇒ Very good performance.
- Comparison to :
  - Individual estimators
  - SuperLearner, van der Laan et al. (2007)
  - Exponentially weighted aggregation.
- High-dimensional data.
- Large number of estimators.

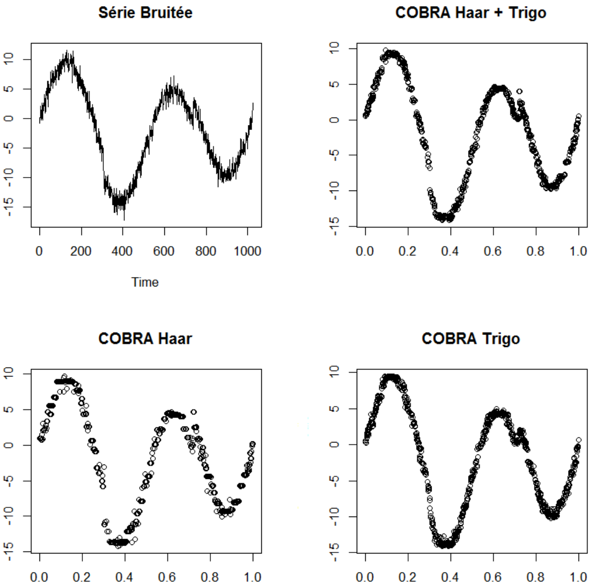
Wavelet 1



Wavelet 2

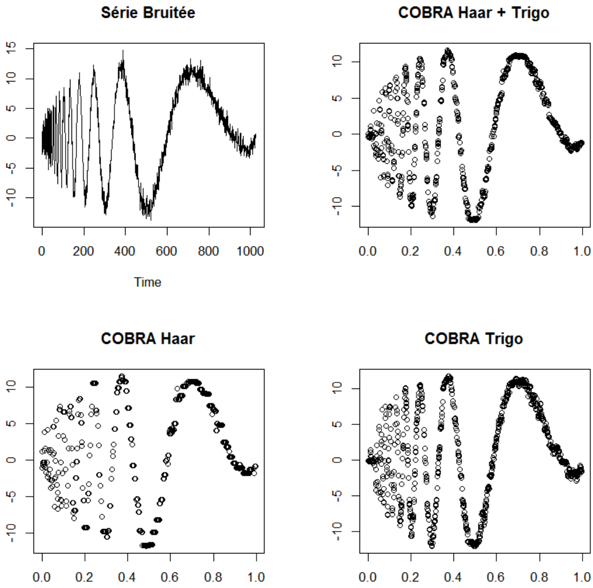

Série Bruitée
0
10
20
30
40
50
60
0
200
400
600
800
1000
Time
COBRA Haar + Trigo
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
COBRA Haar
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
COBRA Trigo
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0

Wavelet 3



Série Bruitée
Time
0
200
400
600
800
1000
10
5
0
-5
-10
-15
COBRA Haar + Trigo
COBRA Haar
COBRA Trigo
0.0
0.2
0.4
0.6
0.8
1.0

Wavelet 4



Série Bruitée
15
10
5
0
-5
-10
0
200
400
600
800
1000
Time
COBRA Haar + Trigo
10
5
0
-5
-10
0.0
0.2
0.4
0.6
0.8
1.0
COBRA Haar
10
5
0
-5
-10
0.0
0.2
0.4
0.6
0.8
1.0
COBRA Trigo
10
5
0
-5
-10
0.0
0.2
0.4
0.6
0.8
1.0

Outline

"Mixed" collective classification

- Back to classification.
- Aim : Improve the collective method by restricting the influence of a possible bad classifier.
- Combining preceding ideas with "geometric" distance between inputs.
- More regular definition of the collective classifier : kernel + plug-in rule.

  $\Rightarrow$ Distances involved :

$$\frac{1}{a}\|\mathbf{X}_i - \mathbf{x}\|^2 + \frac{1}{b}\sum_{m=1}^{M} \mathbf{1}_{\{C_m(\mathbf{X}_i) \neq C_m(\mathbf{x})\}}.$$

"Mixed" collective classification

- Back to classification.
- Aim : Improve the collective method by restricting the influence of a possible bad classifier.
- Combining preceding ideas with "geometric" distance between inputs.
- More regular definition of the collective classifier : kernel + plug-in rule.

  ⇒ Distances involved :

$$\frac{1}{a}\sum_{j=1}^{d}(\mathbf{X}_{ij}-\mathbf{x}_j)^2+\frac{1}{b}\sum_{m=1}^{M}(C_m(\mathbf{X}_i)-C_m(\mathbf{x}))^2.$$

Exponential bound

For instance, Gaussian kernel.

Under appropriate assumptions on bandwidths a, b :

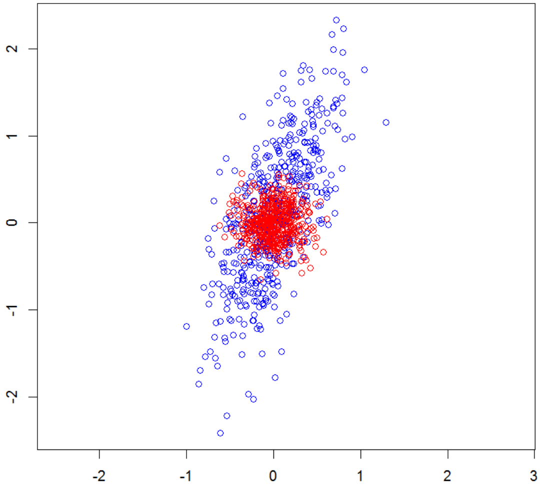
Proposition 4

For any distribution of $(\mathbf{X}, Y)$ and for every $\varepsilon > 0$, there is n_0 such that for $n > n_0$, the error L_n of the mixed collective rule satisfies

$$P(L_n - L^\star > \varepsilon) \leq 2e^{-n\varepsilon^2/C}.$$

$\Rightarrow$ Strong universal consistency.

An illustration



Error rates

Example Mixed 1

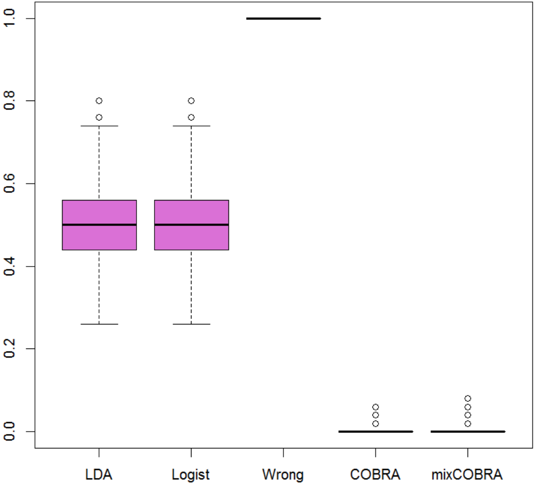
LDA	0.54
Logist	0.54
Wrong	1
COBRA	0
MixCOBRA ($a = b = 0.01$)	0

Example Mixed 2

LDA	0.54
Logist	0.54
Random	0.46
COBRA	0.36
MixCOBRA ($a = 0.01,\ b = 10$)	0.28

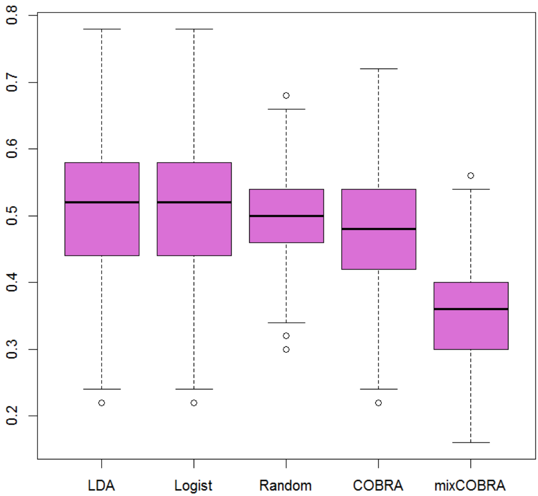
Boxplots Mixed 1

1000 trials



Boxplots Mixed 2

1000 trials



References I

Jean-Yves Audibert. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistique*, 40: 685–736, 2004.

Lucien Birgé. Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 42:273–325, 2006.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for regression learning, 2004. URL http://arxiv.org/abs/math/0410214. Preprint LPMA, Universités Paris 6 - Paris 7.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation and sparsity via ℓ_1-penalized least squares. In Gábor Lugosi and H. U. Simon, editors, *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006), Lecture Notes in Artificial Intelligence*, volume 35, pages 379–391. Springer-Verlag, Berlin-Heidelberg, 2006.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35:1674–1697, 2007a.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 35:169–194, 2007b.

References II

Olivier Catoni. "Universal" aggregation rules with exact bias bounds, 1999. Preprint LPMA, Universités Paris 6 - Paris 7.

Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization*. Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001, Lecture Notes in Mathematics. Springer, 2004.

Arnak Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Anatoli Juditsky and Arkadi Nemirovski. Functional aggregation for nonparametric estimation. *The Annals of Statistics*, 28:681–712, 2000.

Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 45:7–57, 2009.

Pascal Massart. *Concentration Inequalities and Model Selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

Majid Mojirsheibani. Combining classifiers via discretization. *Journal of the American Statistical Association*, 94:600–609, 1999.

References III

Majid Mojirsheibani. A kernel-based combined classification rule. *Statistics & Probability Letters*, 48:411–419, 2000.

Majid Mojirsheibani. An almost surely optimal combined classification rule. *Journal of Multivariate Analysis*, 81:28–46, 2002a.

Majid Mojirsheibani. A comparison study of some combined classifiers. *Communications in Statistics - Simulation and Computation*, 31:245–260, 2002b.

Arkadi Nemirovski. *Topics in Non-Parametric Statistics*. École d'Été de Probabilités de Saint-Flour XXVIII – 1998. Springer, 2000.

Alexandre B. Tsybakov. Optimal rates of aggregation. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Computational Learning Theory and Kernel Machines*, Lecture Notes in Computer Science, pages 303–313. Springer-Verlag, 2003.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32:135–166, 2004.

Sara A. van de Geer. High dimensional generalized linear models and the Lasso. *The Annals of Statistics*, 36:614–645, 2008.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6, September 2007.

References IV

Marten H. Wegkamp. Model selection in nonparametric regression. *The Annals of Statistics*, 31:252–273, 2003.

Yuhong Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74:135–161, 2000.

Yuhong Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96:574–588, 2001.

Yuhong Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10: 25–47, 2004.