Binarsity: Prédiction en grande dimension via la sparsité induite par la binarisation de variables

Mokhtar Zahdi Alaya[1], Stéphane Gaiffas[2] , Agathe Guilloux[3]

47^{e} Journées de Statistique - Lille, 2 juin 2015

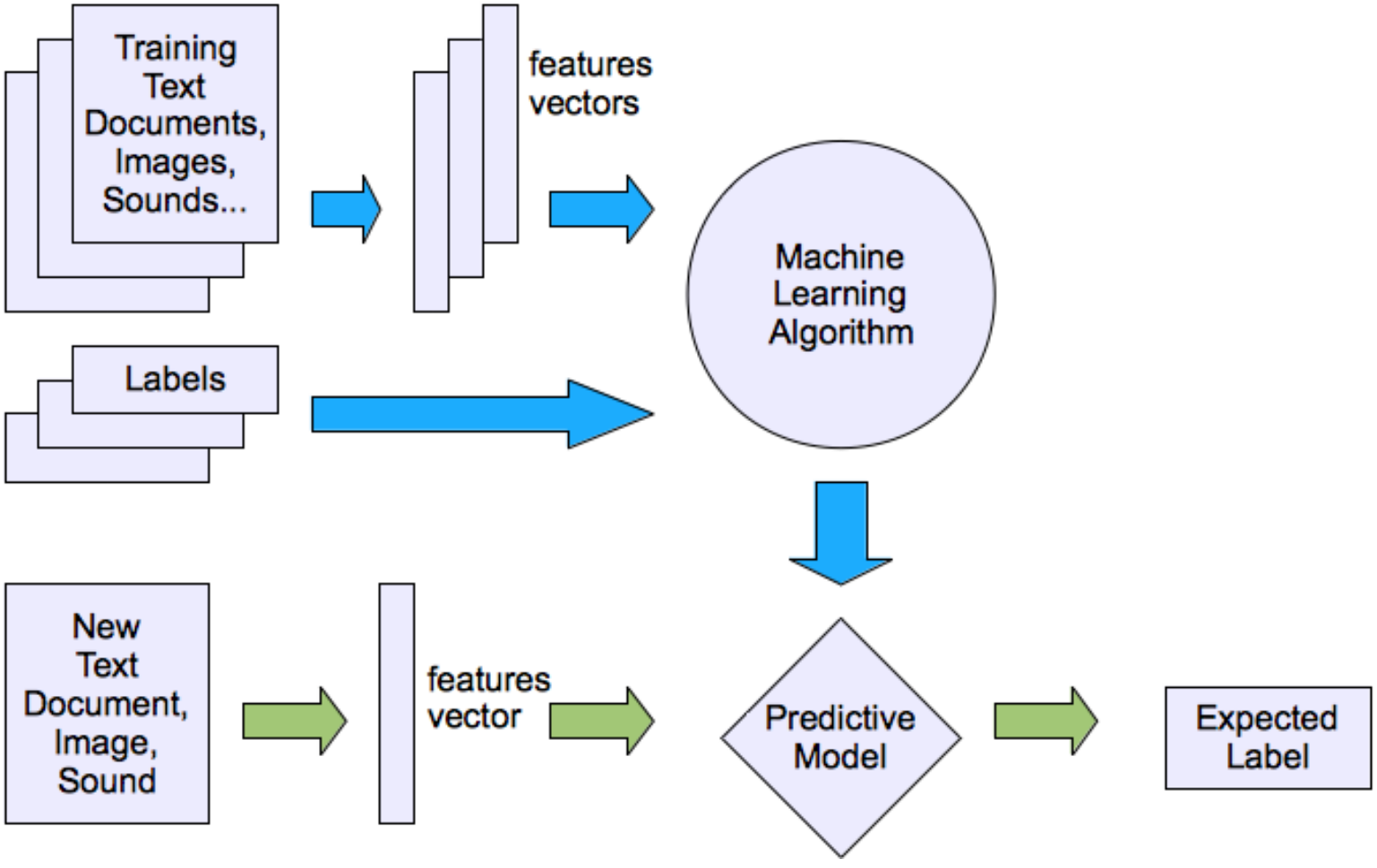
[1]LSTA - UPMC
[2]CMAP – Ecole Polytechnique
[3]LSTA – UPMC

Plan

Statistique en grande dimension: Notation

- Labels $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_n]^\top \in \mathcal{Y}$, $\mathcal{Y} \subset \mathbb{R}$ ou $(\mathcal{Y} \subset \{0, 1\})$.
- Matrice de features $\mathbf{X} = [\mathbf{X}_{\bullet,1}, \ldots, \mathbf{X}_{\bullet,p}]$.
- $\mathbf{X}_{\bullet,j} \in \mathbb{R}^p$ est la j-ème feature.
- $p >> n$.
- $D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ copies $i.i.d.$, à valeurs dans $\mathbb{R}^p \times \mathcal{Y}$.

Statistique en grande dimension: Apprentissage supervisé



Binarisation de variables: Définition

- Si $\mathbf{X}_{\bullet,j}$ prend des valeurs discrètes dans un ensemble de modalités $\{1, \ldots, M_j\}$, on pose $d_j = M_j$ et

$$\mathbf{X}^B_{i,j,k} = \begin{cases} 1 & \text{if } \mathbf{X}_{i,j} = k \\ 0 & \text{sinon} \end{cases}$$

- Si $\mathbf{X}_{\bullet,j}$ est quantitative, on considère une partition d'intervalle $I_{j,1}, \ldots, I_{j,d_j}$, tels que pour tout $k = 1, \ldots, d_j$, $I_{j,k} = \left[q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right)$ avec $q_j(\alpha)$ le quantile d'ordre α de $\mathbf{X}_{\bullet,j}$

$$\mathbf{X}^B_{i,j,k} = \begin{cases} 1 & \text{if } \mathbf{X}_{i,j} \in I_{j,k} \\ 0 & \text{sinon.} \end{cases}$$

Binarisation de variables: Motivation

- Exemple

$$\mathbf{X}_{\bullet,j} = [0.5, 9, 3, -2, 4, 11]^\top \quad n = 6 \quad d_j = 3.$$

Alors,

$$\mathbf{X}^B_{\bullet,j} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- le passage de $\mathbf{X}$ à $\mathbf{X}^B$ s'appelle *binarisation.*
- L'idée est qu'en 'éclatant' une variable en plusieurs variables binaires, on obtient une meilleure attache aux données, par une réponse non-linaire par rapport aux variables d'origine.

Binarsity: Pénalité via la binarisation

- A chaque variable binarisée $\mathbf{X}^B_{\bullet,j,k}$ correspond un coefficient $\theta_{j,k}$.
- Paramètre de binarisation: $\theta_{j,\bullet} = [\theta_{j,1} \cdots \theta_{j,d_j}]^\top$.
- On considère la concaténation de ces vecteurs en un vecteur de taille d.

$$\theta = [\theta_{1,\bullet}^\top \cdots \theta_{p,\bullet}^\top]^\top = \left[\theta_{1,1} \cdots \theta_{1,d_1}\, \theta_{2,1} \cdots \theta_{2,d_2} \cdots \theta_{p,1} \cdots \theta_{p,d_p}\right]^\top.$$

- $d = \sum_{j=1}^{p} d_j$.

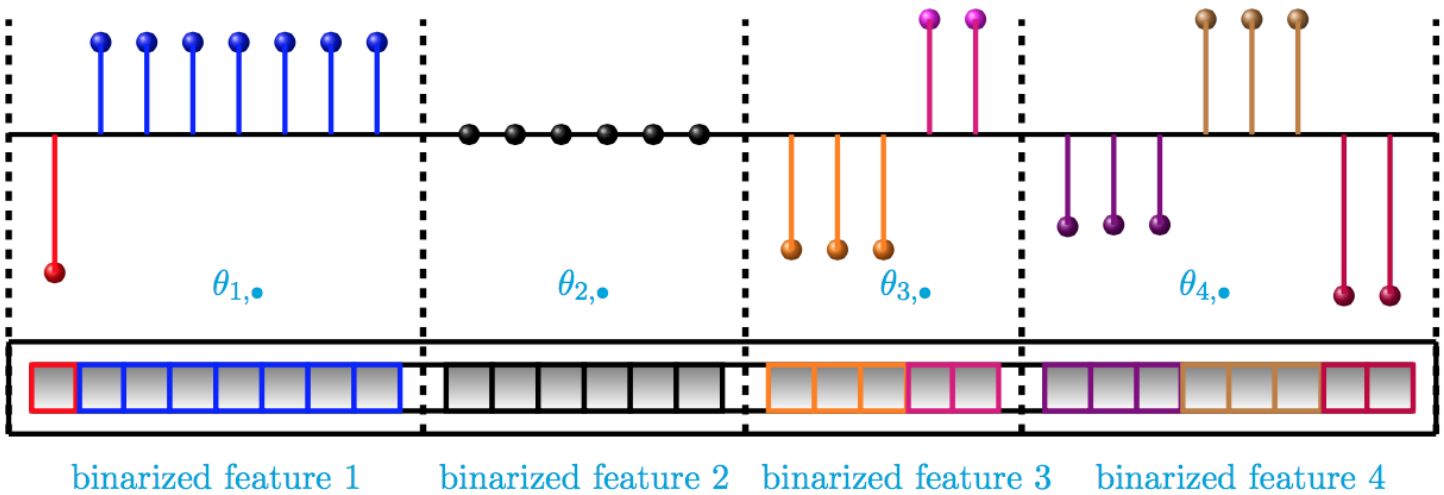


Fig. 1: $p = 4, d_1 = 8,\ d_2 = 6,\ d_3 = 5,\ d_4 = 8$, binarsity$(\theta) = 7$

Binarsity: Pénalité via la binarisation

- Notion de sparsité: Tout bloc $\theta_{j,\bullet}$ peut être nul ou contient un nombre assez petit de valeurs différentes, ce qui est mesuré par la notion de *binarsity*

$$\text{binarsity}(\theta) = \sum_{j=1}^{p} \left(\mathbf{1}_{\theta_{j,1} \neq 0} + \sum_{k=2}^{d_j} \mathbf{1}_{\theta_{j,k} \neq \theta_{j,k-1}} \right).$$

- Si $\mathbf{X}_{j,\bullet}$ est statistiquement non pertinente pour la prédiction, alors le bloc $\theta_{j,\bullet}$ qui lui correspond est nul.
- Si $\mathbf{X}_{j,\bullet}$ est pertinente alors le nombre de valeurs différentes dans le bloc $\theta_{j,\bullet}$ doit être assez petit pour un bon compris biais-variance.

$\|\cdot\|_b$: Relaxation convexe de binarsity

- On définit $\|\cdot\|_b$: une relaxation convexe de binarsity via l'approche de Chandrasekaran et al. (2012).
- Cette approche consiste à considérer la *norme atomique*, $\|\theta\|_{\mathcal{A}}$, obtenue par l'ensemble des atomes $\mathcal{A} \in \mathbb{R}^d$ décrivant binarsity(θ)

$$\|\theta\|_{\mathcal{A}} = \inf\left\{t > 0 : \theta \in t\ \text{conv}(\mathcal{A})\right\},$$

avec conv($\mathcal{A}$) est l'enveloppe convexe de $\mathcal{A}$.

$\|\cdot\|_b$: Relaxation convexe de binarsity

- L'ensemble des atomes de binarsity(θ) est donné par:

$$\mathcal{A} = \Big\{ \pm \left((T_1 e_{1,1})^\top \cdots (T_p e_{p,1})^\top \right)^\top, \ldots, \pm \left((T_1 e_{1,d_1})^\top \cdots (T_p e_{p,d_1})^\top \right)^\top \Big\}$$

- T_j est une $(d_j \times d_j)$ matrice triangulaire inférieure telle que $(T_j)_{r,s} = 0$ si $r < s$ et $(T_j)_{r,s} = 1$ sinon.
- $\left\{ e_{j,k} : k \in \{1, \ldots, d_j\} \right\}$: la base canonique de $\mathbb{R}^{d_j}$.

Lemme

$$\|\theta\|_{\mathcal{A}} = \|\theta\|_b = \sum_{j=1}^{p} \left(|\theta_{j,1}| + \sum_{k=2}^{d_j} |\theta_{j,k} - \theta_{j,k-1}| \right).$$

$\|\cdot\|_{\text{btv},w}$: Version pondérée de $\|\cdot\|_b$

- Dans la suite, on considère une version pondérée notée $\|\cdot\|_{\text{btv},w}$ définie par: $\|\theta\|_{\text{btv},w} = \sum_{j=1}^{p} \|\theta_{j,\bullet}\|_{\text{btv},w_{j,\bullet}}$, avec

$$\|\theta_{j,\bullet}\|_{\text{btv},w_{j,\bullet}} = w_{j,1}|\theta_{j,1}| + \sum_{k=2}^{d} w_{j,k}|\theta_{j,k} - \theta_{j,k-1}|.$$

- $w = [w_{j,\bullet}]_{1\leq j\leq p}$: "Data-driven weights", un vecteur de poids choisi par rapport aux données, tels que

$$w_{j,k} = \lambda\sqrt{\frac{n_{j,k}}{n}}, \quad 1 \leq j \leq p,\, 1 \leq k \leq d_j,$$

avec λ est un paramètre de régularisation dépendant du modèle étudié, et

$$n_{j,k} = \#\left(\left\{i \in \{1,\ldots,n\} : \mathbf{X}_{i,j} \in \bigcup_{r=k}^{d_j} I_{j,r}\right\}\right).$$

Apprentissage avec un scénario de binarsity: Modèle général

- But: prédire un label Y. Pour cela, il nous suffit d'estimer la fonction de régression $x \mapsto \mathbb{E}[Y|X = x]$.
- Procédure d'estimation: approximation par des fonctions constantes par morceaux dans l'espace engendrée par les variables binarisées, $\mathcal{X}^B = \mathrm{span}\{\mathbf{X}^B_{i,\bullet},\, i = 1, \ldots, n\}$.
- Choix d'une fonction de perte ℓ. On écrit le risque empirique:

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(\mathbf{Y}_i, \langle \mathbf{X}^B_{i,\bullet}, \theta \rangle\big).$$

- Problème convexe pénalisé:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ R_n(\theta) + \|\theta\|_{\mathrm{btv},w} \right\}.$$

- Inégalités oracles à vitesse rapide. on utilise une hypothese RE (Restricted Eigenvalue) sur la matrice $\mathbf{X}^B$ (Bickel et al. (2009)).

$$\kappa_{\mathbf{X}^B}(K) = \inf_{u \in \mathbb{R}^d \setminus \{\mathbf{0}_d\} : \, u \in \mathscr{C}_{\text{btv},w}(K)} \left\{ \frac{\|\mathbf{X}^B u\|_2}{\sqrt{n}\|u_K\|_2} \right\} > 0,$$

  avec $\mathscr{C}_{\text{btv},w}(K) = \left\{ u \in \mathbb{R}^d : \|u_{K^c}\|_{\text{btv},w} \leq 3\|u_K\|_{\text{btv},w} \right\}$.
- Pour tout $\theta \in \mathbb{R}^d$, o note par $J(\theta) = \left[J_1(\theta), \ldots, J_p(\theta) \right]$, la concaténation des supports relativement à la pénalité btv tels que $J_j(\theta) = \left\{ \theta_{j,1} \neq 0, \text{ et } \exists\, k = 2, ..., d_j : \theta_{j,k} \neq \theta_{j,k-1} \right\}$.
- Notons que $\text{binarsity}(\theta) = |J(\theta)| = |J_1(\theta)| + \ldots + |J_p(\theta)|$.

Apprentissage avec un scénario de binarsity : Modèle Moindres carrées

- $\mathbf{Y} = f_0(\mathbf{X}) + \varepsilon$, avec $\varepsilon \rightsquigarrow \mathcal{N}(0, \sigma^2 I_n)$
- L'estimateur de f^0 est noté $f_{\hat{\theta}}$, où $\hat{\theta}$ est l'unique solution du problème convexe suivant:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}^B \theta\|_2^2 + \|\theta\|_{\text{btv},w} \right\}.$$

Théorème 1

Supposons que l'hypothèse RE sur $\mathbf{X}^B$ soit vérifiée. Pour tout $A > 0$, fixons

$$\lambda = \sqrt{\frac{8\sigma^2(A + \log d)}{n}}, \text{ et } w_{j,k} = \lambda \sqrt{\frac{n_{j,k}}{n}}.$$

Alors, avec une probabilité supérieure à $1 - 2e^{-A}$, on a

$$\begin{aligned} &\frac{1}{n} \|f_0(\mathbf{X}) - \mathbf{X}^B \hat{\theta}\|_2^2 \\ &\leq \inf_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|f^0(\mathbf{X}) - \mathbf{X}^B \theta\|_2^2 + \frac{608}{\kappa^2_{\mathbf{X}^B}(J(\theta))} |J(\theta)| \max_{j \in [p]} \|w_{j,\bullet}\|_\infty^2 \right\}. \end{aligned}$$

Apprentissage avec un scénario de binarsity: Modèle logistique

- $\mathbf{Y} \in \{0,1\}^n$, et $\pi^0(\mathbf{X}) = \mathbb{E}_{\mathscr{L}(Y|X)}[\mathbf{Y}|\mathbf{X}]$.
- L'estimateur de π^0 est $\pi^{\hat{\theta}}$, avec $\hat{\theta}$ est la solution de problème convexe:

$$\hat{\theta} = \underset{\theta \in B_d(R)}{\operatorname{argmin}} \left\{ R_n(\theta) + \|\theta\|_{\text{btv},w} \right\}.$$

- $R_n(\theta)$ la log-vraisemblance.
- $B_d(R) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$, la boule de rayon $R > 0$.
- Évaluation de la qualité d'estimation en utilisant le risque d'excès

$$\begin{aligned} R(\pi^{\hat{\theta}}) - R(\pi^0) &= \mathbb{E}_{\mathscr{L}(Y|X)}[R_n(\pi^{\hat{\theta}}(\mathbf{X})) - R_n(\pi^0(\mathbf{X}))] \\ &:= KL_n(\pi^0(\mathbf{X}), \pi^{\hat{\theta}}(\mathbf{X})). \end{aligned}$$

- $KL_n(\cdot,\cdot)$: une divergence de Kullback Leibleir empirique.

Théorème 2

Supposons que l'hypothèse RE sur la matrice $\mathbf{X}^B$ soit vérifiée. Fixons $\gamma > 0$. Pour tout $B > 0$, choisissons

$$\lambda = \sqrt{\frac{B + \log d}{n}}, \text{ aet } w_{j,k} = \lambda \sqrt{\frac{n_{j,k}}{n}}.$$

Alors, avec une probabilité supérieure à $1 - 0.25e^{-B}$, on a

$$KL_n(\pi^0(\mathbf{X}), \pi^{\hat{\theta}}(\mathbf{X}))$$
$$\leq (1+\gamma) \inf_{\theta \in B_d(R)} \left\{ KL_n(\pi^0(\mathbf{X}), \pi^{\theta}(\mathbf{X})) + \frac{C(R,\gamma)}{\kappa^2_{\mathbf{X}^B}(J(\theta))} |J(\theta)| \max_{j \in [p]} \|w_{j,\bullet}\|^2_\infty \right\}.$$

Apprentissage avec binarsity scénario: Remarque sur la vitesse

- Dans les Théorèmes 1 et 2, on a un compromis optimal entre la complexité de modèle et l'approximation de la fontion de régression dans $\mathcal{X}^B$.
- Le terme de complexité dépend de la sparsité (binarsity) via le cardinal $|J(\theta)|$ et la constante $\kappa_{\mathbf{X}^B}(J(\theta))$.
- La vitesse de convergence est de l'ordre de $\log d/n$.

Opérateur proximal: Définition

- Soit φ une fonction convexe sur $\mathbb{R}^d$. L'opérateur proximal noté prox_{φ}, est défini par:

$$\text{prox}_{\mu\varphi(v)} := \underset{u\in\mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2}\|v-u\|_2^2 + \mu\varphi(u) \right\}, \forall \mu > 0, \ v \in \mathbb{R}^d.$$

- Les opérateurs proximaux peuvent être interprétés comme des généralisations de la projection car

$$\varphi(u) = \delta_C(u) = \begin{cases} 0, & \text{si } x \in C \\ +\infty, & \text{sinon.} \end{cases} \quad \text{alors} \ \ \text{prox}_{\varphi}(v) = \Pi_C(v).$$

Algorithme FISTA: Rappel

- FISTA: Fast Iterative Shrinkage Thresholding Algorithm.
- Procédure introduite par Beck and Teboulle (2009), qui sert à déterminer

$$\min_{x \in \mathbb{R}^p} F(x) + G(x)$$

où F est une fonction différentiable de gradient Lipschitz et G une fonction avec un opérateur proximal simple à calculer.

Algorithm 1: FISTA

1. Calcul de la constante de Lipschitz L_0 de l'opérateur ∇F.
2. Initialisation: $x_0 \in \mathbb{R}^p$; $v_1 = x_0$; et $t_1 = 1$;
3. **repeat**

$$x_k = \text{prox}_{(1/L)G}\left(v_k - \frac{1}{L}\nabla F(v_k)\right);$$
$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2};$$
$$v_{k+1} = x_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_k - x_{k-1});$$

   until *convergence*
4. **return** x

Algorithme proximale: $\text{prox}_{\text{btv},w}(\cdot)$

- Numériquement, on utilise l'algorithme FISTA pour l'estimation.
- La pénalité $\|\cdot\|_{\text{btv},w}$ est séparable par bloc, donc son prox se calcule par block aussi bien.

Proposition 1

Algorithm 2:

for $j = 1, \ldots, p$ **do**

  set $\beta \leftarrow \theta_{j,\bullet}$; $a \leftarrow w_{j,\bullet}$, $a_{-1} = a \backslash \{a_1\}$;

  $\eta \leftarrow \text{prox}_{\|\cdot\|_{\text{tv},a_{-1}}}(\beta)$;

  $\vartheta_{j,\bullet} \leftarrow \eta - \left(\eta_1 - S_{\frac{a_1}{d_j}}(\eta_1)\right)\mathbf{1}_{d_j}$;

  return $\vartheta_{j,\bullet}$

Calcul de l'estimateur $\pi^{\hat{\theta}}$: $\text{prox}_{\|\cdot\|_{\text{btv},w} + \delta_{B_d(R)}}(\cdot) = ?$

- $\text{prox}_{\text{tv},a_{-1}}(\cdot)$ désigne le prox de la pénalité variation totale pondérée (Alaya et al. (2014)) et S est le soft-thresholding.
- Réécrivons l'estimateur $\hat{\theta}$

$$\hat{\theta} = \underset{\theta \in B_d(R)}{\text{argmin}} \left\{ R_n(\theta) + \|\theta\|_{\text{btv},w} + \delta_{B_d(R)}(\theta) \right\},$$

avec

$$\begin{cases} 0, & \text{si } \theta \in B_d(R) \\ +\infty, & \text{sinon.} \end{cases}$$

- Décomposition de prox,

Proposition 2

$$\text{prox}_{\|\cdot\|_{\text{btv},w} + \delta_{B_d(R)}}(\cdot) = \text{prox}_{\delta_{B_d(R)}} \circ \text{prox}_{\|\cdot\|_{\text{btv},w}}(\cdot).$$

Conclusion

- Présenter une technique de binarisation des variables et la pénalité induite par cette procédure.
- Introduire une relaxation convexe de binarsity.
- Inégalités d'oracles avec une vitesse rapide pour les modèles de régressions moindres carrées et logistique.

Références


M. Z. Alaya, S. Gaïffas, and A. Guilloux. Learning the intensity of time events with change-points. *En Révision pour IEEE Transactions on Information Theory,* 2014.


M. Z. Alaya, S. Gaïffas, and A. Guilloux. Binarsity: features binarization and cuts selection using convex optimization. *Travail en cours,* 2015.


F. Bach. Self-concordance analysis for logistic regression *Electron. J. Statist.,* 4:384-414, 2010.


A. Beck, and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Img. Sci.,* 2(1):183–202, 2009.


P. J. Bickel, Y. Ritov, and , A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig Lasso selector. *Ann. Statist.,* 37(4):1705-1732, 2009.


V., Chandrasekaran, B. Rechet, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Matehamtics,* 12(6), 805-849, 2012.


MERCI.