Robust sequential learning

with applications to the forecasting of air quality and of electricity consumption

Gilles Stoltz

CNRS — École normale supérieure — INRIA, équipe CLASSIC

& HEC Paris

A statistician has to predict a sequence $y_1, y_2, \ldots$ of observations lying in some set $\mathcal{Y}$.

His predictions $\widehat{y}_1, \widehat{y}_2, \ldots$ are picked in a set $\mathcal{X}$.

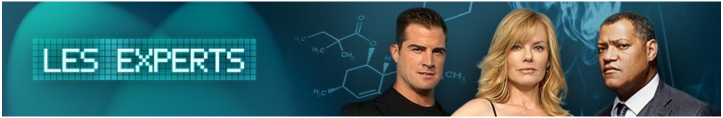
Observations and predictions (1) are made in a sequential fashion and (2) rely on no stochastic modeling.

(1) means that for each instance, the prediction $\widehat{y}_t$ of y_t is determined

- solely based on the past observations $y_1^{t-1} = (y_1, \ldots, y_{t-1})$,
- before getting to know the actual value y_t.

(2) indicates that the methods at hand will not resort to the estimation of some parameters of some stochastic process to build a good model and get some accurate forecasts from it.

To make the problem meaningful, finitely many expert forecasts are called for.



At each instance t, expert $j \in \{1, \ldots, N\}$ outputs a forecast

$$f_{j,t} = f_{j,t}\big(y_1^{t-1}\big) \in \mathcal{X}$$

The statistician now determines $\widehat{y}_t$ based

- on the past observations $y_1^{t-1} = (y_1, \ldots, y_{t-1})$,
- and the current and past expert forecasts $f_{j,s}$, where $s \in \{1, \ldots, t\}$ and $j \in \{1, \ldots, N\}$.

We assume that the set $\mathcal{X}$ of predictions is convex and we restrict the statistician to form convex combinations of the expert forecasts.

At each instance t, the statistician thus picks a convex weight vector $\boldsymbol{p}_t = \bigl(p_{1,t}, \ldots, p_{N,t}\bigr)$ and forms

$$\widehat{y}_t = \sum_{j=1}^{N} p_{j,t}\, f_{j,t}$$

The aim of the statistician is to predict –on average– as well as the best constant convex combination of the expert forecasts.

... But we need first to indicate how to assess the accuracy of a given prediction!

To that end, we consider a convex loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$.

When $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{Y} \subseteq \mathbb{R}$, possible choices are

- the square loss $\ell(x, y) = (x - y)^2$;
- the absolute loss $\ell(x, y) = |x - y|$;
- the absolute percentage of error $\ell(x, y) = |x - y| \big/ |y|$.

The cumulative losses of the statistician and of the constant convex combinations $\boldsymbol{q} = (q_1, \ldots, q_N)$ of the expert forecasts equal

$$\widehat{L}_T = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\ y_t\right) \qquad \text{and} \qquad L_T(\boldsymbol{q}) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\ y_t\right)$$

The regret is defined as the difference

$$R_T = \widehat{L}_T - \min_{\boldsymbol{q}} L_T(\boldsymbol{q})$$

Recall that the regret R_T is defined as the difference

$$\widehat{L}_T - \min_{\boldsymbol{q}} L_T(\boldsymbol{q}) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) - \min_{\boldsymbol{q}} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right)$$

We are interested in aggregation rules with (uniformly) vanishing per-round regret,

$$\limsup_{T\to\infty} \quad \frac{1}{T} \sup \left\{ \widehat{L}_T - \min_{\boldsymbol{q}} L_T(\boldsymbol{q}) \right\} \leqslant 0$$

where the supremum is over all possible sequences of observations and of expert forecasts.

This is why this framework is referred to as prediction of individual sequences or as robust aggregation of expert forecasts.

Note that the best convex combination $\boldsymbol{q}^\star$ can only be determined in hindsight whereas the statistician has to predict in a sequential fashion.

This framework leads to a meta-statistical interpretation:

- each series of expert forecasts may be given by a statistical forecasting method, possibly tuned with some given set of parameters;
- these base forecasts relying on some stochastic model are then combined in a robust and deterministic manner.

The cumulative loss of the statistician can be decomposed as

$$\widehat{L}_T = \min_{\boldsymbol{q}} L_T(\boldsymbol{q}) + R_T$$

This leads to the following interpretations:

- the term indicating the performance of the best convex combination of the expert forecasts is an approximation error;
- the regret term measures a sequential estimation error.

A simple strategy

Let's do some maths. But simple maths, and for 10 minutes only!

Reminder of the aim:

Uniformly bound the regret with respect to all convex weight vectors $\boldsymbol{q}$,

$$\sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\; y_t\right) - \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\; y_t\right)$$

When $\mathcal{X} \subseteq \mathbb{R}^d$ and when ℓ is convex in its first argument, sub-gradients exist, i.e.:

For all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, there exists $\nabla \ell(x,y)$ such that

$$\forall x' \in \mathcal{X}, \qquad \ell(x,y) - \ell(x',y) \leqslant \nabla \ell(x,y) \cdot (x - x')$$

To uniformly bound the regret with respect to all convex weight vectors $\boldsymbol{q}$, we write

$$\begin{aligned}
& \max_{\boldsymbol{q}} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) - \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right) \\
\leqslant \quad & \max_{\boldsymbol{q}} \sum_{t=1}^{T} \nabla \ell\left(\sum_{k=1}^{N} p_{k,t} f_{k,t},\, y_t\right) \cdot \left(\sum_{j=1}^{N} p_{j,t} f_{j,t} - \sum_{j=1}^{N} q_j f_{j,t}\right) \\
= \quad & \max_{\boldsymbol{q}} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \sum_{j=1}^{N} q_j \widetilde{\ell}_{j,t}\right) \\
= \quad & \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t}
\end{aligned}$$

where we denoted

$$\widetilde{\ell}_{j,t} = \nabla \ell\left(\sum_{k=1}^{N} p_{k,t} f_{k,t},\, y_t\right) \cdot f_{j,t}$$

Via the (signed) pseudo-losses $\widetilde{\ell}_{j,t}$, it suffices to consider the following simplified framework.

At each round $t = 1, 2, \ldots,$

- the statistician picks a convex weight vector $\mu_t = \bigl(\mu_{1,t}, \ldots, \mu_{N,t}\bigr)$;
- the environment simultaneously determines a loss vector $\ell_t = \bigl(\ell_{1,t}, \ldots, \ell_{N,t}\bigr)$;
- the values of μ_t and ℓ_t are both revealed.

The aim is to bound uniformly the regret

$$R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} \mu_{j,t} \ell_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell_{i,t}$$

Lemma. Consider two real numbers $m \leqslant M$.

For all $\eta > 0$ and for all individual sequences of elements $\ell_{j,t} \in [m, M]$, where $j \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T\}$,

$$R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} \mu_{j,t} \ell_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell_{i,t} \leqslant \frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{8} T \,,$$

where for all $j \in \{1, \ldots, N\}$, we picked $\mu_{j,1} = 1/N$ and for all $t \geqslant 2$,

$$\mu_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

This strategy is known as performing exponentially weighted averages of the past cumulative losses of the experts (with fixed learning rate η).

References: Vovk '90; Littlestone and Warmuth '94

Proof of the regret bound

It relies on Hoeffding's lemma: for all random variables X with range $[m, M]$, for all $s \in \mathbb{R}$,

$$\ln \mathbb{E}\left[e^{sX}\right] \leqslant s\,\mathbb{E}[X] + \frac{s^2}{8}(M-m)^2$$

For all $t = 1, 2, \ldots$,

$$\begin{aligned} -\eta \sum_{j=1}^{N} \mu_{j,t}\ell_{j,t} \quad &= \quad -\eta \sum_{j=1}^{N} \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)} \ell_{j,t} \\ &\geqslant \quad \ln \frac{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{t} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)} - \frac{\eta^2}{8}(M-m)^2 \end{aligned}$$

A telescoping sum appears and leads to

$$\sum_{t=1}^{T} \sum_{j=1}^{N} \mu_{j,t}\ell_{j,t} \leqslant \underbrace{-\frac{1}{\eta} \ln \frac{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{T} \ell_{j,s}\right)}{N}}_{\leqslant \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell_{i,t} + \frac{\ln N}{\eta}} + \eta \frac{(M-m)^2}{8} T \,.$$

We now discuss the obtained bound.

Recall that $[m, M]$ is the loss range.

The stated bound can be optimized in η:

$$R_T \leqslant \min_{\eta>0} \left\{ \frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{8} T \right\} = (M-m)\sqrt{\frac{T}{2} \ln N}$$

for the (theoretical) optimal choice

$$\eta^\star = \frac{1}{M-m}\sqrt{\frac{8 \ln N}{T}}$$

This choice depends on M and m, which are not necessarily known beforehand, as well as on T, which may not be bounded (if the prediction game goes forever).

Since no fixed value of $\eta > 0$ ensures that $R_T = o(T)$, we still have no fully sequential strategy... but this can be taken care of.

The possibles patches are, first, to resort to the "doubling trick."

Alternatively, the learning rates of the exponentially weighted average strategy may vary over time, depending on the past: for $t \geqslant 2$,

$$\mu_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

By a careful such adaptive choice of the η_t, the following regret bound can be obtained:

$$R_T \leqslant \square\,(M-m)\sqrt{T \ln N} + \square\,(M-m)\ln N$$

where the $\square$ denote some universal constants.

We thus recover the same orders of magnitude for the regret bound.

References: Auer, Cesa-Bianchi and Gentile '02; Cesa-Bianchi, Mansour and Stoltz '07

However, these theoretically satisfactory solutions would not work well in practice. This is what we do instead.

The exponentially weighted average strategy $\mathcal{E}_\eta$ with fixed learning rate η picks the convex combination $\mu_t(\eta)$, where

$$\mu_{j,t}(\eta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

We denote its cumulative loss $\widehat{L}_t(\eta) = \sum_{s=1}^{t} \sum_{j=1}^{N} \mu_{j,s}(\eta)\, \ell_{j,s}$

Based on the family of the $\mathcal{E}_\eta$, we build a data-driven meta-strategy which at each instance $t \geqslant 2$ resorts to

$$\mu_t(\eta_t) \qquad \text{where} \qquad \eta_t \in \operatorname*{arg\,min}_{\eta > 0} \widehat{L}_{t-1}(\eta)$$

Reference: An idea of Vivien Mallet

Non stationarity

Competing against sequences of experts with few shifts

In changing environments the performance of a given fixed convex combination $\boldsymbol{p}$ can be poor.

A more ambitious goal is to mimic the performance of sequences of the form

$$\underline{\boldsymbol{p}} = \left(\boldsymbol{p}^1, \ldots, \boldsymbol{p}^1,\ \ \boldsymbol{p}^2, \ldots, \boldsymbol{p}^2,\ \ \ldots,\ \ \boldsymbol{p}^{m+1}, \ldots, \boldsymbol{p}^{m+1}\right),$$

where among the T rounds up to m shifts can occur.

The cumulative loss $L^{\star}_{T,m}$ of the best such sequence $\underline{\boldsymbol{p}}$ is usually much smaller than the cumulative loss of the best fixed convex combination in hindsight, $\min\limits_{\boldsymbol{q}} L_T(\boldsymbol{q})$.

The cumulative loss can be decomposed as

$$\widehat{L}_T = L^{\star}_{T,m} + R_{T,m}\,,$$

where $R_{T,m}$ is the corresponding regret. And the question is:

How much larger gets the regret bound?

The fixed-share algorithm resembles the exponentially weighted average algorithm, except that at the end of each round the weights are redistributed, via a mixing with the uniform distribution:

$$p_{i,t} \qquad \text{becomes} \qquad \alpha + (1 - N\alpha) p_{i,t}$$

Fixed-share thus relies on two parameters $\alpha \geqslant 0$ and $\eta > 0$.

When these are optimally tuned, the regret bound is

$$R_{T,m} \leqslant \Box \sqrt{T m \ln N} + \ldots$$

where $\Box$ is some constant depending on the scale of the problem.

We will see that in practice –when indeed breaks occur– this worsening of the regret (by a factor of $\sqrt{m}$) is more than compensated by the better approximation error.

Two empirical studies

- Prediction of air quality
- Forecasting of the electricity consumption

Two empirical studies

The methodology of our studies is in four steps:

1. Build the experts (possibly on a training data set) and pick another data set for the evaluation of our methods;
2. Compute some benchmarks and some reference oracles;
3. Evaluate our strategies when run with fixed parameters (i.e., with the best parameters in hindsight);
4. The performance of interest is actually the one of the data-driven meta-strategies.

First study:

Prediction of air quality

Joint work with Vivien Mallet (INRIA) and M.Sc. students; published in the Journal of Geophysical Research

Some characteristics of one among the studied data sets:

- 126 days during summer ’01; one-day ahead prediction
- 241 stations in France and Germany
- Typical ozone concentrations between 40 $\mu\mathrm{g\,m}^{-3}$ and 150 $\mu\mathrm{g\,m}^{-3}$; sometimes above the values 180 $\mu\mathrm{g\,m}^{-3}$ or 240 $\mu\mathrm{g\,m}^{-3}$
- 48 experts, built in Mallet et Sportisse ’06 by choosing a physical and chemical formulation, a numerical approximation scheme to solve the involved PDEs, and a set of input data (among many)

The stations of the network are indexed by $\mathcal{S}$.

Each model $j = 1, \ldots, 48$ outputs a prediction $f_{j,t}^s$ for the ozone peak at station s and day t, which is then compared to the actual peak y_t^s.

The statistician chooses at each round a single convex weight vector $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$ of the experts to be used at all stations; this leads to prediction fields.

The strategies are assessed based on their RMSEs, which amounts to considering the convex loss functions

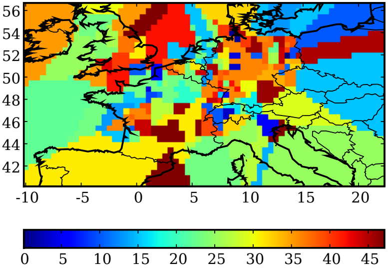
$$\ell_t(\boldsymbol{p}_t) = \ell\big(\boldsymbol{p}_t,\, (y_t^s)_{s \in \mathcal{S}_t}\big) = \sum_{s \in \mathcal{S}_t} \left(\sum_{j=1}^{48} p_{j,t}\, f_{j,t}^s - y_t^s \right)^2$$

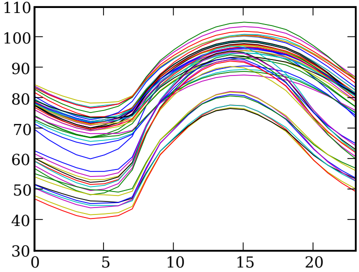
where $\mathcal{S}_t$ is the subset of active stations at day t.

Then, the RMSE equals $\sqrt{\dfrac{\sum_{t=1}^T \ell_t(\boldsymbol{p}_t)}{\sum_{t=1}^T |\mathcal{S}_t|}}$

Left: There are several good and useful experts.

Right: Their forecasting profiles are quite different (the experts are not clones the ones of the others!).





Left: Coloring of Europe according to the index of the locally best expert

Right: Average forecasting profiles during a day (averages over time and space)

RMSE / Performance of the experts

Uniform mean	Best expert	Best $\boldsymbol{p}$
24.41	22.43	21.45

RMSE / Performance of the exponentially weighted average strategies
(tuned with optimal parameters in hindsight)

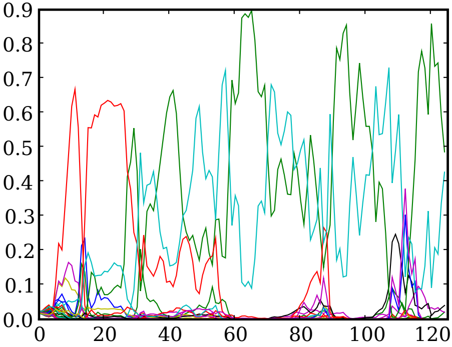
Original version	Fixed history length	Discounted version
21.47	21.37	21.31

The version with fixed history length H only uses the losses encountered in the past H rounds.

The version with discounted losses puts more weight on more recent losses (while still considering all past losses).

Our strategies do not focus on a single expert.

The weights associated with the experts can change quickly and significantly over time (which illustrates in passing that the performance of the considered experts varies over time).



Convex weight vectors output by the exponentially weighted average strategy.

Comments and extensions

The good performance is preserved for the data-driven meta-strategies.

Other strategies are useful and exhibit good performance (e.g., ridge regression or sequential LASSO).

We should also compare our robust methods to more classical stochastic ones (e.g., BMA or CART / random forests).

But we are running out of time to detail this... and instead will switch to a different application.

Second study:

Forecasting of the electricity consumption

Joint work with Yannig Goude (EDF R&D) and M.Sc. students (Marie Devaine, Pierre Gaillard); submitted

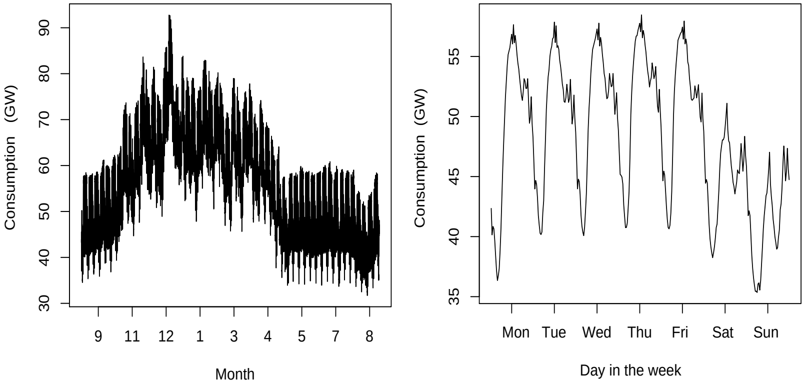
Specialized experts are available: each of them only outputs a forecast when specific conditions are met (working day vs. week end, temperature, etc.).

The definitions and strategies need to be generalized to this setting.

Exhaustive list of references: Blum '97; Freund et al. '97; Cesa-Bianchi and Lugosi '03; Blum and Mansour '07... This is it!

On our data set,

- 3 families of experts, 24 experts in total;
- [operational constraint:] one-day ahead prediction at a half-hour step, i.e., the next 48 half-hour instances are to be predicted every day at noon



Electricity consumption in France

- Year 2007–08 (left)
- Typical summer week (right)

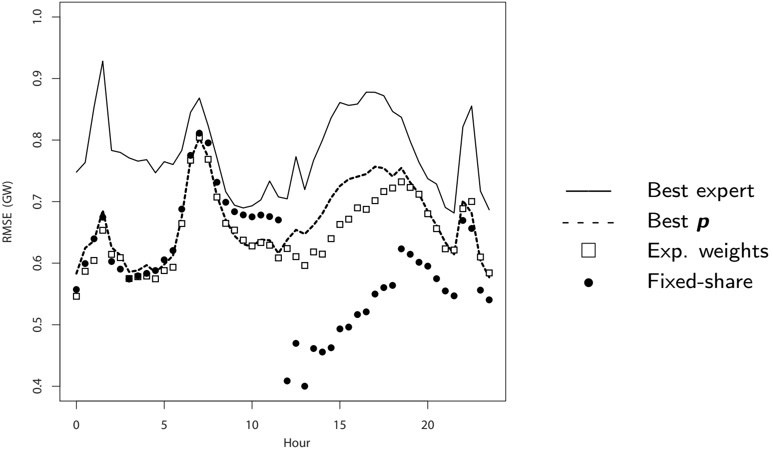
Some orders of magnitude for the prediction problem at hand are indicated below.

Time intervals	Every 30 minutes
Number of days D	320
Time instances T	$15\,360$ $(= 320 \times 48)$
Number of experts N	24 $(= 15 + 8 + 1)$
Median of the y_t	$56\,330$ MW
Bound B on the y_t	$92\,760$ MW

We indicate RMSE (average errors and 95 % standard errors).

Best expert	Uniform mean	Best p
782 ± 10	724 ± 11	658 ± 9
Exp. weights	Best parameter	Adaptive
	629 ± 8	637 ± 9

Shifts	$m = T - 1 = 15\,359$	$m = 200$	$m = 50$
	$223 \pm ?$	$414 \pm ?$	$534 \pm ?$
	Fixed-Share	Best parameter	Adaptive
		599 ± 9	629 ± 8



Average RMSEs (in GW / not in MW) according to the half hours

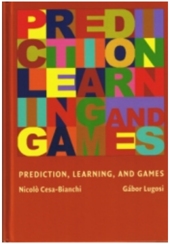
A picture is worth thousand tables, right?

The average RMSE were similar but the behaviors seem different by the half-hours.

References

In case you're not bored to death (yet) by this topic!

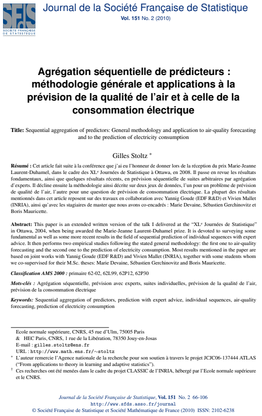
The so-called “red bible!”


PREDICTION, LEARNING, AND GAMES
Nicolò Cesa-Bianchi
Gábor Lugosi

Prediction, Learning, and Games

Nicolò Cesa-Bianchi et Gábor Lugosi

I published a survey paper (containing this talk!) one year ago in the Journal de la Société Française de Statistique



Journal de la Société Française de Statistique
Vol. 151 No. 2 (2010)

Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique

Title: Sequential aggregation of predictors: General methodology and application to air-quality forecasting and to the prediction of electricity consumption

Gilles Stoltz *

Résumé : Cet article fait suite à la conférence que j'ai eu l'honneur de donner lors de la réception du prix Marie-Jeanne Laurent-Duhamel, dans le cadre des XL^e Journées de Statistique à Ottawa, en 2008. Il passe en revue les résultats fondamentaux, ainsi que quelques résultats récents, en prévision séquentielle de suites arbitraires par agrégation d'experts. Il décline ensuite la méthodologie ainsi décrite sur deux jeux de données, l'un pour un problème de prévision de qualité de l'air, l'autre pour une question de prévision de consommation électrique. La plupart des résultats mentionnés dans cet article reposent sur des travaux en collaboration avec Yannig Goude (EDF R&D) et Vivien Mallet (INRIA), ainsi qu'avec les stagiaires de master que nous avons co-encadrés : Marie Devaine, Sébastien Gerchinovitz et Boris Mauricette.

Abstract: This paper is an extended written version of the talk I delivered at the "XL^e Journées de Statistique" in Ottawa, 2004, when being awarded the Marie-Jeanne Laurent-Duhamel prize. It is devoted to surveying some fundamental as well as some more recent results in the field of sequential prediction of individual sequences with expert advice. It then performs two empirical studies following the stated general methodology: the first one to air-quality forecasting and the second one to the prediction of electricity consumption. Most results mentioned in the paper are based on joint works with Yannig Goude (EDF R&D) and Vivien Mallet (INRIA), together with some students whom we co-supervised for their M.Sc. theses: Marie Devaine, Sébastien Gerchinovitz and Boris Mauricette.

Classification AMS 2000 : primaire 62-02, 62L99, 62P12, 62P30

Mots-clés : Agrégation séquentielle, prévision avec experts, suites individuelles, prévision de la qualité de l'air, prévision de la consommation électrique

Keywords: Sequential aggregation of predictors, prediction with expert advice, individual sequences, air-quality forecasting, prediction of electricity consumption

---

Ecole normale supérieure, CNRS, 45 rue d'Ulm, 75005 Paris
& HEC Paris, CNRS, 1 rue de la Libération, 78350 Jouy-en-Josas
E-mail : `gilles.stoltz@ens.fr`
URL : `http://www.math.ens.fr/~stoltz`

[*] L'auteur remercie l'Agence nationale de la recherche pour son soutien à travers le projet JCJC06-137444 ATLAS ("From applications to theory in learning and adaptive statistics").

[†] Ces recherches ont été menées dans le cadre du projet CLASSIC de l'INRIA, hébergé par l'Ecole normale supérieure et le CNRS.

Journal de la Société Française de Statistique, **Vol. 151** No. 2 66-106
`http://www.sfds.asso.fr/journal`

Even better (or worse)—it is in French!