Actualisation en ligne d'un score d'ensemble

Benoît Lalloué, Jean-Marie Monnez, Eliane Albuisson

51e Journées de Statistiques

Juin 2019, Nancy

Introduction

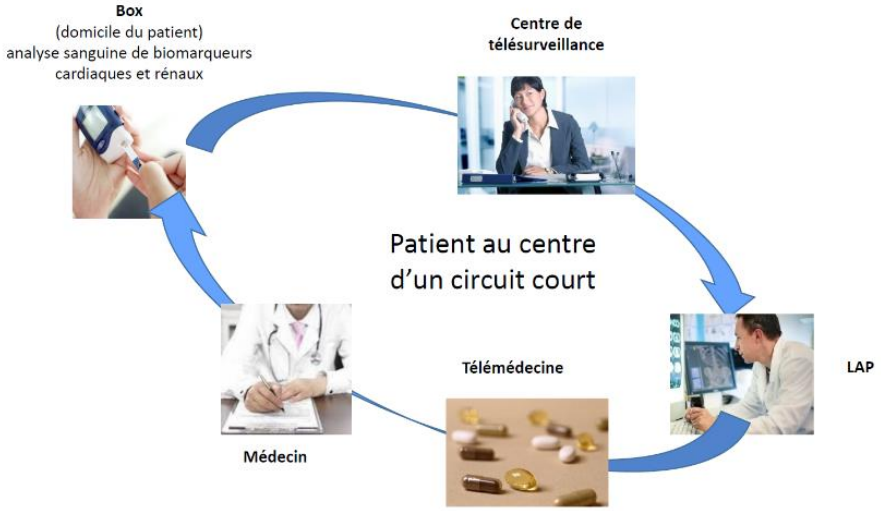
- *Problématique générale* : prédiction des valeurs d'une variable dépendante y à partir de variables observées $x^1, \ldots, x^p$

- *Illustration :* identifier les patients ayant un risque d'hospitalisation ou de décès à court terme (≤ 30 jours) pour progression de leur insuffisance cardiaque.

- Une possibilité : construire un score à l'aide d'une méthode d'ensemble

Apprentissage en ligne

- *Problématique :* comment mettre à jour les paramètres du score quand les données arrivent en flux continu ?



- Stocker et réutiliser à chaque fois toutes les données déjà obtenues jusqu'à présent : peu pratique (voire impossible)

Score d'ensemble « batch »

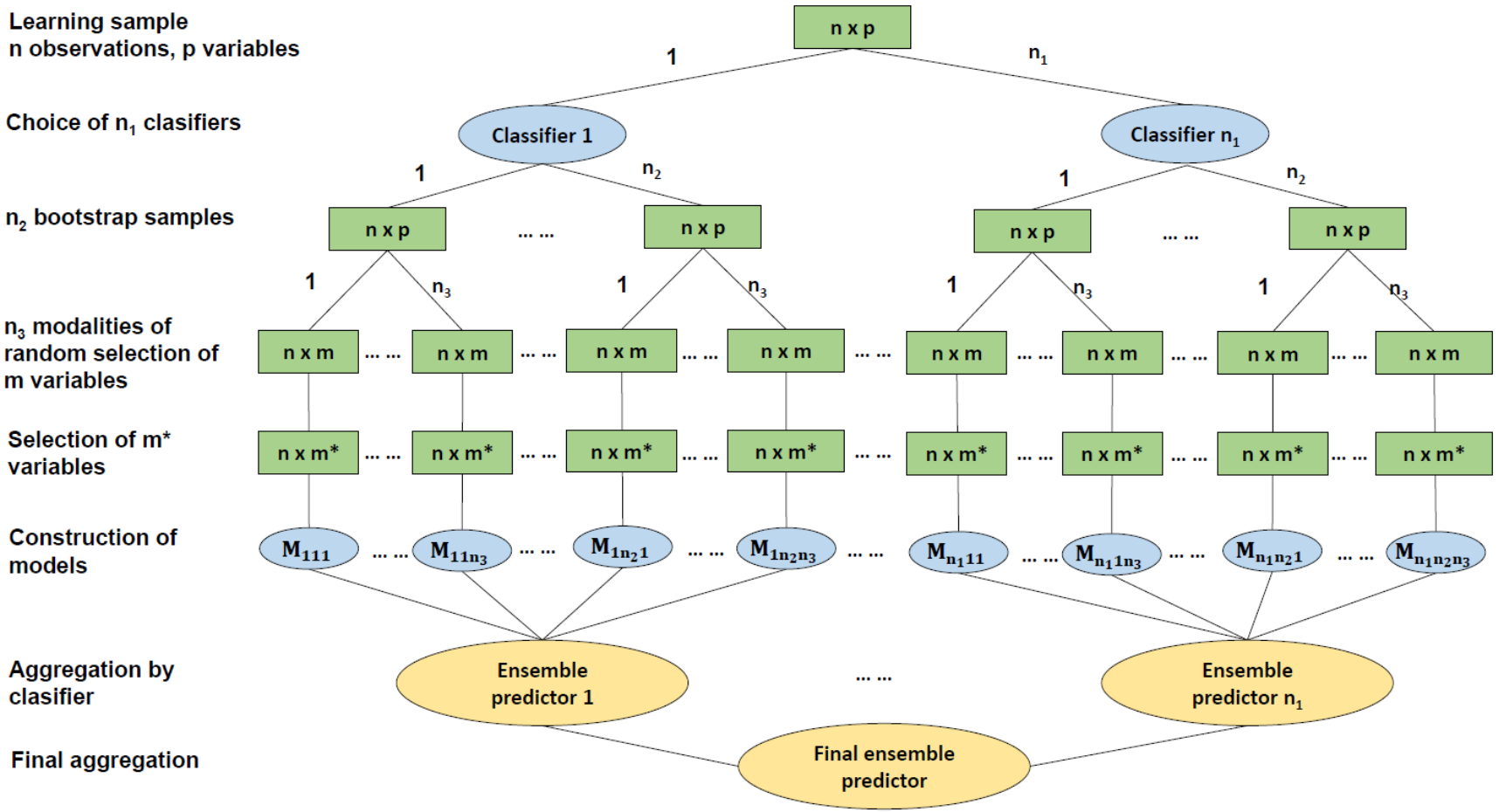
- Méthode de construction d'un score d'ensemble proposée par Duarte, Monnez et Albuisson

- Inspirations : *bagging* (Breiman 1996), forêts aléatoires de modèles logistiques (Tufféry 2015), *random generalized linear models* (Song, Langfelder & Horvath 2013)
  → mais en utilisant plusieurs règles de prédiction

- Application sur des données cliniques de patients atteints d'insuffisance cardiaque

Duarte, K., Monnez J.-M. and Albuisson E. (2018). Methodology for Constructing a Short-Term Event Risk Score in Heart Failure Patients. *Appl Math*, 09(08):954-74.

Score d'ensemble « batch » (2)



Duarte, K., Monnez J.-M. and Albuisson E. (2018a). Methodology for Constructing a Short-Term Event Risk Score in Heart Failure Patients. *Appl Math*, 09(08):954-74.

Score d'ensemble « en ligne »

- Choix des règles de classifications : définies par le score initial qu'on cherche à actualiser

- Mise à jour des échantillons par bootstrap Poisson (Oza & Russel 2001) :
  Pour chaque nouvelle observation et chaque échantillon bootstrap b_i :
  - simuler $k_i \sim \mathcal{P}(1)$
  - ajouter k_i fois l'observation à l'échantillon b_i

- Variables sélectionnées : définies par le score initial

Score d'ensemble « en ligne » (2)

- Mise à jour des prédicteurs :
  - Méthode dépendant des règles de prédiction
  - Pour la régression logistique et la LDA :
    - Processus de gradient stochastique
    - Standardisation en ligne des données pour éviter les explosions numériques

Régression linéaire en ligne

Soit :

- $R\ (p, 1)$ et $S\ (q, 1)$ deux vecteurs aléatoires
- A l'étape *n* : arrivée de m_n nouvelles données (R_i, S_i) échantillon iid de (R, S)
- $M_n = \sum_{i=1}^{n} m_i$; $I_n = \{M_{n-1} + 1, \dots, M_n\}$
- $\bar{R}_{M_n} = \frac{1}{M_n} \sum_{i=1}^{M_n} R_i$
- $\bar{S}_{M_n} = \frac{1}{M_n} \sum_{i=1}^{M_n} S_i$
- Γ_n (resp. Γ_n^1) la matrice diagonale des inverses des écarts-types des composantes de *R* (resp. *S*) calculées récursivement à partir des données $(R_i, S_i), i \leq n$

On cherche à estimer le vecteur θ des coefficients de la régression linéaire de *S* par rapport à *R*.

Duarte, K., Monnez, J.M. and Albuisson, E. (2018b). Sequential linear regression with online standardized data, *PloS One*, 13 (1) e0191186.

Régression linéaire en ligne (2)

Si :

- $a_n \rightarrow 0^+ ; \sum_{n=1}^{\infty} a_n = \infty$
- $B_n \rightarrow B = Covar(R) ; F_n \rightarrow F = Covar(R, S)$
- (et quelques autres conditions classiques)

Alors le processus stochastique suivant converge vers θ :

$$X_{n+1} = X_n - a_n(B_n X_n - F_n)$$

On utilise toutes les observations jusqu'au pas *n* :

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

Duarte, K., Monnez, J.M. and Albuisson, E. (2018b). Sequential linear regression with online standardized data, *PloS One*, 13 (1) e0191186.

Régression logistique en ligne

Soit :

- R vecteur aléatoire réel et S v.a. à valeurs dans $\{0, 1\}$
- A l'étape *n* : m_n nouvelles données (R_i, S_i) échantillon iid de (R, S)
- $\bar{R}_{M_n}$ le vecteur des moyennes des R_i jusqu'au pas *n*
- Γ_{M_n} la matrice de diagonale de l'inverse des écart-types des R_i jusqu'au pas *n* (calculés récursivement)
- $\tilde{Z}_j = \Gamma_{M_{n-1}}\left(R_j - \bar{R}_{M_{n-1}}\right)$ $(j \in I_n)$ les données standardisées en ligne
- $h(u) = \frac{e^u}{1+e^u}$ la fonction logistique

On cherche à estimer le vecteur θ des coefficients de la régression logistique de *S* par rapport à *R*.

Lalloué, B., Monnez, J.-M., Albuisson, E. (2019). Streaming constrained binary logistic regression with online standardized data. Application to scoring heart failure., *Working paper.*

Régression logistique en ligne (2)

Si :

- $a_n > 0, \sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} \frac{a_n}{\sqrt{n}} < \infty, \sum_{n=1}^{\infty} a_n^2 < \infty$
- (et quelques autres conditions classiques)

Le processus de gradient stochastique suivant converge vers θ :

$$X_{n+1} = X_n - a_n \frac{1}{m_n} \sum_{j \in I_n} \tilde{Z}_j \left(h\left(\tilde{Z}_j X_n\right) - S_j \right)$$

Lalloué, B., Monnez, J.-M., Albuisson, E. (2019). Streaming constrained binary logistic regression with online standardized data. Application to scoring heart failure., *Working paper.*

Résultats d'application

- Application sur des données issues d'un essai clinique (EPHESUS) :
  - 21382 couples patients-visites
  - 317 événements à 30 jours
  - 27 variables explicatives

- Score « batch » construit sur ces données avec :
  - $n_1 = 1$ (Régression logistique)
  - $n_2 = 100$ échantillons bootstrap
  - $n_3 = 3$ modalités de sélection de variables

  → 300 prédicteurs

Résultats d'application (2)

- Score « en ligne » construit avec :
  - Simulation d'un flux de données par tirage au sort dans le jeu de données
  - Régression logistique
  - Même nombre d'échantillons bootstrap ($n_2 = 100$)
  - Même nombre de variables sélectionnées (avec les mêmes modalités) pour chaque couple échantillon-modalité
  - Initialisation des 300 processus avec le vecteur nul

Résultats d'application (3)

Comparaison entre scores « batch » et « en ligne » :

- Cosinus entre les vecteurs des coefficients des scores d'ensemble
- Corrélations entre les scores obtenus

Processus simple		N	2N	3N	4N	5N
10 nouvelles obs par étape	*Cos*	0.9995	0.9998	0.9998	0.9998	0.9998
	Cor	0.9443	0.9642	0.9683	0.9721	0.9728
100 nouvelles obs par étape	*Cos*	0.9998	0.9998	0.9998	0.9998	0.9998
	Cor	0.9507	0.9599	0.9638	0.9646	0.9684

Processus moyennisé à pas constant par paliers (50)		N	2N	3N	4N	5N
10 nouvelles obs par étape	*Cos*	0.9998	0.9999	0.9999	0.9999	0.9999
	Cor	0.9715	0.9923	0.9930	0.9948	0.9950
100 nouvelles obs par étape	*Cos*	0.9999	0.9999	0.9999	0.9999	0.9999
	Cor	0.9865	0.9861	0.9900	0.9926	0.9933

Conclusion

- Méthode de mise à jour d'un score d'ensemble en ligne grâce au bootstrap Poisson et à des processus de gradient stochastique

- Résultats satisfaisants de convergence de la méthode sur un jeu de données réelles

- Perspectives :
  - Ajouter la régression linéaire/LDA pour reproduire le score de Duarte et al.
  - Optimiser le temps de calcul

Merci !

Méthodes d'ensemble

- Méthode d'ensemble ?
  → Construire de nombreux prédicteurs à partir de variantes (échantillons bootstrap, variables sélectionnées, …) autour de méthodes simples (régression linéaire ou logistique, k-nn, arbres de décision, …)

  → Les agréger (par moyenne, par vote, …)

  → On s'attend à ce que le prédicteur d'ensemble donne de meilleurs résultats que chaque prédicteur individuel

- Rarement utilisées en médecine : souvent jugées trop « boite noire »