Bagging et Forêts Aléatoires

Nicolas Verzelen, Joseph Salmon (et Pierre Pudlo)

INRAE / Université de Montpellier

Plan

Bagging

Forêts aléatoires

Importance des variables

Agrégation d'algorithmes de prédiction

Méthodes d'agrégation

- Construction d'un grand nombre de prédicteurs $\widehat{f}_b$, $b = 1 \dots B$
- Agrégation ou combinaison de ces algorithmes :
  $\widehat{f} = \sum_{b=1}^{B} w_b \widehat{f}_b$ ou $\mathrm{sign}\left(\sum_{b=1}^{B} w_b \widehat{f}_b\right)$.

$\hookrightarrow$ En particulier : agrégation par *bagging/boosting*

Exemples :

Bagging Breiman (1996) : pour des algorithmes instables, de variance forte (provient de **B**ootstrap **agg**regat**ing**)

Boosting Freund et Schapire (1997) : pour des algorithmes fortement biaisés, mais de faible variance

Bagging

Rappel : $\eta^*(x) = \mathbb{E}[Y|X = x]$ (fonction de régression)

Principe du bagging : agréger un ensemble d'algorithmes $\widehat{\eta}_1, \ldots, \widehat{\eta}_B$ sous la forme $\widehat{\eta}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\eta}_b$

Décomposition biais/variance :

$$\mathbb{E}\left[(\widehat{\eta}(x) - \eta^*(x))^2\right] \leq \left(\mathbb{E}\left[\widehat{\eta}(x)\right] - \eta^*(x)\right)^2 + \mathrm{Var}\left(\widehat{\eta}(x)\right)$$

Cas $\widehat{\eta}_1, \ldots, \widehat{\eta}_B$ i.i.d. : $\begin{cases} \mathbb{E}\left[\widehat{\eta}(x)\right] = \mathbb{E}\left[\widehat{\eta}_b(x)\right] \\ \mathrm{Var}\left(\widehat{\eta}(x)\right) = \mathrm{Var}\left(\widehat{\eta}_b(x)\right)/B \end{cases}$

⚠ les $\widehat{\eta}_B$ sont non i.i.d. car construits sur le même échantillon !

Sans indépendance : avec $\rho_x = \mathrm{corr}(\widehat{\eta}_b(x), \widehat{\eta}_{b'}(x)) > -1/(B-1)$
http://pages.stern.nyu.edu/~rengle/Dynamic_Equicorrelation.pdf

$$\mathrm{Var}\left(\widehat{\eta}(x)\right) = \rho_x \mathrm{Var}\left(\widehat{\eta}_b(x)\right) + \frac{1-\rho_x}{B}\mathrm{Var}\left(\widehat{\eta}_b(x)\right) \overset{B\to+\infty}{\to} \rho_x \mathrm{Var}\left(\widehat{\eta}_b(x)\right)$$

Idée : construire des prédicteurs avec échantillons *bootstrap*

Algorithme du bagging

Considérons :

- un type de prédicteur η_{D^n} associé à un échantillon D^n
- un nombre B (grand) d'échantillons bootstrap de D^n : $D^{*1}_{m_n} \dots D^{*B}_{m_n}$, de taille $m_n \leq n$, indépendants les uns des autres conditionnellement à D^n.

$$\text{Pour } b = 1 \dots B, \quad \widehat{\eta}_b = \eta_{D^{*b}_{m_n}}$$

<u>Principe du bagging</u> : agréger les algorithmes $\widehat{\eta}_1, \dots, \widehat{\eta}_B$:

- $\widehat{\eta} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\eta}_b$; moyenne/ régression
- $\widehat{\eta} = \mathrm{sign}(\sum_{b=1}^{B} \widehat{\eta}_b)$; vote à la majorité / classification binaire

Plan

Bagging

Forêts aléatoires

Importance des variables

Forêts aléatoires Breiman (2001)

Forêts aléatoires : Bagging d'arbres maximaux construits sur des échantillons bootstrap de taille $m_n = n$, par une variante de la méthode CART consistant, pour chaque nœud, à

- tirer au hasard un sous-échantillon de taille $p' < p$ de variables explicatives,
- partition fils à gauche / fils à droite : sur la base de la "meilleure" de ces p' variables explicatives

Algorithme : Forêts aléatoires

input :

x : l'entrée dont on veut prédire la sortie

D^n : l'échantillon observé

p' : le nombre de variables explicatives sélectionnées à chaque nœud

B : le nombre d'itérations

pour $b = 1, \ldots, B$ **faire**

- Tirer un échantillon bootstrap D_n^{*b} de D^n
- Construire un arbre maximal $\widehat{\eta}_b$ sur l'échantillon bootstrap D_n^{*b} par la variante de CART suivante :
- **pour chaque nœud** *de* 1 *à* N_b **faire**
  - Tirer un sous-échantillon de p' variables explicatives
  - Partitionner le nœud à partir de ces p' variables

output : $\frac{1}{B}\sum_{b=1}^{B}\widehat{\eta}_b(x)$ ou $\mathrm{sign}\left(\sum_{b=1}^{B}\widehat{\eta}_b(x)\right)$

Ajustement des paramètres : erreur Out Of Bag

Si p' diminue, la variance diminue (la corrélation diminue) et le biais augmente (moins bonne qualité d'ajustement)

Compromis biais/variance $\Rightarrow$ choix optimal de p' lié aussi au nombre d'observations dans les nœuds terminaux

$\hookrightarrow$ Ajustement par validation croisée hold-out ou K fold ou par l'estimation Out Of Bag du risque

Erreur Out Of Bag :

- $\frac{\sum_{i=1}^{n} I_i^b (\widehat{\eta}_b(x_i) - y_i)^2}{\sum_{i=1}^{n} I_i^b}$ (régression)
- $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\text{sign}\left(\sum_{b=1}^{B} I_i^b \widehat{\eta}_b(x_i)\right) \neq y_i}$ (classification binaire)

où $I_i^b = \begin{cases} 1, & \text{si l'observation } i \notin D_n^{*b} \\ 0, & \text{sinon} \end{cases}$

Avantages / inconvénients

Avantages

- meilleure prédiction
- Implémentation facile
- Adaptée à la parallélisation

Inconvénients

- perte de l'interprétation (effet "boîte noire")

$\hookrightarrow$ mesures d'importance des variables, même si on perd l'interprétation avec des seuils sur ces variables.

Plan

Importance des variables

Méthode rudimentaire : regarder la fréquence des variables explicatives sélectionnées pour découper les arbres de la forêt

Méthode recommandée par Breiman (2001) : pour chaque variable explicative $X^{(j)}$ et pour tout b :

- Calculer l'erreur Out Of Bag de l'arbre $\widehat{\eta}_b$ sur l'échantillon Out Of Bag correspondant :

$$OOB_b = \frac{\sum_{i=1}^n I_i^b(\widehat{\eta}_b(x_i)-y_i)^2}{\sum_{i=1}^n I_i^b} \text{ ou } \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\text{sign}\left(\sum_{b=1}^B I_i^b \widehat{\eta}_b(x_i)\right) \neq y_i}$$

- Créer un échantillon Out Of Bag permuté (en permutant aléatoirement les valeurs de la variable explicative $X^{(j)}$ dans l'échantillon Out Of Bag) et calculer l'erreur Out Of Bag OOB_b^j de l'arbre $\widehat{\eta}_b$ sur cet échantillon Out Of Bag permuté

L'importance de la variable $X^{(j)}$ est finalement mesurée par

$$\frac{1}{B}\sum_{b=1}^B (OOB_b^j - OOB_b).$$

Résumé

- Les arbres de décision sont des modèles simples et interprétables.
- Cependant, ils fournissent souvent de mauvais résultats comparés à d'autres méthodes.
- Le bagging est une bonne méthode pour améliorer la prédiction des arbres de décision, et agrège de nombreux arbres pour les combiner pour obtenir une décision finale
- Les forêts aléatoires (et le boosting) font partie de l'état de l'art actuel des méthodes d'apprentissage supervisé. <u>Limite</u> : difficile à interpréter

Bibliographie

- Breiman, L. “Bagging Predictors”. In : *Mach. Learn.* 24.2 (1996), p. 123-140.
- – .“Random Forests”. In : *Mach. Learn.* 45.1 (2001), p. 5-32.
- Freund, Y. et R. E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In : *Journal of computer and system sciences* 55.1 (1997), p. 119-139.