

EXTENSION DE CART AUX DONNEES CENSUREES

On observe un échantillon iid de v.a. $(Y_i, \delta_i, X_i)_{1 \leq i \leq n}$ de distribution (Y, δ, X) , où

$$\begin{cases} Y &= \inf(T, C) \\ \delta &= \mathbf{1}_{T \leq C} \end{cases}$$

Observation courante Y , données incomplète : $\delta = 0$.

C : variable de censure.

- On cherche $T^* = E[T \mid \delta = 0, Y, \mathbf{X}]$.
- But : trouver un estimateur de T^* , sachant que l'on n'a pas d'observations iid de $T \Rightarrow$ pas de LGN, ...

PROBLEME CAUSE PAR LA CENSURE

- Considérons le problème plus simple d'estimer $t = E[T]$.
- Si j'observe (T_1, \dots, T_n) i.i.d., je peux estimer t par

$$\tilde{t} = \frac{1}{n} \sum_{i=1}^n T_i \rightarrow_{p.s.} t.$$

- Que se passe-t-il si je n'observe que $(Y_1, \delta_1, \dots, Y_n, \delta_n)$?
- Mauvaise idée 1 : $\tilde{t}_1 = \frac{1}{n} \sum_{i=1}^n Y_i.$
- Mauvaise idée 2 : $\tilde{t}_2 = \frac{1}{\sum_{j=1}^n \delta_j} \sum_{i=1}^n \delta_i Y_i.$

ILLUSTRATION PAR L'EXEMPLE

- Exemple naïf : $T \sim \mathcal{E}(\lambda)$, et $D \sim \mathcal{E}(\mu)$, avec T et D indépendants.
- Dans ce cas, \tilde{t}_1 tend vers

$$E[\inf(T, D)] = \frac{1}{\lambda + \mu}.$$

- De plus, \tilde{t}_2 tend vers

$$\frac{E[\delta T]}{E[\delta]} = \frac{1}{\lambda + \mu}$$

- Dans les deux cas, on **sous-estime** la valeur moyenne de T .
- **Solution** : corriger la présence de la censure en essayant de compenser cette sous-estimation.

COMMENT GÉRER LES OBSERVATIONS CENSUREES ?

- *Mauvaise solution* : ne considérer que les observations complètes pour construire l'arbre de décision afin d'estimer la réponse.
→ On sous-estimera la quantité finale...

Cependant, les observations incomplètes donnent également une information biaisée \Rightarrow à corriger !

- **Une solution possible** : surpondérer les sinistres clos avec dével. long pour compenser leur sous-représentation...

\Rightarrow **Question : quels poids ?**

INGREDIENTS : ESTIMATEUR KAPLAN-MEIER ET IPCW

L'algorithme CART peut être adapté ([LMT16]) avec les outils suivants. Hypothèse : T est indépendant de C .

- Soit $\hat{F}(t) = 1 - \prod_{Y_i \leq t} \left(1 - \frac{\delta_i}{\sum_{j=1}^n \mathbf{1}_{Y_j \geq Y_i}}\right)$.
→ Cet estimateur tend vers $F(t) = \mathbb{P}(T \leq t)$.
- **Version additive** : $\hat{F}(t) = \sum_{i=1}^n W_{i,n} \mathbf{1}_{Y_i \leq t}$, avec les poids Kaplan-Meier

$$W_{i,n} = \frac{\delta_i}{n[1 - \hat{G}(Y_i -)]},$$

où $\hat{G}(t)$ est l'estimateur Kaplan-Meier de $G(t) = \mathbb{P}(C \leq t)$.

LES POIDS KAPLAN-MEIER

En regardant bien l'expression des poids KM, on réalise que :

- les observations censurées ont un poids nul, mais leur impact se joue au niveau de la loi de la censure au dénominateur ;
- la somme des poids fait 1 ;
- le poids est d'autant plus grand que Y_i est grand : en effet,

$$\mathbb{P}(C \leq Y_i) \rightarrow 1...$$

- le saut de proba. est lié au nombre d'observations censurées entre 2 observations non censurées !

POURQUOI CA MARCHE ?

① $W_{i,n} = \frac{1}{n} \frac{\delta_i}{1 - \hat{G}(Y_{i-})}$ est "proche" de $W_{i,n}^* = \frac{1}{n} \frac{\delta_i}{1 - G(Y_{i-})}$.

② De plus (LGN),

$$\sum_{i=1}^n W_{i,n}^* \phi(Y_i) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i \phi(Y_i)}{1 - G(Y_{i-})} \rightarrow_{p.s.} E \left[\frac{\delta \phi(Y)}{1 - G(Y-)} \right].$$

Proposition

Pour toute fonction ϕ telle que $E[\phi(T)] < \infty$,

$$E \left[\frac{\delta \phi(Y)}{1 - G(Y-)} \right] = E[\phi(T)].$$

ADAPTATION A NOTRE CONTEXTE

On estime des quantités comme $E[\phi(T, X)]$ (voir equation (2)).

Proposition

Supposons que C est indépendant de (T, X) .

Alors

$$E\left[\frac{\delta\phi(Y, X)}{n(1 - G(Y-))}\right] = E[\phi(T, X)],$$

Et

$$E\left[\frac{\delta\phi(Y, X)}{n(1 - G(Y-))} \mid X\right] = E[\phi(T, X) \mid X].$$

- Donc pour estimer $E[\phi(T, X)]$, on utilise

$$\frac{1}{n} \sum_{i=1}^n \frac{\delta_i \phi(Y_i, X_i)}{1 - \hat{G}(Y_{i-})} = \sum_{i=1}^n W_{i,n} \phi(Y_i, X_i).$$

- Ainsi, pour estimer des quantités comme

$$E[(\Phi(T_i) - a)^2 \mathbf{1}_{X_i \in \mathcal{X}}],$$

où \mathcal{X} est un sous-espace, on calcule

$$\sum_{i=1}^n W_{i,n} (\Phi(Y_i) - a)^2 \mathbf{1}_{X_i \in \mathcal{X}}.$$

QUALITE DE L'ESTIMATEUR CART REPONDERE : SIMULATIONS

- 1 Simuler $n + v$ obs. *iid* $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ de \mathbf{X} , où $\mathbf{X}_i \sim \mathcal{U}(0, 1)$;
- 2 Simuler $n + v$ durées de vie *iid* (T_1, \dots, T_n) de loi exponentielle telle que $T_i \sim \mathcal{E}(\beta = \alpha_1 \mathbb{1}_{\mathbf{x}_i \in [a, b[} + \alpha_2 \mathbb{1}_{\mathbf{x}_i \in [b, c[} + \alpha_3 \mathbb{1}_{\mathbf{x}_i \in [c, d[} + \alpha_4 \mathbb{1}_{\mathbf{x}_i \in [d, e]})$.
(remarquer qu'il existe ainsi 4 sous-groupes dans la population)
- 3 Simuler $n + v$ durées de censure *iid*, telle que $C_i \sim \mathcal{Pareto}(\lambda, \mu)$;
- 4 En déduire la durée de vie observée pour chaque individu, $Y_i = \inf(T_i, C_i)$, et l'indicatrice de censure $\delta_i = \mathbf{1}_{T_i \leq C_i}$;
- 5 Calculer l'estimateur \hat{G} à partir de l'échantillon $(Y_i, \delta_i)_{1 \leq i \leq n+v}$.
- 6 Faire tourner l'algorithme CART en pondérant les calculs de la fonction de perte à chaque division envisagée.

PARAMÈTRES DE SIMULATION

On simule plusieurs échantillons de différentes tailles, puis on construit un arbre CART pondéré que l'on élague ensuite.

TABLE 2
Parameters involved in the simulation scheme.

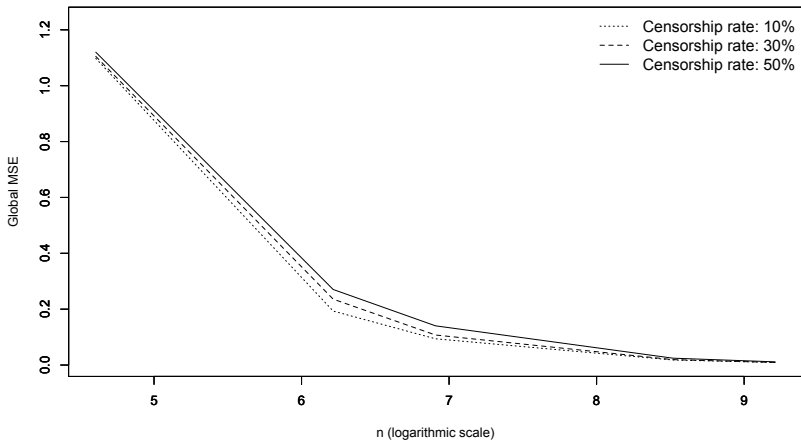
Group-specific means				Component probabilities				Censorship rate		
α_1	α_2	α_3	α_4	$[a, b[$	$[b, c[$	$[c, d[$	$[d, e]$	10%	30%	50%
0.08	0.05	0.16	0.5	$[0, 0.3[$	$[0.3, 0.6[$	$[0.6, 0.8[$	$[0.8, 1]$	(λ, μ)	(λ, μ)	(λ, μ)
12.5	20	6.25	2	30%	30%	20%	20%	(80,1.03)	(20,1.2)	(14,2)

TABLE 3
Descriptive statistics of simulated datasets.

Sample size n	Group-specific exposure				Sample mean
	Group 1	Group 2	Group 3	Group 4	
100	35%	28%	17%	20%	11.08
500	26.8%	31.6%	20%	21.6%	11.37
1 000	30.1%	28.7%	20.6%	20.6%	11.33
5 000	31.42%	29.96%	19.5%	19.12%	11.53
10 000	30.25%	30.19%	19.79%	19.77%	11.52

% of censored observations	Sample size n	Group-specific MWSE				Global MWSE
		Group 1 MWSE	Group 2 MWSE	Group 3 MWSE	Group 4 MWSE	
10%	100	0.19516	0.42008	0.17937	0.30992	<i>1.10454</i>
	500	0.03058	0.07523	0.03183	0.06029	<i>0.19796</i>
	1 000	0.01509	0.03650	0.01517	0.02619	<i>0.09306</i>
	5 000	0.00295	0.00714	0.00289	0.00530	<i>0.01804</i>
	10 000	0.00105	0.00378	0.00117	0.00292	<i>0.00910</i>
30%	100	0.20060	0.43664	0.17448	0.29022	<i>1.10765</i>
	500	0.03736	0.07604	0.04301	0.06584	<i>0.22217</i>
	1 000	0.01748	0.04095	0.01535	0.02674	<i>0.10043</i>
	5 000	0.00319	0.00758	0.00291	0.00547	<i>0.01904</i>
	10 000	0.00117	0.00372	0.00125	0.00292	<i>0.00930</i>
50%	100	0.19784	0.45945	0.17387	0.28363	<i>1.11476</i>
	500	0.04906	0.08993	0.05301	0.06466	<i>0.25668</i>
	1 000	0.02481	0.05115	0.01788	0.03004	<i>0.12387</i>
	5 000	0.00520	0.00867	0.00389	0.00516	<i>0.02299</i>
	10 000	0.00153	0.00407	0.00162	0.00308	<i>0.01057</i>

$$WSE_i = W_{i,n} (\hat{\gamma}_{l(i)} - \pi_0(\mathbf{X}_i))^2, \quad \text{avec } \pi_0(\mathbf{X}_i) = 1/\beta.$$



Quelques articles de référence I



Y. Benjamini and Y. Hochberg.

Controlling the false discovery rate : a practical and powerful approach to multiple testing.

Journal of the Royal Statistical Society, Series B, 57 :289–300, 1995.



Leo Breiman.

Bagging predictors.

Technical report, UC Berkeley, 1994.



J.H. Friedman and P. Hall.

On bagging and nonlinear estimation.

pages –, 2000.



J. Friedman.

Greedy function approximation : A gradient boosting machine.

The Annals of Statistics, 29 :119–139, 2001.

Quelques articles de référence II



Y. Freund and R.E. Shapire.

A decision-theoretic generalization of on-line learning and an application to boosting.

J. Comput. Syst. Sci., 55(1) :119–139, 1997.



Olivier Lopez, Xavier Milhaud, and Pierre-Emmanuel Therond.

Tree-based censored regression with applications in insurance.

Electron. J. Stat., 10 :2685–2716, 2016.



R.E. Shapire and Y. Freund.

Boosting.

MIT Press, Cambridge, 1st edition, 2012.



R.E. Shapire.

The boosting approach to machine learning : An overview.

Technical report, 2003.