Introduction aux forêts aléatoires

Robin Genuer

5 février 2019
IME204 Aide à la décision
Master 2 SITIS, ISPED, Université de Bordeaux

Plan

Forêts aléatoires

- introduites par Breiman (2001)
- famille des méthodes d'ensemble, Dietterich (1999,2000)
- algorithme d'apprentissage statistique très performant, à la fois pour des problèmes de classification et de régression.

Notations :

$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ v.a. i.i.d. de même loi que (X, Y).

$X \in \mathbb{R}^p$ (variables)
$Y \in \mathcal{Y}$ (réponse)

- $\mathcal{Y} = \mathbb{R}$: régression
- $\mathcal{Y} = \{1, \ldots, L\}$: classification

But : construire un prédicteur $\widehat{h} : \mathbb{R}^p \to \mathcal{Y}$

Données spam

- 4601 emails : 2788 emails souhaitables, 1813 spams.
- $p = 57$ variables : proportions d'occurences de mots ou de caractères, comme par exemple `$`, `!`, `free`, `money`...
- La réponse est binaire : spam ou non-spam.
- But : construire un "bon" filtre anti-spam

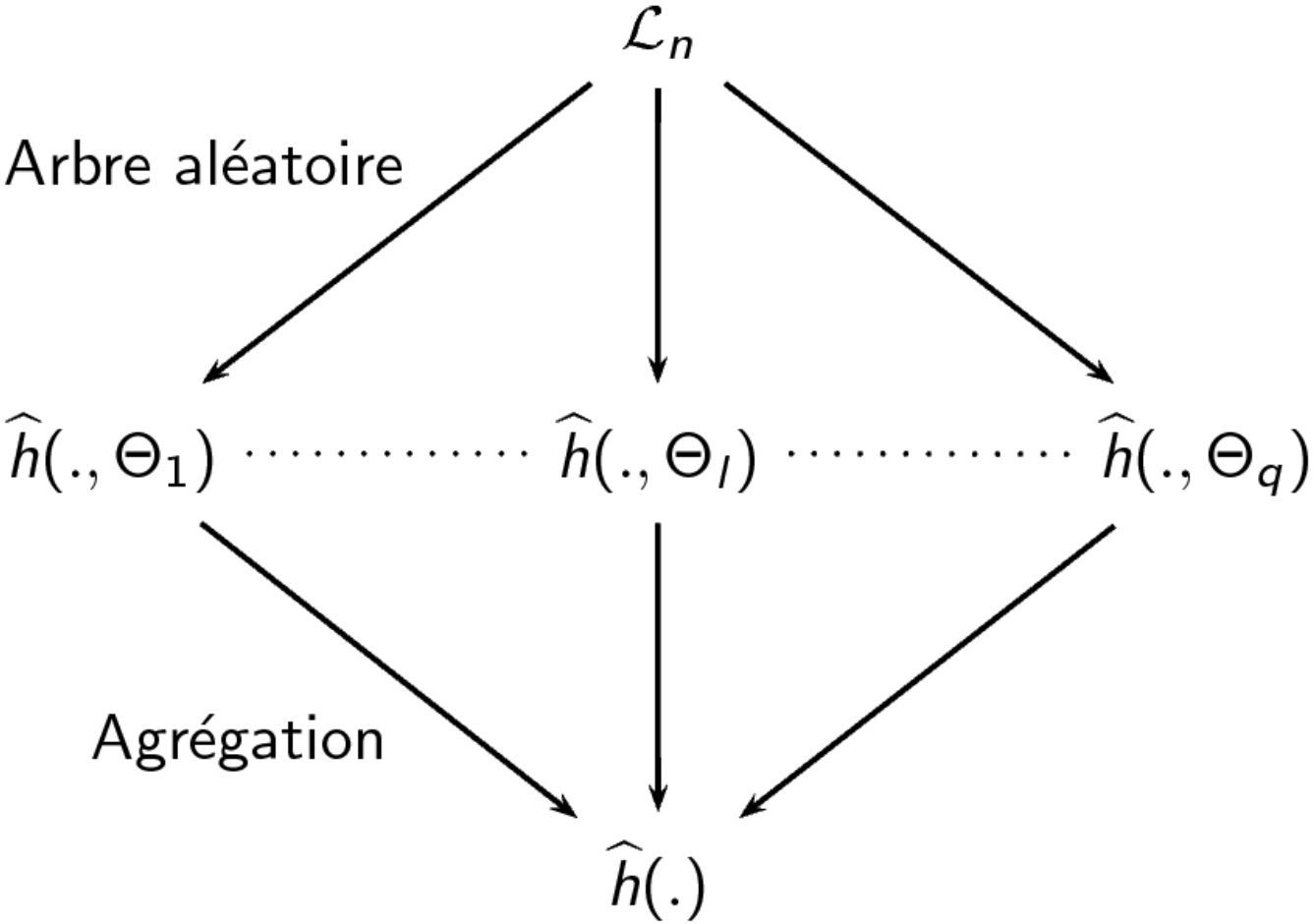
  (Un nouveau mail arrive, il faut réussir à prédire si c'est un spam ou non.)
- Performance du filtre anti-spam calculée par erreur test :
  - $n = 2300$ mails pour l'apprentissage
  - 2301 mails pour tester les prédictions

Définition : Forêts aléatoires (Breiman 2001)

$\left\{\widehat{h}(.,\Theta_\ell), 1 \leq \ell \leq q\right\}$ collection de prédicteurs par arbre,
$(\Theta_\ell)_{1\leq\ell\leq q}$ v.a. i.i.d. indépendantes de $\mathcal{L}_n$.
Prédicteur des forêts aléatoires $\widehat{h}$ obtenu en agrégeant la collection d'arbres.

Agrégation :

- $\widehat{h}(x) = \frac{1}{q}\sum_{\ell=1}^{q}\widehat{h}(x,\Theta_\ell)$ en régression
- $\widehat{h}(x) = \underset{1\leq c\leq L}{\operatorname{argmax}} \sum_{\ell=1}^{q}\mathbb{1}_{\widehat{h}(x,\Theta_\ell)=c}$ en classification


$\mathcal{L}_n$
Arbre aléatoire
$\widehat{h}(.,\Theta_1)$
$\widehat{h}(.,\Theta_l)$
$\widehat{h}(.,\Theta_q)$
Agrégation
$\widehat{h}(.)$

Bilan sur les arbres

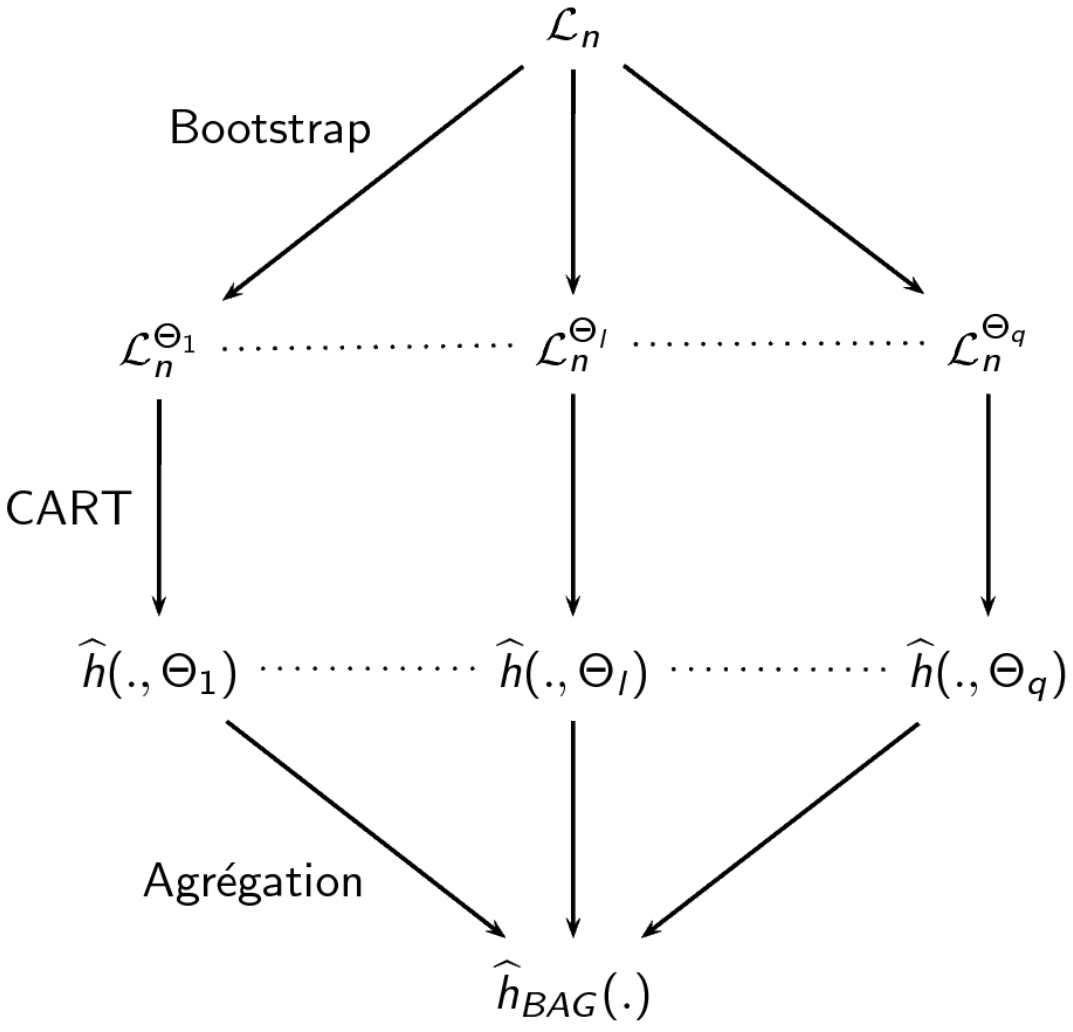
Avantages :

- Méthode très générale, facile à comprendre et à mettre en oeuvre, s'applique à beaucoup de type de données (variables quali et/ou quanti) et de problèmes (régression, classification)
- Représentation graphique sous forme d'arbre très visuelle, très utile pour interpréter les résultats
- Méthode non-linéaire et non-paramétrique, donc méthode très souple.

Inconvénients :

- Méthode instable : si on change un peu les données d'apprentissage, l'arbre peut changer complètement
- Il existe des méthodes avec des performances bien meilleures

Bagging (Breiman 1996)



Instabilité de CART $\Rightarrow$ amélioration des performances

Random Forests-Random Inputs (Breiman 2001)

Définition : Arbre RI

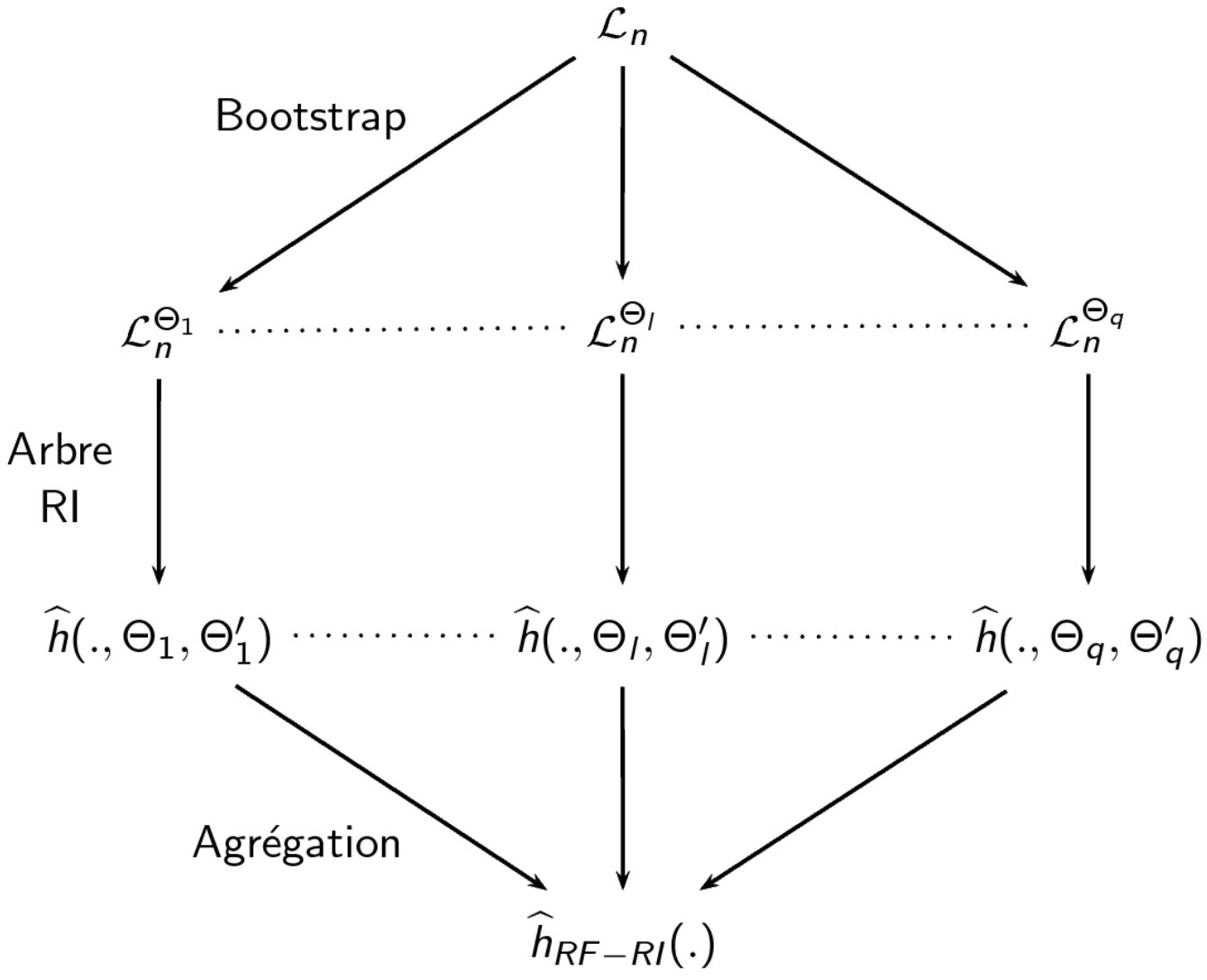
Un arbre RI consiste à tirer aléatoirement, à chaque noeud `mtry` variables, puis à chercher la meilleure coupure uniquement parmi les variables sélectionnées. De plus, un arbre RI n'est pas élagué.

`mtry` est le même pour tous les noeuds de tous les arbres de la forêt.

Définition : Random Forests-RI

Une forêt Random Forests-RI est obtenue en effectuant du Bagging avec des arbres RI.

Random Forests-RI



Aléa supplémentaire $\Rightarrow$ amélioration des performances

Random Forests-RI

Paquet R `randomForest`:

- basé sur le code de Breiman, Cutler (2000)
- décrit dans Liaw, Wiener (2002)

Principaux paramètres de l'algorithme `randomForest` :

- `ntree` : nombre d'arbres dans la forêt
- `mtry` : le nombre de variables tirées aléatoirement à chaque noeud

Données spam

Prédicteur	arbre optimal	bagging	rf-ri
Erreur test	0.08	0.061	0.053

Table: Taux d'erreurs test de l'arbre optimal, du bagging et des rf-ri pour les données spam

Estimation de l'erreur de prédiction

OOB = Out Of Bag ($\approx$ "En dehors du Bootstrap")

Erreur OOB

Pour prédire X_i, on agrège uniquement les prédicteurs $\widehat{h}(.,\Theta_\ell)$ construits sur des échantillons bootstrap ne contenant pas (X_i, Y_i).

$\Rightarrow$ Erreur OOB :

- $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$ en régression
- $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i \neq \widehat{Y}_i}$ en classification

Importance des variables

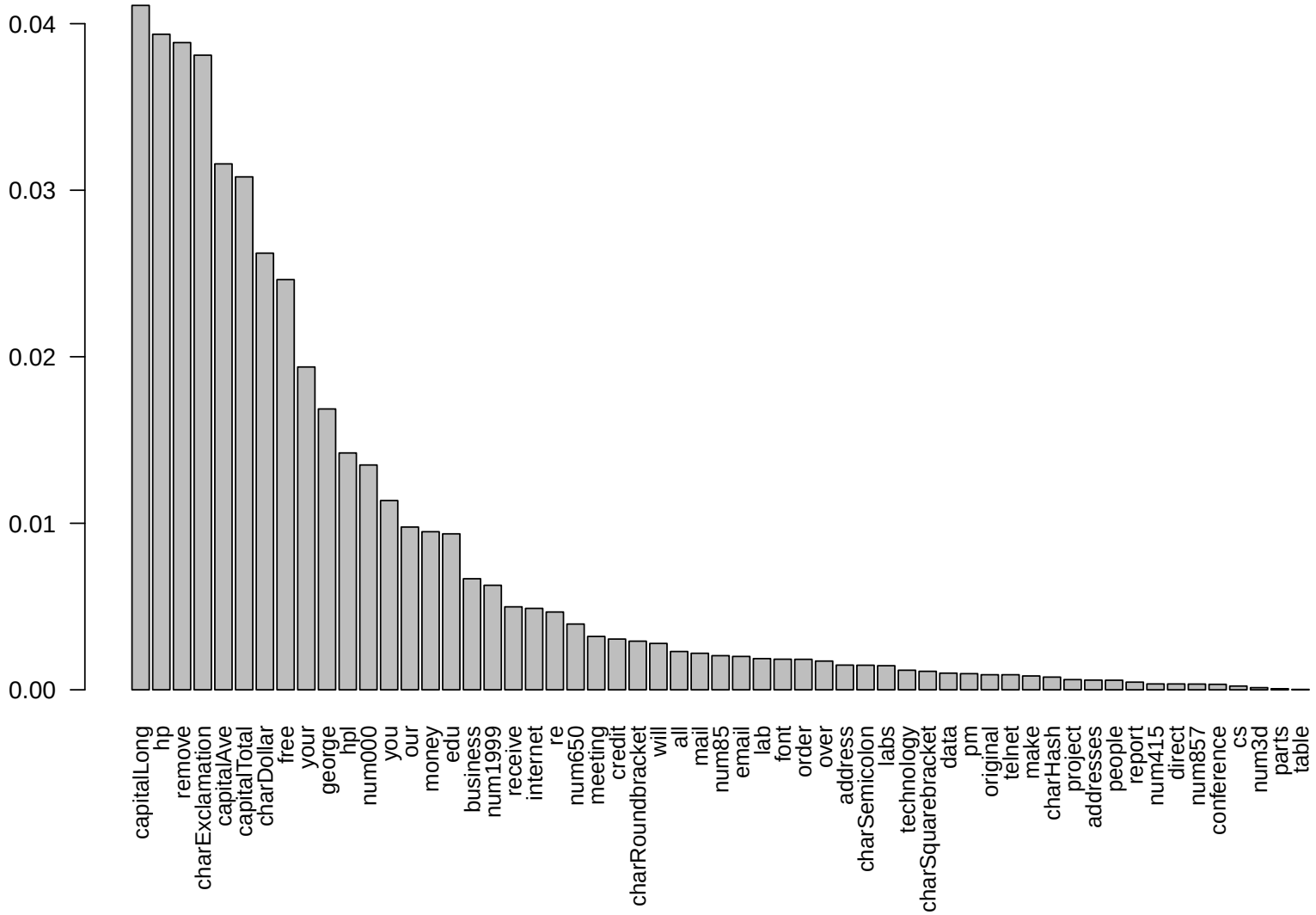
Importance des variables

Soit $j \in \{1, \ldots, p\}$. Pour chaque échantillon OOB, on permute aléatoirement les valeurs de la j-ième variable des données.

Importance de la j-ième variable = augmentation moyenne de l'erreur d'un arbre après permutation.

Plus l'augmentation d'erreur est forte,
plus la variable est importante.

Variables importantes dans spam

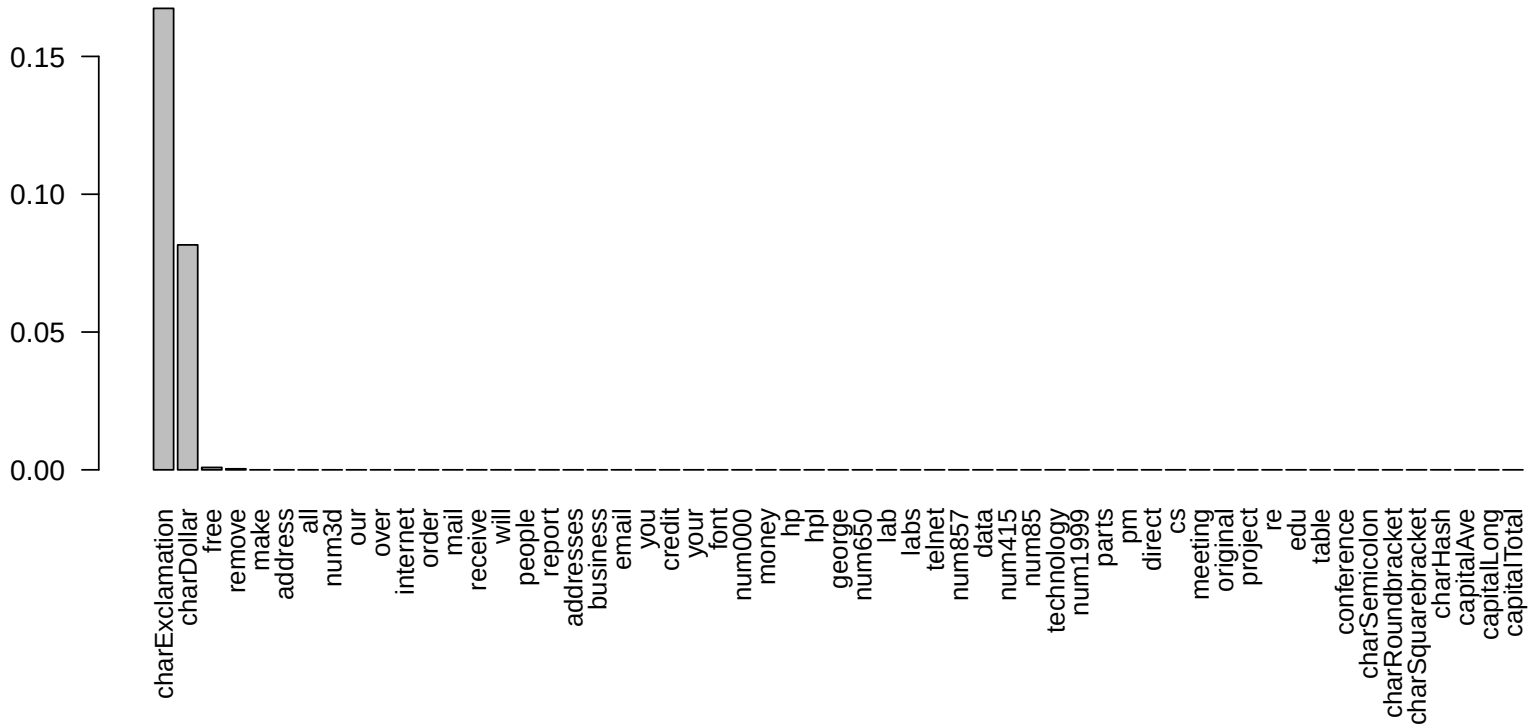


Variables importantes dans spam

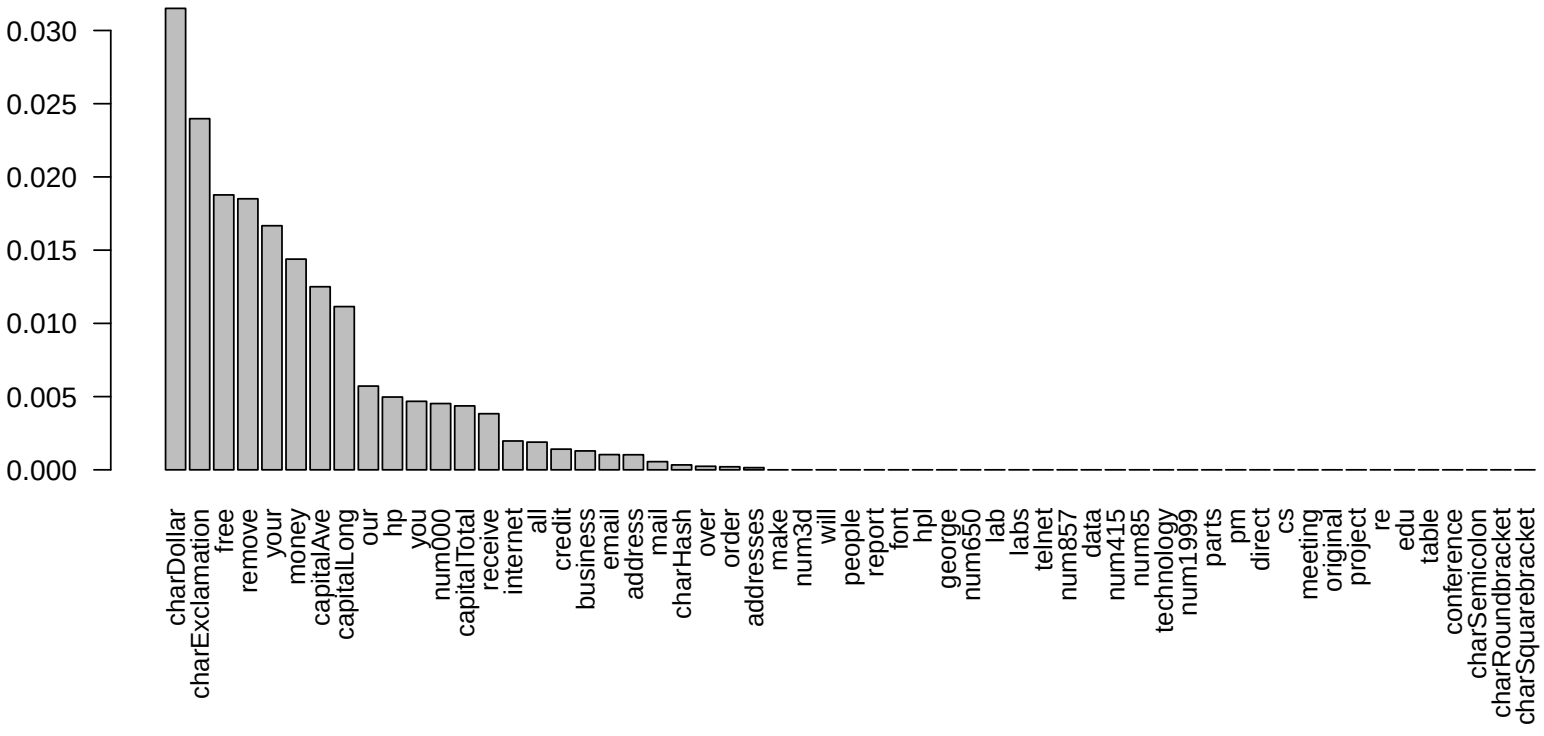
Forêts	sans select	select
Erreur test	0.053	0.06

Table: Taux d'erreurs des forêts sans sélection et avec sélection de variables pour les données spam

Bagging d'arbres à 2 feuilles



Forêts d'arbres à 2 feuilles



Performances avec des arbres 2 feuilles

Prédicteur	arbre	bagging	forêts
Erreur test	0.209	0.207	0.173

Table: Taux d'erreurs du bagging et des forêts avec des arbres à 2 feuilles pour les données `spam`

Short bibliography


Breiman, L., Friedman J., Olshen R., Stone C. *Classification And Regression Trees*. Chapman & Hall (1984)


Breiman, L. *Bagging*. Machine Learning (1996)


Breiman, L. *Random Forests*. Machine Learning (2001)


Genuer R., Poggi J.-M. *Arbres CART et Forêts aléatoires, Importance et sélection de variables*. Preprint (2016)


Genuer R., Poggi J.-M. and Tuleau-Malot C. *VSURF: An R Package for Variable Selection Using Random Forests*. The R Journal (2015)