Score de risque d'événement et score en ligne pour des insuffisants cardiaques

Kévin Duarte[*,**], Jean-Marie Monnez[*,**], Eliane Albuisson[***,****,♯]

[*]Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France
[**]CHRU Nancy, INSERM, Université de Lorraine, CIC, Plurithématique, F-54000 Nancy, France
[***]Université de Lorraine, CNRS, IECL, F-54000 Nancy, France
[****]Université de Lorraine, CHRU Nancy, Biobase, Pôle S^2R, F-54000 Nancy, France
[♯]Université de Lorraine, Faculté de Médecine, InSciDenS, F-54000 Nancy, France
k.duarte@chru-nancy.fr, jean-marie.monnez@univ-lorraine.fr,
eliane.albuisson@univ-lorraine.fr

Résumé. On présente la construction d'un score de risque d'événement à court terme pour des insuffisants cardiaques. On suppose ensuite que les données de patients arrivent de façon continue et que l'on veut actualiser en ligne la fonction de score. On étudie en particulier l'estimation en ligne des paramètres d'un modèle de régression linéaire par un processus de gradient stochastique en utilisant des données standardisées en ligne au lieu des données brutes.

1 Introduction

Dans Duarte et al. (2018a), nous présentons une méthodologie de construction d'un score d'événement cardio-vasculaire (décès ou hospitalisation) à court terme pour des insuffisants cardiaques, basé sur un prédicteur d'ensemble dont la qualité est mesurée par l'aire sous la courbe ROC (AUC). Deux règles de classification ont été utilisées, la régression logistique et l'analyse discriminante linéaire, 1000 échantillons bootstrap et un nombre optimal de variables explicatives sélectionnées par un tirage aléatoire selon trois modalités différentes. Ceci donne 6000 prédicteurs différents. Une fonction de score combinaison affine des variables explicatives a été associée à chaque prédicteur. La fonction de score finale, obtenue par une moyennisation en deux étapes, a été ramenée à une échelle de 0 à 100. Nous appliquons cette méthodologie aux données provenant de l'étude EPHESUS (Pitt et al., 2003). Nous présentons dans la suite l'actualisation en ligne de cette fonction de score dans le cas d'un flux de données, en estimant en ligne les paramètres des modèles de régression linéaire (Duarte et al., 2018b) ou logistique (Monnez, 2018).

2 Introduction aux scores en ligne

Supposons maintenant que les données de patients arrivent de façon continue. Le problème est d'actualiser en ligne la fonction de score définie dans le paragraphe précédent. Des

algorithmes stochastiques récursifs peuvent être utilisés pour des observations arrivant séquentiellement, par exemple pour estimer des centres de classes en classification non supervisée (Cardot et al., 2012), ou des composantes principales d'une analyse factorielle (Monnez et Skiredj, 2018), ou des paramètres d'un modèle de régression. Quand on utilise de tels algorithmes, il n'est pas nécessaire de stocker les données et beaucoup plus de données qu'avec les méthodes classiques peuvent être prises en compte pendant la même durée de temps. Nous présentons ici des algorithmes stochastiques utilisables pour l'estimation en ligne des paramètres d'un modèle de régression linéaire, en particulier d'une fonction discriminante linéaire binaire (Duarte et al., 2018b). Le cas de la régression logistique est étudié dans Monnez (2018).

A chaque étape d'un tel algorithme, un lot de nouvelles données est pris en compte et affecté aux échantillons bootstrap en utilisant le bootstrap Poisson (Oza et Russell, 2001). L'ensemble des variables explicatives tirées au hasard étant fixé pour chaque échantillon bootstrap, un prédicteur basé sur l'analyse discriminante linéaire ou sur la régression logistique peut être actualisé en ligne. Donc la fonction de score, obtenue à partir d'un ensemble de 6000 prédicteurs, peut être actualisée en ligne.

3 Régression linéaire en ligne

Soit R $(p,1)$ et S $(q,1)$ deux vecteurs aléatoires. Dans la régression linéaire multidimensionnelle de S par rapport à R, on cherche à déterminer des paramètres matriciels θ (p,q) et η $(q,1)$ qui rendent minimale $\mathbb{E}\left[\|S-\theta' R-\eta\|^2\right]$. Dans le cas d'un flux de données $((R_n,S_n),n\geq 1)$ constituant un échantillon i.i.d. de (R,S), on sera amené à estimer ces paramètres en ligne, en utilisant un processus de gradient stochastique.

Pour éviter une explosion numérique (Pascanu et al., 2012) due en particulier au choix du pas ou à la présence de données extrêmes, nous proposons d'utiliser des données standardisées (centrées réduites). Soit $\mathbb{E}[R]$ le vecteur-espérance mathématique de R, $\mathbb{E}[S]$ celui de S. Soit Γ la matrice diagonale des inverses des écarts-types des composantes de R, Γ^1 celle de S. Soit $R_1=\Gamma(R-\mathbb{E}[R])$, $S_1=\Gamma^1(S-\mathbb{E}[S])$. Déterminons $\theta_1=\Gamma^{-1}\theta\Gamma^1$ qui rend minimale $\mathbb{E}\left[\|S_1-\theta_1' R_1\|^2\right]$. θ_1 est solution d'un système d'équations linéaires $B\theta_1=F$:

$$\nabla_{\theta_1}\mathbb{E}\left[\|S_1-{\theta_1}' R_1\|^2\right]=0\Leftrightarrow\mathbb{E}\left[R_1{R_1}'\right]\theta_1=\mathbb{E}\left[R_1{S_1}'\right].$$

On peut utiliser un processus de gradient stochastique pour estimer en ligne θ_1. Le problème est que l'on ne connaît pas a priori, dans le cas d'un flux de données, les moments de R et S. On estime alors récursivement en ligne les espérances mathématiques et les écart-types des composantes de R et S.

Soit à résoudre un système $B\theta=F$. Le processus de gradient stochastique classique (SGD) (X_n) convergeant vers la solution θ est défini de façon récursive par

$$\begin{aligned} &X_{n+1}=X_n-a_n(B_nX_n-F_n),\ \text{avec :}\\ &\quad\mathbb{E}[B_n]=B,\mathbb{E}[F_n]=F,\\ &\quad a_n>0,\sum_{n=1}^{\infty}a_n=\infty,\sum_{n=1}^{\infty}{a_n}^2<\infty. \end{aligned}\tag{1}$$

Un pas a_n trop grand dans les premières étapes peut entraîner une explosion numérique, un pas a_n trop petit une convergence trop lente.

Le processus ASGD (SGD moyennisé) (Y_n) utilise sous (1) un pas constant a :

$$\begin{aligned} X_{n+1} &= X_n - a(B_n X_n - F_n), \\ Y_{n+1} &= \frac{1}{n+1}\sum_{i=1}^{n+1} X_i. \end{aligned}$$

On constate que dans notre cas l'hypothèse (1) n'est pas vérifiée, puisqu'on ne dispose pas d'un échantillon i.i.d. de (R_1, S_1).

Supposons alors qu'un lot de m_n données (R_i, S_i) arrive à l'étape n et notons

$$\begin{aligned} M_n &= \sum_{i=1}^{n} m_i, \\ I_n &= \{M_{n-1}+1, ..., M_n\}. \end{aligned}$$

Notons Γ_n, respectivement Γ_n^1, la matrice diagonale des inverses des estimations des écarts-types des composantes de R, respectivement S, calculée récursivement à partir des données (R_i, S_i), $i \leq n$. Deux définitions de B_n et F_n sont utilisées dans cette étude (utilisation du lot de données arrivant ou de toutes les données (R_i, S_i), $i \leq n$) :

Définition 1)

$$\begin{aligned} B_n &= \Gamma_{M_{n-1}} \left(\frac{1}{m_n} \sum_{j \in I_n} \left(R_j - \bar{R}_{M_{n-1}}\right) \left(R_j - \bar{R}_{M_{n-1}}\right)' \right) \Gamma_{M_{n-1}}, \\ F_n &= \Gamma_{M_{n-1}} \left(\frac{1}{m_n} \sum_{j \in I_n} \left(R_j - \bar{R}_{M_{n-1}}\right) \left(S_j - \bar{S}_{M_{n-1}}\right)' \right) \Gamma^1_{M_{n-1}} \end{aligned}$$

avec $\bar{R}_{M_{n-1}} = \dfrac{1}{M_{n-1}} \displaystyle\sum_{i=1}^{M_{n-1}} R_i$, $\bar{S}_{M_{n-1}} = \dfrac{1}{M_{n-1}} \displaystyle\sum_{i=1}^{M_{n-1}} S_i$.

Définition 2)

$$\begin{aligned} B_n &= \Gamma_{M_n} \left(\frac{1}{M_n} \sum_{i=1}^{n} \sum_{j \in I_i} R_j {R_j}' - \bar{R}_{M_n} (\bar{R}_{M_n})' \right) \Gamma_{M_n}, \\ F_n &= \Gamma_{M_n} \left(\frac{1}{M_n} \sum_{i=1}^{n} \sum_{j \in I_i} R_j {S_j}' - \bar{R}_{M_n} \bar{S}'_{M_n} \right) \Gamma^1_{M_n}. \end{aligned}$$

Dans le cas 2), les variables aléatoires R_j et S_j n'étant pas centrées, on peut éventuellement remplacer R_j par $R_j - m$ et S_j par $S_j - m_1$, m et m_1 étant des estimations respectives

de $\mathbb{E}[R]$ et $\mathbb{E}[S]$ calculées a priori avec un nombre restreint de données. Ces processus à données standardisées en ligne, dont la convergence est établie, sont comparés à des processus classiques dans le cas $q = 1$ sur 11 fichiers de données dont les caractéristiques sont données dans Duarte et al. (2018b), en régression linéaire ou analyse discriminante linéaire binaire. Les meilleurs résultats sont obtenus pour un processus à données standardisées en ligne, à pas constant, utilisant toutes les observations jusqu'à l'étape courante avec introduction d'un lot de nouvelles observations à chaque étape (cas 2).

Ce travail a bénéficié d'une aide de l'Etat gérée par l'ANR au titre du programme d'investissements d'avenir portant la référence ANR-15-RHU 0004.

Références

Cardot, H., P. Cénac, et J.-M. Monnez (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics & Data Analysis 56*(6), 1434–1449.

Duarte, K., J.-M. Monnez, et E. Albuisson (2018a). Methodology for constructing a short-term event risk score in heart failure patients. Accepted in *Applied Mathematics*, hal-01813130v1.

Duarte, K., J.-M. Monnez, et E. Albuisson (2018b). Sequential linear regression with online standardized data. *PLOS ONE 13*(1), e0191186.

Monnez, J.-M. (2018). Online logistic regression process with online standardized data. *Working paper*.

Monnez, J.-M. et A. Skiredj (2018). Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream. Under review in *Journal of Multivariate Analysis*.

Oza, N. C. et S. Russell (2001). Online bagging and boosting. In *Proceedings of Eighth International Workshop on Artificial Intelligence and Statistics, Key West, Florida, USA, 4-7 January 2001*, pp. 105–112.

Pascanu, R., T. Mikolov, et Y. Bengio (2012). Understanding the exploding gradient problem. *arXiv :1211.5063v1*.

Pitt, B., W. Remme, F. Zannad, J. Neaton, F. Martinez, B. Roniker, R. Bittman, S. Hurley, J. Kleiman, et M. Gatlin (2003). Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *New England Journal of Medicine 348*(14), 1309–1321.

Summary

A methodology for constructing a short-term event risk score in heart failure patients is presented. Suppose now that patient data arrive continuously and that the score function must be updated online. The particular case of online estimation of parameters of a linear regression function is presented, using online standardized data instead of raw data.