Régression logistique sous contrainte avec standardisation en ligne pour flux de données

Benoît Lalloué Jean-Marie Monnez Eliane Albuisson

26èmes Rencontres de la Société Francophone de Classification
Nancy, 3-5 septembre 2019

Contexte

Apprentissage en ligne

- Analyse d'un jeu de données massives ou d'un flux.
- Éviter de stocker les données.
- Mettre à jour les résultats par étapes successives, en prenant en compte de nouvelles données à chaque étape.
- Possibilité : utiliser des algorithmes stochastiques récursifs.
- Exemples :
  - régression linéaire.
  - analyse en composantes principales [1].
  - k-médianes [2].

---

1. Monnez JM, Skiredj A. Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream. *HAL*. 2018

2. Cardot H, Cénac P, Monnez JM. A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics & Data Analysis*. 2012.

Standardisation en ligne des données

- On peut standardiser les données pour :
  - éviter une explosion numérique.
  - appliquer une méthode de pénalisation (par ex. LASSO).
- Problème dans le cas d'un flux de données : moyennes et variances des variables inconnues a priori.
- Possibilité : effectuer une standardisation en ligne.
- Cas étudié pour la régression linéaire, avec meilleures performances que sur données brutes [3].
- On adopte une approche similaire pour la régression logistique.

3. Duarte K, Monnez JM, Albuisson E. Sequential linear regression with online standardized data. *PLOS ONE*. 2018.

Processus de gradient stochastique

Notations

- On observe les réalisations d'un vecteur aléatoire $(R^1, ..., R^p, S)$ dans $\mathbb{R}^p \times \{0, 1\}$.
- Soit :
  - R le vecteur aléatoire $\left(R^1 \ ... \ R^p \ 1\right)'$
  - $m = \left(E\left[R^1\right] \ ... \ E\left[R^p\right] \ 0\right)'$
  - $R^c = R - m$ (r^c réalisation de R^c)
  - σ^k l'écart-type de R^k, $k = 1, ..., p$
  - Γ la matrice de diagonale $\frac{1}{\sigma^1}, ..., \frac{1}{\sigma^p}, 1$ (convention : $\sigma^k = 1$ pour une variable discrète)
  - $Z = \Gamma R^c$ ($z = \Gamma r^c$ réalisation de Z), vecteur R standardisé
  - $\theta = \left(\theta^1 \ ... \ \theta^p \ \theta^{p+1}\right)'$ un vecteur de paramètres réels.

Régression logistique

- Modèle logistique avec variables explicatives standardisées :

$$P(S = s \mid R = r) = f(s; z, \theta) = \frac{e^{z'\theta s}}{1+e^{z'\theta}}.$$

- $E[S \mid R] = h(Z'\theta)$ avec $h(u) = \frac{e^u}{1+e^u}$.
- Fonction de perte : $-\ln f(s; z, x) = -z'xs + \ln\left(1 + e^{z'x}\right)$
- On cherche θ tel que la fonction de coût

$$F(x) = -E[\ln f(S; Z, x)] = E\left[-Z'xS + \ln\left(1 + e^{Z'x}\right)\right]$$

soit minimale.

- θ est l'unique solution de :

$$F'(x) = E\left[-ZS + \frac{Ze^{Z'x}}{1 + e^{Z'x}}\right] = E\left[Z\left(h\left(Z'x\right) - S\right)\right] = 0.$$

Notations (2)

Soit :

- $\left((R_n^1, ..., R_n^p, S_n), n \geqslant 1\right)$ un échantillon i.i.d. de $(R^1, ..., R^p, S)$
- $R_n = \left(R_n^1 \ ... \ R_n^p \ 1\right)', \ n \geqslant 1$
- $R_n^c = R_n - m, \ n \geqslant 1$
- $Z_n = \Gamma R_n^c, \ n \geqslant 1$, vecteur R_n standardisé
- Pour $k = 1, ..., p$:
  $\overline{R}_n^k$ la moyenne de l'échantillon $\left(R_1^k, ..., R_n^k\right)$ de R^k et $\left(V_n^k\right)^2$ sa variance, calculées récursivement
- $\overline{R}_n = \left(\overline{R_n^1} \ ... \ \overline{R_n^p} \ 0\right)'$ et
  Γ_n la matrice de diagonale $\frac{1}{\sqrt{\frac{n}{n-1}} V_n^1}, ..., \frac{1}{\sqrt{\frac{n}{n-1}} V_n^p}, 1.$

Processus de gradient stochastique

- Supposons que m_n observations (R_i, S_i) soient prises en compte à l'étape n, avec $\mu_n = \sum_{i=1}^{n} m_i$, $I_n = \{\mu_{n-1} + 1, ..., \mu_n\}$
- Pour $j \in I_n$, $\widetilde{Z}_j = \Gamma_{\mu_{n-1}} \left(R_j - \overline{R}_{\mu_{n-1}}\right)$
- Supposons que θ appartienne à un sous-ensemble convexe K de $\mathbb{R}^{p+1}$. Soit Π l'opérateur de projection sur K.
- Définissons récursivement les processus d'approximation stochastique (X_n) et $(\overline{X}_n)$:

$$
\begin{aligned}
X_{n+1} &= \Pi\left(X_n - a_n \frac{1}{m_n} \sum_{j \in I_n} \widetilde{Z}_j \left(h\left(\widetilde{Z}_j' X_n\right) - S_j\right)\right), \\
\overline{X}_{n+1} &= \frac{1}{n+1} \sum_{1}^{n+1} X_i.
\end{aligned}
$$

Processus de gradient stochastique (2)

Posons les hypothèses suivantes :

- (H1a) Il n'y a pas de relation affine entre les composantes de R.
- (H1b) Les moments d'ordre 4 de R existent.
- (H2) $a_n > 0$, $\sum_{n=1}^{\infty} a_n = \infty$, $\sum_{n=1}^{\infty} \frac{a_n}{\sqrt{n}} < \infty$, $\sum_{n=1}^{\infty} a_n^2 < \infty$.

Théorème

Supposons que H1a,b et H2 soient vérifiées. Alors (X_n) et $(\overline{X}_n)$ convergent vers θ p.s.

Démonstrations et détails dans : Lalloué, B., Monnez, J.-M., Albuisson, E. Streaming constrained binary logistic regression with online standardized data. Application to scoring heart failure. 2019. *hal-02156324*

Choix du pas a_n

- Choix du pas : crucial pour obtenir de bonnes performances.
- Pas trop petit : convergence trop lente
  Pas trop grand : possibilité d'explosion numérique.
- Plusieurs possibilités :
  - Processus à pas variable : $a_n = \frac{c}{(b+n)^\alpha}$
  - Processus moyennisé à pas constant : $\forall n, a_n = a$ (non adapté ici [4]).
  - Processus moyennisé à pas constant par paliers : $a_n = \frac{c}{(b+\lfloor \frac{n}{\tau} \rfloor)^\alpha}$, avec $\lfloor . \rfloor$ la partie entière et τ la taille des paliers (suggéré par Bach [5]).
  - Ici : $\alpha = 2/3$, $b = 1$ et $c = 1$.

---

4. Bach F, Moulines E. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In : *Advances in Neural Information Processing Systems 26*. 2013.

5. Bach F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*. 2014.

Expérimentations

Processus testés

24 processus testés :

Method type	Abbreviation	Type of data	Number of observations used at each step of the process	Step-size	Levels size	Use of the averaged process
Classic (C)	CR1V	Raw data (R)	1	Variable (V)	-	No
	CR10V		10			
	CR100V		100			
ASGD with piecewise constant step-size (A)	AR1P50		1	Piecewise constant (P)	50	Yes
	AR10P50		10			
	AR100P50		100			
	AR1P100		1		100	
	AR10P100		10			
	AR100P100		100			
	AR1P200		1		200	
	AR10P200		10			
	AR100P200		100			
Classic (C)	CS1V	Online Standardized data (S)	1	Variable (V)	-	No
	CS10V		10			
	CS100V		100			
ASGD with piecewise constant step-size (A)	AS1P50		1	Piecewise constant (P)	50	Yes
	AS10P50		10			
	AS100P50		100			
	AS1P100		1		100	
	AS10P100		10			
	AS100P100		100			
	AS1P200		1		200	
	AS10P200		10			
	AS100P200		100			

Données testées

- Chaque processus est testé sur 6 jeux de données :

Dataset name	N_a	N	p_a	p	Source
Twonorm	7400	7400	20	20	www.cs.toronto.edu/~delve/data/datasets.html
Ringnorm	7400	7400	20	20	www.cs.toronto.edu/~delve/data/datasets.html
Quantum	50000	15798	78	12	derived from www.osmot.cs.cornell.edu/kddcup
Adult2	45222	45222	14	38	derived from www.cs.toronto.edu/~delve/data/datasets.html
EEG	14980	14977	14	14	https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State
HOSPHF30D	21382	21382	29	13	derived from EPHESUS study

N_a: number of available observations; N: number of selected observations; p_a: number of available parameters; p: number of selected parameters.

- Flux de données : simulé par tirage au sort avec remise.
- Enregistrement des valeurs des critères pour des nombres d'observations utilisées (1N à 100N) et des temps de calcul (1 à 120s) fixes.
- Pour chaque jeu de données et point d'enregistrement : classement des processus.
- Comparaison du classement moyen sur tous les jeux de données.

Initialisation

- Tous les processus initialisés avec $X_1 = 0$.
- Initialisation de la standardisation en ligne : première estimation des moyennes et variances avec un échantillon aléatoire de 1000 observations, puis mise à jour à chaque itération.
- Moyennisation : rodage de 1000 itérations (non inclues dans la moyenne).

Critères de convergence

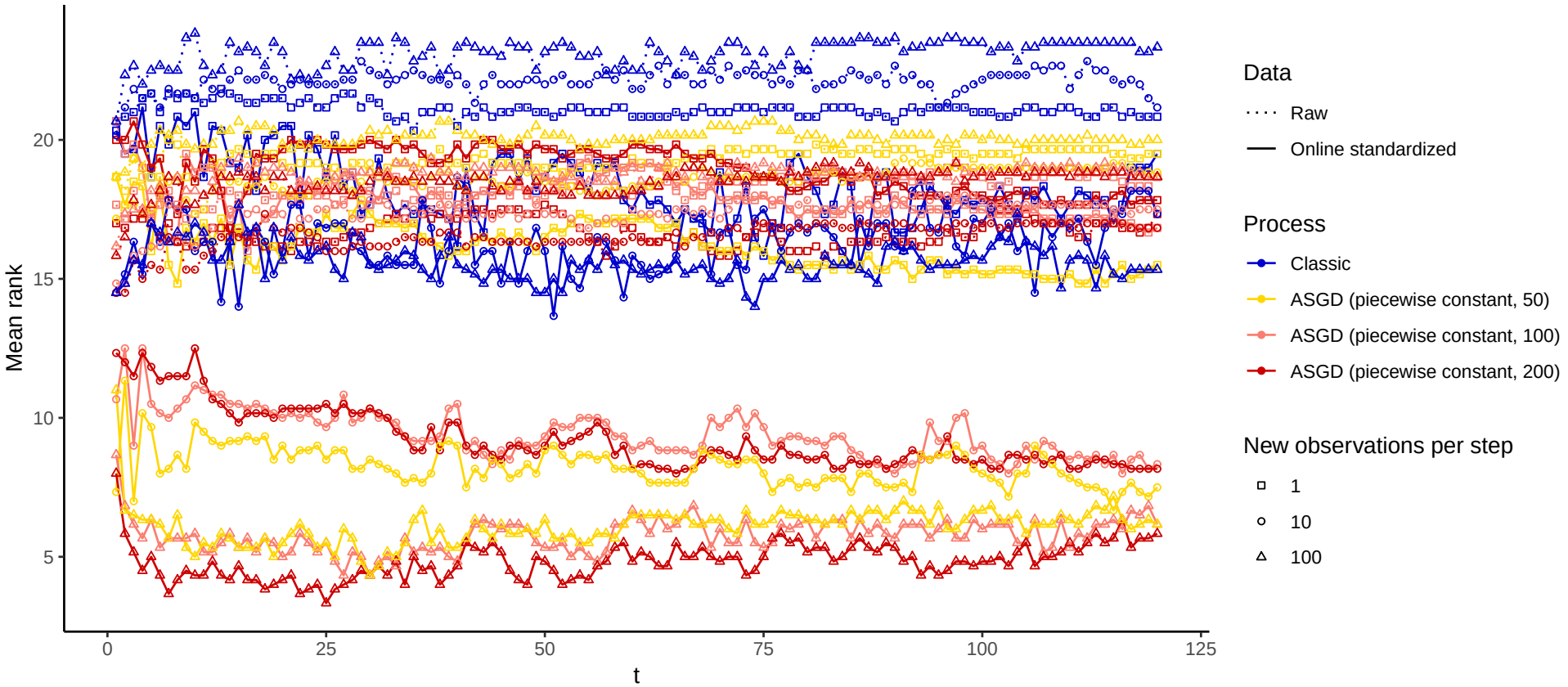
- Régression logistique "classique" (fonction `glm` de R) comme référence : vecteur de coefficients θ^c.
- Soit $\hat{\theta}_{n+1}$ le vecteur estimé obtenu par un processus après n itérations.
- Critères de convergence :
  - Cosinus entre les deux vecteurs : $cos(\theta^c, \hat{\theta}_{n+1}) = \frac{\theta^{c\prime}\hat{\theta}_{n+1}}{\|\theta^c\|\|\hat{\theta}_{n+1}\|}$.
  - Coefficient de corrélation entre les prédictions obtenues par les deux méthodes (non présenté).
  - Rapport $\frac{\hat{F}(\hat{\theta}_{n+1})-\hat{F}(\theta^c)}{\hat{F}(\theta^c)}$
    avec $\hat{F}(\hat{\theta}_{n+1}) = \frac{1}{N}\sum_{i=1}^{N}\left(-r_i'\hat{\theta}_{n+1}s_i + ln(1+e^{r_i'\hat{\theta}_{n+1}})\right)$
    estimation de la fonction de coût F en $\hat{\theta}_{n+1}$ (non présenté).

Comparaison à temps de calcul fixé (60s)

Process	Twonorm	Ringnorm	Quantum	Adult	EEG	HOSPHF30D	Mean rank
CR1V	0.9999	0.9999	0.9709	EXPL	EXPL	EXPL	20.8
CR10V	1.0000	1.0000	0.9683	EXPL	EXPL	EXPL	21.8
CR100V	1.0000	1.0000	0.9659	EXPL	EXPL	EXPL	22.5
AR1P50	1.0000	1.0000	0.9978	EXPL	EXPL	EXPL	19.3
AR10P50	1.0000	1.0000	0.9960	EXPL	EXPL	EXPL	18.3
AR100P50	1.0000	1.0000	0.9948	EXPL	EXPL	EXPL	20.0
AR1P100	1.0000	1.0000	0.9991	EXPL	EXPL	EXPL	18.0
AR10P100	1.0000	1.0000	0.9972	EXPL	EXPL	EXPL	17.5
AR100P100	1.0000	1.0000	0.9959	EXPL	EXPL	EXPL	19.0
AR1P200	1.0000	1.0000	0.9999	EXPL	EXPL	EXPL	16.8
AR10P200	1.0000	1.0000	0.9981	EXPL	EXPL	EXPL	16.5
AR100P200	1.0000	1.0000	0.9970	EXPL	EXPL	EXPL	18.2
CS1V	0.9997	0.9998	0.9987	0.9979	0.9988	0.9898	17.5
CS10V	0.9996	1.0000	0.9989	0.9968	0.9994	0.9932	15.8
CS100V	0.9994	1.0000	0.9992	0.9953	0.9993	0.9840	15.3
AS1P50	0.9999	1.0000	0.9959	0.9964	0.9993	0.9854	17.2
AS10P50	1.0000	1.0000	0.9999	0.9998	0.9997	0.9986	8.2
AS100P50	1.0000	1.0000	0.9999	0.9999	1.0000	0.9999	6.5
AS1P100	0.9999	1.0000	0.9948	0.9888	0.9992	0.9841	19.2
AS10P100	1.0000	1.0000	0.9999	0.9998	0.9996	0.9987	8.8
AS100P100	1.0000	1.0000	0.9999	0.9999	1.0000	0.9999	6.7
AS1P200	0.9999	0.9999	0.9934	0.9823	0.9987	0.9812	19.8
AS10P200	1.0000	1.0000	0.9999	0.9996	0.9996	0.9986	8.2
AS100P200	1.0000	1.0000	0.9999	1.0000	1.0000	0.9999	4.8

EXPL: numerical explosion.

Évolution selon le temps de calcul



- Meilleur processus : moyennisé, pas constant par paliers (200), données standardisées en ligne, 100 nouvelles obs par étape.
- N.B. : les processus sur données brutes conduisent à une explosion numérique pour `Adult`, `EEG`, `HOSPHF30D`.

Conclusion

- Processus de gradient stochastique pour réaliser une régression logistique en ligne.
- Standardisation en ligne pour éviter les explosions numériques (entre autres).
- Les expérimentations confirment l'intérêt de processus moyennisés à pas constant par paliers, sur données standardisées en ligne.
- Utilisation de ce processus dans un score en ligne appliqué à l'insuffisance cardiaque [6].

6. Lalloué, B., J.-M. Monnez, and E. Albuisson. Actualisation en ligne d'un score d'ensemble. *51e Journées de Statistique. hal-02152352.* 2019

Merci de votre attention !