Prévision séquentielle déterministe par agrégation de prédicteurs

Sébastien Gerchinovitz

Institut de Mathématiques de Toulouse, Université Toulouse 3

Introduction

Dans certains problèmes de prévision séquentielle, plusieurs prévisions sont disponibles simultanément. Les techniques d'agrégation séquentielle sont un moyen de combiner ces prévisions.

Plan de l'exposé :

1. Cadre de l'agrégation
2. Un algorithme et ses garanties théoriques
3. Différentes applications

Première formulation du problème

Enjeu : un statisticien cherche à prévoir tour après tour les valeurs d'une suite $y_1, y_2, \ldots \in \mathcal{Y}$. Ses prévisions sont notées $\widehat{a}_1, \widehat{a}_2, \ldots \in \mathcal{D}$ ($\mathcal{D}$ convexe).

Agrégation séquentielle : à chaque date $t \geqslant 1$, le statisticien dispose de K prévisions fondamentales $a_{1,t}, a_{2,t}, \ldots, a_{K,t} \in \mathcal{D}$ qu'il peut combiner pour choisir sa prévision $\widehat{a}_t$.

Mesure de qualité des prévisions : fonction de perte $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}_+$.
L'objectif du statisticien est de prévoir sur le long terme presque aussi bien que le meilleur des prédicteurs frondamentaux, i.e., minimiser le regret

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \ .$$

Prévision de suites individuelles

Nature des observations y_t et des prédicteurs fondamentaux $a_{i,t}$

- Cadre méta-statistique : prédicteurs = méthodes statistiques associées à des modèles différents.
- Cadre déterministe : prédicteurs = modèles physiques (ex : prévision de pics d'ozone, d'électricité).
- Cadre antagoniste : observations et prédicteurs réagissant aux prévisions du statisticien (ex : finance, détection de spams).

Théorie des suites individuelles

Contrairement au cadre statistique classique, pas d'hypothèse stochastique sur la suite $(y_t)_{t \geqslant 1}$. On cherche des garanties pour chacune de ces suites.

Prévision de suites individuelles

Nature des observations y_t et des prédicteurs fondamentaux $a_{i,t}$

- Cadre méta-statistique : prédicteurs = méthodes statistiques associées à des modèles différents.
- Cadre déterministe : prédicteurs = modèles physiques (ex : prévision de pics d'ozone, d'électricité).
- Cadre antagoniste : observations et prédicteurs réagissant aux prévisions du statisticien (ex : finance, détection de spams).

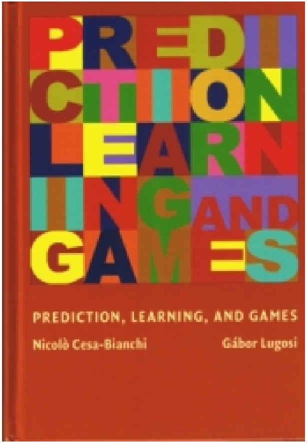
Théorie des suites individuelles

Contrairement au cadre statistique classique, pas d'hypothèse stochastique sur la suite $(y_t)_{t \geqslant 1}$. On cherche des garanties pour chacune de ces suites.

Les 3 cadres ci-dessus sont couverts par la théorie.

Ouvrage de référence

Pour ceux intéressés, je recommande vivement l'ouvrage :
Prediction, learning, and games, Cesa-Bianchi et Lugosi (2006).



Objectif : minimiser le regret

La perte cumulée du statisticien peut se décomposer comme suit :

$$\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) = \underbrace{\min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t)}_{\sim \text{ erreur d'approximation}} + \underbrace{\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t)}_{\text{regret } \sim \text{ erreur d'estimation}} .$$

Dans la suite, on s'attache à minimiser le regret uniformément en toutes les suites $y_1, y_2, \ldots \in \mathcal{Y}$ et $(a_{i,1})_{1 \leqslant i \leqslant K}, (a_{i,2})_{1 \leqslant i \leqslant K}, \ldots \in \mathcal{D}^K$. Typiquement :

$$\sup_{\substack{\boldsymbol{y_t} \in \mathcal{Y} \\ \boldsymbol{a_{i,t}} \in \mathcal{D}}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) - \min_{1 \leqslant i \leqslant K} \frac{1}{T} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) \right\} \leqslant o(1) \quad \text{quand} \quad T \to +\infty .$$

Pour résumer

A chaque date $t \in \mathbb{N}^*$,

1. Les prévisions fondamentales $a_{1,t}, \ldots, a_{K,t} \in \mathcal{D}$ sont accessibles.
2. Le statisticien formule sa prévision $\widehat{a}_t \in \mathcal{D}$ à l'aide des prévisions fondamentales courantes $a_{i,t}$ et des données passées $(a_{\bullet,s}, y_s)$, $1 \leqslant s \leqslant t-1$.
3. Le statisticien observe $y_t \in \mathcal{Y}$ et il encourt la perte $\ell(\widehat{a}_t, y_t)$.

Objectif : sur le long terme, prévoir presque aussi bien que le meilleur prédicteur fondamental, i.e., minimiser le regret

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \,.$$

Prélude : sur la nécessité de convexifier

Exemple :

- $\mathcal{Y} = \{0,1\}$, $\mathcal{D} = [0,1]$ et $\ell(a,y) = |y-a|$.
- Deux prédicteurs fondamentaux : $a_{1,t} = 0$ et $a_{2,t} = 1 \quad \forall t \geqslant 1$.

Sans convexification : si le statisticien choisit ses prévisions dans $\{0,1\}$, il existe une suite $y_1, \ldots, y_t \in \{0,1\}$ telle que $\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) = T$.

Puisque l'un des 2 experts est correct au moins la moitié du temps :

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \geqslant \frac{T}{2} \,.$$

Ordre de grandeur du regret :
Ce regret croît linéairement en T... Comme on le verra ci-après, il existe des prévisions $\widehat{a}_t = \sum_{i=1}^{2} p_{i,t} a_{i,t} \in [0,1]$ qui assurent un regret en $\mathcal{O}\big(\sqrt{T}\big)$.

Agrégation par pondération exponentielle

Exemple classique d'algorithme séquentiel introduit en *machine learning* par Littlestone et Warmuth (1994) et Vovk (1990).

Algorithme (Prédicteur par pondération exponentielle)

Paramètre : $\eta > 0$

A chaque date $t \geqslant 1$*,*

- *A l'aide des données passées, calculer le vecteur de poids* $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{K,t})$ *donné par*

$$p_{i,t} \triangleq \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell(a_{i,s}, y_s)\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \ell(a_{j,s}, y_s)\right)}, \quad 1 \leqslant i \leqslant K ;$$

- *Combiner les prévisions fondamentales* $a_{1,t}, \ldots, a_{K,t} \in \mathcal{D}$ *via*

$$\widehat{a}_t \triangleq \sum_{j=1}^{K} p_{i,t}\, a_{i,t} \in \mathcal{D} .$$

Une première garantie théorique

Théorème (Cesa-Bianchi 1999)

Hypothèses *:* $a \mapsto \ell(a, y)$ *convexe et* $|\ell| \leqslant B$.

Garantie *: pour tout* $T \in \mathbb{N}^*$ *et toute suite de prévisions fondamentales* $a_{i,t} \in \mathcal{D}$ *et d'observations* $y_t \in \mathcal{Y}$,

$$\begin{aligned}\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) &\leqslant \frac{\ln K}{\eta} + \frac{\eta T B^2}{2} \\ &\leqslant B\sqrt{2T \ln K}\,,\end{aligned}$$

où la dernière borne est obtenue pour le choix $\eta = B^{-1}\sqrt{2\ln(K)/T}$.

Peut-on améliorer la borne ?

Non, pas dans le pire des cas. D'après les travaux de Cesa-Bianchi et al. (1997, 2005), la vitesse $\sqrt{T \ln K}$ est optimale au sens minimax.

Une première garantie théorique

Théorème (Cesa-Bianchi 1999)

Hypothèses *:* $a \mapsto \ell(a,y)$ *convexe et* $|\ell| \leqslant B$.

Garantie *: pour tout* $T \in \mathbb{N}^*$ *et toute suite de prévisions fondamentales* $a_{i,t} \in \mathcal{D}$ *et d'observations* $y_t \in \mathcal{Y}$,

$$\begin{aligned} \sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) &\leqslant \frac{\ln K}{\eta} + \frac{\eta T B^2}{2} \\ &\leqslant B\sqrt{2T \ln K}\ , \end{aligned}$$

où la dernière borne est obtenue pour le choix $\eta = B^{-1}\sqrt{2\ln(K)/T}$.

Comment calibrer η si on ne connaît pas B et T ?

On recourt à des techniques de calibration séquentielle :

$$\eta_t \approx {B_t}^{-1}\sqrt{2(\ln K)/t} \qquad \text{où} \quad B_t \triangleq \max_{1 \leqslant s \leqslant t-1} \max_i \left|\ell(a_{i,s}, y_s)\right|\ .$$

Preuve

La preuve repose sur le contrôle de $\ln(W_{T+1}/W_1)$, où

$$W_t \triangleq \frac{1}{K}\sum_{i=1}^{K} e^{-\eta L_{i,t-1}} \qquad \text{avec} \quad L_{i,t} \triangleq \sum_{s=1}^{t} \ell(a_{i,s}, y_s) \ .$$

Minoration :

$$\ln \frac{W_{T+1}}{W_1} \geqslant -\eta \min_{1 \leqslant i \leqslant K} L_{i,T} - \ln K \ .$$

Majoration :

$$\begin{aligned}\ln \frac{W_{T+1}}{W_1} &= \sum_{t=1}^{T} \ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta \ell(a_{i,t}, y_t)}\right) \\ &\leqslant -\eta \sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) + T\eta^2 B^2/2 \qquad \text{(lemme d'Hoeffding \& convexité)} \ .\end{aligned}$$

On conclut en réarrangeant les termes et en divisant par η.

Cas des pertes exp-concaves

Déf. La fonction de perte $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}_+$ est dite η_0-exp-concave en son premier argument ssi $a \in \mathcal{D} \mapsto e^{-\eta_0 \ell(a,y)}$ est concave pour tout $y \in \mathcal{Y}$.

Exemple : $\mathcal{D} = \mathcal{Y} = [-B, B]$, $\ell(a, y) = (y - a)^2$ est $1/(8B^2)$-exp-concave.

Théorème (Kivinen et Warmuth 1999)

Hypothèse *: $\ell(\cdot, \cdot)$ est η_0-exp-concave en son premier argument.*

Garantie *: pour tout $T \in \mathbb{N}^*$ et toute suite de prévisions fondamentales $a_{i,t} \in \mathcal{D}$ et d'observations $y_t \in \mathcal{Y}$,*

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta}$$

pourvu que $\eta \leqslant \eta_0$.

Cas des pertes exp-concaves

Déf. La fonction de perte $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}_+$ est dite η_0-exp-concave en son premier argument ssi $a \in \mathcal{D} \mapsto e^{-\eta_0 \ell(a,y)}$ est concave pour tout $y \in \mathcal{Y}$.

Exemple : $\mathcal{D} = \mathcal{Y} = [-B, B]$, $\ell(a, y) = (y - a)^2$ est $1/(8B^2)$-exp-concave.

Théorème (Kivinen et Warmuth 1999)

Hypothèse *: $\ell(\cdot,\cdot)$ est η_0-exp-concave en son premier argument.*

Garantie *: pour tout $T \in \mathbb{N}^*$ et toute suite de prévisions fondamentales $a_{i,t} \in \mathcal{D}$ et d'observations $y_t \in \mathcal{Y}$,*

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta}$$

pourvu que $\eta \leqslant \eta_0$.

Calibration de η : on adapte séquentiellement η_t en fonction des données.

Vers une borne à coloration PAC-Bayésienne

Rappel (cas d'une perte ℓ exp-concave) :

$$\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) \leqslant \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) + \frac{\ln K}{\eta} \ .$$

Plaçons maintenant des poids a priori π_i sur les prédicteurs fondamentaux :

$$p_{i,t} \triangleq \frac{\pi_i \exp\left(-\eta \sum_{s=1}^{t-1} \ell(a_{i,s}, y_s)\right)}{\sum_{j=1}^{K} \pi_j \exp\left(-\eta \sum_{s=1}^{t-1} \ell(a_{j,s}, y_s)\right)} \ , \quad 1 \leqslant i \leqslant K \ .$$

La borne de regret devient :

$$\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) \leqslant \min_{1 \leqslant i \leqslant K} \left\{ \sum_{t=1}^{T} \ell(a_{i,t}, y_t) + \frac{1}{\eta} \ln \frac{1}{\pi_i} \right\} \ .$$

Vers une borne à coloration PAC-Bayésienne (2)

Rappel (cas d'une perte ℓ exp-concave) :

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) \leqslant \min_{1 \leqslant i \leqslant K} \left\{ \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) + \frac{1}{\eta} \ln \frac{1}{\pi_i} \right\} .$$

On peut obtenir une borne plus fine via la divergence de Kullback-Leibler.
Rappel : pour 2 vecteurs de poids $\boldsymbol{p} = (p_1, \ldots, p_K)$ et $\boldsymbol{q} = (q_1, \ldots, q_K)$,

$$\mathcal{K}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{i=1}^{K} p_i \ln \frac{p_i}{q_i} .$$

On a la borne plus fine suivante (saveur PAC-Bayésienne, cf. les travaux de Catoni 1999, 2004 dans la communauté statistique) :

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) \leqslant \min_{\boldsymbol{\rho}=(\rho_1, \ldots, \rho_K)} \left\{ \sum_{i=1}^{K} \rho_i \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) + \frac{\mathcal{K}(\boldsymbol{\rho}, \pi)}{\eta} \right\} .$$

Extension au cas continu

Si l'ensemble des prédicteurs fondamentaux Θ est continu (e.g., $\Theta = \mathbb{R}^d$), on peut généraliser le prédicteur par pondération exponentielle :

$$p_t(d\theta) \triangleq \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell(a_{\theta,s}, y_s)\right) \pi(d\theta)}{\int_\Theta \exp\left(-\eta \sum_{s=1}^{t-1} \ell(a_{\theta',s}, y_s)\right) \pi(d\theta')} .$$

Sous les mêmes hypothèses d'exp-concavité de ℓ, la borne de regret s'écrit :

$$\sum_{t=1}^{T} \ell\left(\widehat{a}_t, y_t\right) \leqslant \inf_{\rho \in \mathcal{M}_1^+(\Theta)} \left\{ \int_\Theta \sum_{t=1}^{T} \ell\left(a_{\theta,t}, y_t\right) \rho(d\theta) + \frac{\mathcal{K}(\rho, \pi)}{\eta} \right\} .$$

Régression linéaire séquentielle (perte carrée, $\Theta = \mathbb{R}^d$)

Pour de bons choix de la loi a priori π, cela permet de contrôler :

- une variante du prédicteur ridge (Vovk, 2001) ;
- un prédicteur favorisant la parcimonie (Gerchinovitz, 2011).

Régression linéaire séquentielle

Protocole de prévision

A chaque date $t \in \mathbb{N}^*$,

entrée		prévisions fondamentales		prév. du statisticien		observation
$x_t \in \mathcal{X}$	$\rightsquigarrow$	$\boldsymbol{\varphi}(x_t) = \big(\varphi_j(x_t)\big)_{1 \leqslant j \leqslant d} \in \mathbb{R}^d$	$\rightsquigarrow$	$\widehat{y}_t \in \mathbb{R}$ ex : $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{\varphi}(x_t)$	$\rightsquigarrow$	$y_t \in \mathbb{R}$

Objectif : obtenir des garanties de la forme (avec un regret $\Delta_{T,d}(\boldsymbol{u})$ petit)

$$\sum_{t=1}^{T} \big(y_t - \widehat{y}_t\big)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + \Delta_{T,d}(\boldsymbol{u}) \right\} .$$

On cherche des bornes qui valent pour toute suite déterministe $(x_t, y_t)_{1 \leqslant t \leqslant T}$.
Cela conduit à des algorithmes séquentiels robustes.

L'algorithme *ridge* séquentiel

L'algorithme *ridge*, initialement étudié par Hoerl et Kennard (1970) en statistique, a été étendu au cadre déterministe séquentiel par Azoury et Warmuth (2001) et Vovk (2001).

Pour un paramètre $\lambda > 0$, l'algorithme *ridge* séquentiel produit à l'instant t la prévision $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \varphi(x_t)$, où

$$\widehat{\boldsymbol{u}}_t \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} \bigl(y_s - \boldsymbol{u} \cdot \varphi(x_s)\bigr)^2 + \lambda \, \|\boldsymbol{u}\|_2^2 + \bigl(\boldsymbol{u} \cdot \varphi(x_t)\bigr)^2 \right\} .$$

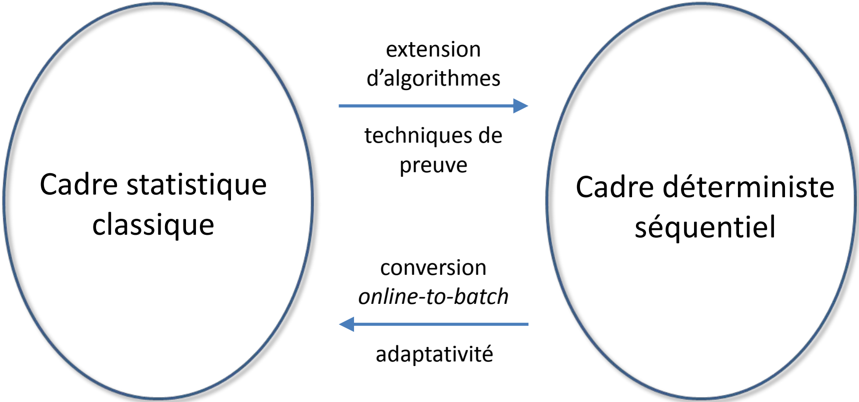
Cet algorithme vérifie, pour toute suite $(x_1, y_1), \ldots, (x_T, y_T) \in \mathcal{X} \times \mathbb{R}$,

$$\sum_{t=1}^{T} \bigl(y_t - \widehat{y}_t\bigr)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \bigl(y_t - \boldsymbol{u} \cdot \varphi(x_t)\bigr)^2 + \lambda \, \|\boldsymbol{u}\|_2^2 + d \, C_y \ln T \right\} + \ldots$$

La vitesse $d \ln T$ correspond à la vitesse paramétrique d/T dans le cadre statistique.

Liens entre les cadres statistique et déterministe

Protocoles de prévision et hypothèses associées radicalement différents, mais liens étroits entre les cadres statistique et déterministe.



Illustrons brièvement ces liens en régression parcimonieuse.

Cadre séquentiel déterministe

Via un algorithme de pondération exponentielle, nous avons obtenu des inégalités oracle de parcimonie déterministes (cf. Gerchinovitz 2011) :

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^{\boldsymbol{d}}} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \varphi(x_t)\big)^2 + \|\boldsymbol{u}\|_0 \, C_y \ln\big(dT \, \|\boldsymbol{u}\|_1\big) \right\} + \dots \quad (1)$$

> Interprétation : même en grande dimension d, il est possible d'encourir un regret faible pour des vecteurs $\boldsymbol{u}$ parcimonieux ($\|\boldsymbol{u}\|_0 \triangleq \big|\{j : u_j \neq 0\}\big|$ petit).

De telles garanties théoriques n'étaient connues que dans un cadre statistique (cf., par ex., Birgé et Massart 2001; Bunea, Tsybakov, et Wegkamp 2007).

Corollaire dans le cadre statistique classique

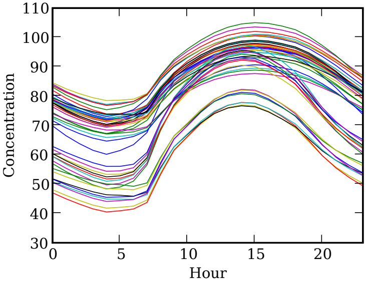
Si $(x_t, y_t)_{1 \leqslant t \leqslant T}$ est aléatoire i.i.d., la borne déterministe (1) implique, par intégration, des garanties statistiques similaires à celles de Dalalyan et Tsybakov (2012).

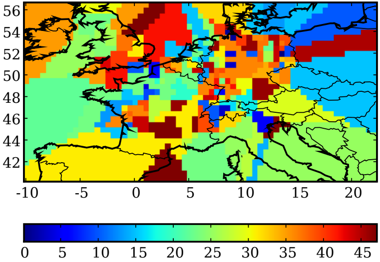
> De surcroît, grâce à des techniques de calibration séquentielle, notre borne de risque est adaptative en la variance inconnue du bruit (gaussien).

Prévision journalière des pics d'ozone

Application initiée par Vivien Mallet (INRIA CLIME) et Gilles Stoltz (ENS).

Enjeu : prévision journalière des pics d'ozone en Europe à partir de 48 simulations numériques simultanées (différents modèles physico-chimiques, différents schémas numériques).





A gauche : profil moyen de concentration d'ozone (en $\mu g/m^3$).

A droite : couleur du meilleur prédicteur local.

Prévision journalière des pics d'ozone (2)

Cadre : régression linéaire séquentielle.

Apport des techniques d'agrégation
Erreurs quadratiques moyennes sur une année, en $\mu g/m^3$:

Moyenne des prédicteurs fondamentaux	24.41
Meilleur prédicteur fondamental	22.43
Algorithme de l'exponentielle des gradients	21.47
Algorithme ridge escompté	19.45

Voir l'article de Mallet, Mauricette, et Stoltz (2009).

Applications actuelles des techniques d'agrégation

Problèmes déjà traités en France :

- Prévision des pics d'ozone journaliers (équipe INRIA CLIME, Vivien Mallet et Gilles Stoltz).
- Prévision de la consommation électrique (EDF R&D Clamart, Yannig Goude et Gilles Stoltz).
- Prévision pour la production d'hydrocarbures (IFP Energies nouvelles, Sébastien Da Veiga et Gilles Stoltz).

Applications théoriques des techniques d'agrégation :

Obtention de méthodes adaptatives. Les techniques de calibration séquentielle permettent une adaptation en la variance inconnue du bruit.

Conclusion

Cadre séquentiel déterministe

Les techniques de suites individuelles permettent de combiner séquentiellement des prévisions élémentaires de façon robuste (déterministe). Inclut les cadres méta-statistique, déterministe et antagoniste.

Techniques

- Agrégation (mélange) de prédicteurs.
- Liens étroits avec le cadre statistique classique.
- Calibration *automatique* des paramètres possible.

Applications

- Plusieurs applications en France (ozone, électricité, hydrocarbures).
- D'autres suggestions d'applications ?

Bibliographie I

K. S. Azoury et M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3) :211–246, 2001.

L. Birgé et P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3 :203–268, 2001.

F. Bunea, A. B. Tsybakov, et M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4) :1674–1697, 2007.

O. Catoni. Universal aggregation rules with exact bias bounds. Technical Report PMA-510, Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris, 1999.

O. Catoni. *Statistical learning theory and stochastic optimization*. Springer, New York, 2004.

N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *J. Comput. System Sci.*, 59(3) :392–411, 1999.

N. Cesa-Bianchi et G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, et M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3) :427–485, 1997.

N. Cesa-Bianchi, G. Lugosi, et G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Trans. Inform. Theory*, 51(6), 2005.

A. Dalalyan et A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3) : 914–944, 2012.

Bibliographie II

S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *JMLR Workshop and Conference Proceedings*, 19 (COLT 2011 Proceedings) :377–396, 2011.

A. E. Hoerl et R. W. Kennard. Ridge regression : biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970.

J. Kivinen et M. K. Warmuth. Averaging expert predictions. In *Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT'99)*, pages 153–167, 1999.

N. Littlestone et M. K. Warmuth. The weighted majority algorithm. *Inform. and Comput.*, 108 :212–261, 1994.

V. Mallet, B. Mauricette, et G. Stoltz. Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research*, 114(D05307), 2009.

V. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT'90)*, pages 371–383, 1990.

V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69 :213–248, 2001.