SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

SORBONNE UNIVERSITES

Cours M2

Bagging

Élèves :
Salim AMOUKOU
Mohamed-Jad Bouchra

Enseignant :
Gérard Biau

13 janvier 2019

Table des matières

1 Introduction

Le Bagging, introduit par Breiman en 1994, est une puissante technique qui permet d'améliorer la performance des modèles simples et de réduire l'overfitting des modèles plus complexes. Le principe est très simple : au lieu d'entraîner notre modèle sur tout l'échantillon, on entraîne plusieurs copies de notre modèle sur différents sous-échantillons (tirés avec/sans remise). Ces modèles sont alors agrégés en prenant la moyenne ou par un système de vote.

Bien que le Bagging est très effectif dans la pratique, il n'est pas encore bien compris dans la théorie. Le principal but de ce cours est d'essayer de comprendre pourquoi le Bagging fonctionne quitte à se restreindre à des cas particuliers pour simplifier le problème. Afin d'atteindre cet objectif, il nous faut poser les bases de l'apprentissage statistique.

1.1 Qu'est ce que l'apprentissage statistique ?

Objectifs : Prédire une donnée de sortie y à partir d'une donnée d'entrée x, en construisant la meilleure règle de décision g^*

Exemples :

- Classifier des images à partir des pixels.
- Prédire la présence d'une maladie à partir d'un génome.
- Traduire un texte.

Difficultés :

- Y n'est pas une fonction déterministe de X.
- Il peut y avoir un bruit, par exemple $Y = f(X) + \epsilon$.
- Plus généralement, $Y = f(X, Z)$ où Z n'est pas observé.

Approches possibles :

- Essayer de faire bien dans le pire cas $\rightarrow$ théorie des jeux, stratégie minmax...
- Essayer de faire bien en moyenne $\rightarrow$ apprentissage statistique.

Idée : Modéliser X et Y comme des variables aléatoires ainsi la meilleure décision en moyenne pourrait être prise à partir de $\mathbb{P}(Y = .|X = x)$.

- On ne connaît pas la loi de $Y|X$.
- X et Y peuvent être de très grande dimension $\rightarrow$ fléau de la dimension.

Information disponible : Des réalisations (X, Y) : $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$.

Solution : Apprendre ! c'est à dire utiliser une stratégie qui marche pour un ensemble d'observations existantes et qui puisse se généraliser aux autres observations.

Formalisation : Soit $(X, Y) \in \mathbb{R}^d \mathcal{X}\mathbb{N}$, un échantillon de variables aléatoires indépendantes identiquement distribuées $\mathcal{D}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ qui suivent la même loi que (X, Y) tel que $\mathbb{E}[Y^2] < \infty$. Notre but sera de trouver un bon estimateur de $r(x) = \mathbb{E}[Y|X = x]$ et de construire une règle de décision g^*.

Question : Pourquoi modéliser par des variables aléatoires ?

Formalisme de la solution : Comme nous l'avons dit précédemment, notre but est de trouver une règle de décision, bien entendu toute fonction borèlienne $f : \mathbb{R}^d \to \mathbb{R}$ est une règle. Il nous faut alors un critère de qualité. Dans un modèle de classification, on considère qu'une erreur se produise si $g(x) \neq y$. C'est donc tout maternellement qu'on pose $L(g) = \mathbb{P}(g(x) \neq y)$, la probabilité qu'une erreur se produise, qui sera notre mesure de qualité. On a ainsi transformer notre problème en un problème d'optimisation, c'est à dire trouver g^* tels que $g^* = argmin_{g:\mathbb{R}^d \to \mathbb{R}} L(g)$.

Remarque. *Un classifieur de bayes n'existe pas toujours et si il existe, il n'est pas forcément unique. De plus, dans la pratique il est inatteignable car on ne connait pas $r(x)$. Ainsi notre but de construire un classifieur qui se rapproche le plus de la règle de bayes g^* ou qui a un petit risque dans le pire des cas.*

Maintenant qu'on a bien définit notre problème, on va introduire un critère de qualité. On considère que notre estimateur g^* est performant s'il converge vers l'estimateur de bayes aux sens suivants.

Consistance :

- La règle de décision g_n est dite **consistante** pour une distribution de (X, Y) si : $\mathbb{E}[L(g_n)] \to L(g^*)$
- La règle de décision g_n est dite **fortement consistante** pour une distribution de (X, Y) si : $L(g_n) \to L(g^*)$

Remarque. *Il est légitime de se poser la question de l'existence de règle consistante quelle que soit la distribution de (X, Y). On dit alors que la règle est* ***universellement consistante****. Dans ce qui suit, on présente un théorème qui assure ce type de consistance sous certaines conditions.*

1.2 Théorème de Stone

On rappelle que notre but est d'estimer la fonction de régression $r(x)$ à partir de $D_n = (X_1, Y_1), ..., (X_n, Y_n)$. Une façon naturelle de procéder est de poser :

$$\hat{r_n}(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i \quad \forall x \in \mathbb{R}^d$$

Où $W_{ni}(x)$ est une fonction borélienne de x, et $X_1, ..., X_n$(pas de $Y_1, ..., Y_n$). Il est intuitivement clair que les couples (X_i, Y_i) où X_i est proche de x devront apporter plus

d'information que leurs voisins plus éloignés. Ainsi les poids devront être plus grand autour de x de tel sorte que $\hat{r}_n(x)$ soit une moyenne des Y_i correspondants aux X_i qui se situent dans un certain voisinage de x. Un estimateur de cette forme est appelé estimateur de type moyenne locale.

Exemple :

L'estimateur du k-PPV est un estimateur à moyenne locale où pour un x fixé.

$$W_{n,i}(x) = \begin{cases} \frac{1}{k} & \text{Pour les k plus proches voisins de } x \\ 0 & \text{Pour les autres.} \end{cases}$$

Théorème 1.1. *Supposons que, pour toute loi de X, les poids satisfassent les conditions suivantes :*
1 : On a,

$$\mathbb{E}(\max_{1\leq i\leq n} W_{n,i}(X)) \to 0.$$

2 : Pour tout a>0,

$$\mathbb{E}(\sum_{i=1}^{n} W_{n,i}(X)\mathbb{1}_{[||X_i - X||>a]}) \to 0.$$

3 : Il existe une constante C, telle que, pour toute fonction borélienne, $f : \mathbb{R}^d \to \mathbb{R}$ avec $\mathbb{E}|f(x)| < \infty$,

$$\mathbb{E}[\sum_{i=1}^{n} W_{n,i}(X)|f(X_i)|] \leq C\mathbb{E}|f(X)|$$

Alors la règle g_n associée à la fonction de régression $\hat{r_n}$ est ***universellement consistante****.*

La condition 1 permet de garantir qu'il n'y a pas que le poids d'un seul point qui influe l'estimateur. La condition 2 exprime le fait que seuls les poids des points situés dans un voisinage de X sont importants pour l'estimation. Enfin, la dernière condition, appelée hypothèse de Stone, est principalement de nature technique (pour la démonstration de ce théorème voir [cours apprentissage statistique G.Biau]).
Maintenant que nous avons tous les outils en main, nous pouvons rentrer dans le vif du sujet : l'étude du Bagging.

2 Généralités : Bagging

Le terme Bagging vient de la contraction de **B**oostrap **Agg**regating. Rappelons que nous disposons d'un n-échantillon iid $\mathcal{D}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ de même loi que (X, Y) et notre but est d'estimer la fonction de régression $r(x) = \mathbb{E}(Y|X = x)$.

2.1 Réduction de la variance

On considère B échantillons indépendants $\{Z_b\}_{b=1...n}$:$Z_b \subset \mathcal{D}_n$ tirés avec ou sans remise.

L'idée est de construire B estimateurs $\hat{r}_1(x), ...\hat{r}_B(x)$ sur chacun des échantillons puis de prendre leur moyenne comme estimateur final : $\hat{r}_B(x) = \frac{\sum_{k=0}^{B} \hat{r}_k(x)}{B}$. On a alors :

$$\mathbb{E}[\hat{r}(x)] = \mathbb{E}[\hat{r}_1(x)] \qquad \mathbb{V}[\hat{r}(x)] = \frac{\mathbb{E}[\hat{r}_1(x)]}{B}$$

Le biais de l'estimateur agrégé est le même que ceux qu'on agrège mais la variance est plus petite. Cependant, on n'a jamais B échantillons indépendants dans la pratique. L'idée derrière le Bagging consiste à diminuer la dépendance entre les estimateurs qu'on agrège en construisant des échantillons boostrap.

2.2 Algorithme

On construit B estimateurs sur des échantillons boostrap. Deux types de tirages sont généralement utilisés pour le boostrap.

1. On tire n observations avec remise dans D_n.
2. On tire k observations (avec ou sans remise) tel que $k < n$ dans D_n

Les échantillons boostrap introduisent de l'aléa dans nos estimateurs, on considère θ_k l'échantillon boostrap de l'étape k et $\hat{r}(., \theta_k)$ l'estimateur construit à cette étape. Notons que les $\theta_1, ..., \theta_k$ sont iid de meme loi que θ (loi du tirage boostrap). On peut alors estimer $\hat{r}(x) = \mathbb{E}_\theta[\hat{r}(x, \theta]$ par monte-carlo, car grâce à la loi forte des grands nombres on a :

$$\hat{r}(x) = \lim_{B \to +\infty} \frac{\sum_{k=0}^{B} \hat{r}(x, \theta_k)}{B} = \mathbb{E}_\theta[\hat{r}(x, \theta)]$$

L'espérance est calculée selon la loi de θ, on en déduit de cette équation que le paramètre B n'est pas crucial pour la performance. Il est recommandé de le prendre aussi grand que possible. On verra dans la suite que cet estimateur est universellement convergent dans le cas des 1-plus proche voisin sous certaines conditions.

Algorithm 1: Bagging

input :
- x l'observation à prédire
- un régrésseur (arbre de décision, k-plus proche voisins...)
- d_n l'échantillon
- B le nombre d'estimateurs que l'on agrège

for $k \leftarrow 1$ **to** B **do**

1. tirer un échantillon boostrap d_n^k de d_n
2. Ajuster le régresseur sur cet échantillon boostrap : $\hat{r}_k$

end

output: $\hat{r}(x) = \frac{\sum_{k=0}^{B} \hat{r}_k(x)}{B}$

2.3 Biais et variance

Dans cette partie, nous allons comparer l'estimateur agrégé et ceux qu'on agrège. On considère la notation suivante :

- $\hat{r}_B(x) = \frac{\sum_{k=0}^{B} \hat{r}(x,\theta_k)}{B}$ et $\hat{r}(x) = \lim_{B\to+\infty} \hat{r}_B(x)$
- $\hat{\sigma^2}(x) = V(\hat{r}(x,\theta_k)$ la variance des estimateurs
- $\rho(x) = corr(\hat{r}(x,\theta_1), \hat{r}(x,\theta_2))$ le coefficient de corrélation entre deux estimateurs.

Les estimateurs $\hat{r}(x,\theta_1), ..., \hat{r}(x,\theta_k)$ sont iid, la variance et le coefficient de corrélation sont calculés par rapport à θ et D_n.

Proposition 1. *On a :*

$$V(\hat{r}_B(x)) = \rho(x)\sigma^2(x) + \frac{1-\rho(x)}{B}\sigma^2(x)$$

En passant à la limite, on a alors :

$$V(\hat{r}(x)) = \rho(x)\sigma^2(x)$$

Démonstration. On note $X_k = \hat{r}(x,\theta_k)$, les X_k étant iid on a :

$$\begin{aligned} V(\hat{r}(x)) &= \frac{\sum_{k=1}^{B} V(X_k)}{B^2} + \frac{\sum_{1\le k\neq l\le B} \mathrm{cov}(X_k, X_l)}{B^2} \\ &= \frac{1}{B}\sigma^2(x) + \frac{B(B-1)}{B^2}\rho(x)\sigma^2(x) \\ &= \rho(x)\sigma^2(x) + \frac{1-\rho(x)}{B}\sigma^2(x) \end{aligned}$$

□

Interprétation :
On en déduit de la proposition que si $\rho(x) < 1$, alors la variance est plus petite, c'est alors la dé-corrélation entre les estimateurs qui entraîne l'efficacité du Bagging. Cependant, on peut naïvement penser qu'une bonne stratégie serait d'agréger des estimateurs avec un petit biais mais cette stratégie entraîne une variance plus grande et on a aucune garantie que $\rho(x)\sigma^2(x)$ soit petit. Ainsi il est nécessaire que les estimateurs soient sensibles au boostrap, comme les arbres de décisions ou les k-plus proche voisins. Le Bagging n'améliore pas les performances des estimateurs trop stables comme les régressions linéaires ou l'analyse discriminante.

3 Bagging et estimateur linéaire

Dans cette partie, on considère B échantillons boostrap $\{Z_b\}_{b=0...B}$, $L(y,\theta)$ une fonction d'erreur négative ou la fonction log jointe de vraisemblance, $\mathbf{x} = (x_1, ..., x_n)$, $\mathbf{z} = \{(x_1, y_1), ...(x_n, y_n)\}$ notre échantillon. On note :

— $\hat{\theta}_b(\mathbf{x}) = \text{argmax}_{\theta\in\theta} \sum_{i\in z_b} L(y_i, \theta(x_i))$ l'estimateur sur l'échantillon boostrap z_b.
— $\hat{\theta}_B(x) = \frac{\sum_{k=1}^{B} \hat{\theta}_b(x)}{B}$ l'estimateur du bagging.
— $\theta^*_{bag,B}(x) = \lim_{B\to\infty} \frac{\sum_{k=1}^{B} \hat{\theta}_b(x)}{B} = \mathbb{E}^*[\hat{\theta}_b(x)]$ l'estimateur du bagging théorique calculé selon la loi du boostrap.

On a vu précédemment que le Bagging n'influe pas sur les estimateurs stables. En effet, quand l'estimateur est linéaire l'estimateur du Bagging reste le même.

Example 3.1. *Bagging linéaire*

Regardons le cas simple de l'estimateur empirique : Soit $\hat{\theta}_n(x) = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y_n}$, si on considère Y_1^* un tirage boostrap d'un élément de $\mathbf{y}= \{Y_1, ..., Y_n\}$

$$Alors \quad \theta^*_{bag,1} = \mathbb{E}^*[Y_1^*] = \bar{Y_n}$$

On en conclut que les cas intéressants du bagging sont ceux avec des estimateurs non linéaires.

4 Bagging et estimateur non linéaire

On peut distinguer deux types d'estimateurs. Le premier type sont les estimateurs qui sont infiniment dérivables. On peut alors avoir une expression asymptotique de l'estimateur qui sera décomposée en une partie linéaire et une autre de termes de hauts degrés. Ces estimateurs sont obtenus généralement en cherchant le θ qui maximise une certaine fonction objective comme la log de vraisemblance. Il existe toute une théorie [voir H.Friedman, Peter Hall, 2000] qui montre que le Bagging réduit la variabilité de la partie non linéaire en la remplaçant par une estimation et garde la partie linéaire inchangée. Ce qui explique l'efficacité du Bagging dans les modèles non linéaires comme les arbres de décisions.

Cependant, dans ce qui suit nous allons nous intéresser à des estimateurs non dérivables voir non continus. Ces estimateurs sont très difficiles à étudier dans la théorie, nous allons alors étudier quelques cas particuliers.

4.1 Cas 1 : Bagging appliqué à une indicatrice

Une façon de voir comment le bagging fonctionne peut se démontrer avec un exemple jouet. Considérons l'estimateur :

$$\theta_n(x) = \mathbb{1}_{\bar{Y_n}\leq x}, x \in \mathbb{R}$$

Pour un x fixé, $\hat{\theta}_n(x)$ est une indicatrice avec $\bar{Y_n}$ comme seuil. L'estimateur résultant du bagging sera alors une somme d'indicatrice avec un seuil $\bar{Y_n}^*$ qui varie autour de $\bar{Y_n}$.

Par le théorème central limite, on voit facilement que :

$$\sqrt{n}(\bar{Y_n} - \mu) \to \mathcal{N}(0, \sigma^2)$$

avec $\mu = \mathbb{E}(Y_1)$, et $\sigma^2 = \mathbb{V}(Y_1)$. Si on choisi un x dans le $\sqrt{n}$ voisins de μ par exemple pour c constante, alors :

$$x = x_n(x) = \mu + \sqrt{n}c\sigma$$

Alors d'après le TCL on a :

$$\theta_n(x_n(c)) \approx \mathbb{1}_{Z \leq c}, \qquad Z \sim \mathcal{N}(0,1)$$

En notant ϕ la fonction de répartion de la loi normale centrée réduite, on a :

$$\mathbb{E}[\hat{\theta}_n(x)] = \mathbb{E}[\mathbb{1}_{\hat{Y}_n \leq x_n(x)}] = \mathbb{P}(Z \leq c) = \phi(c) \quad (n \to \infty)$$

$$\mathbb{V}(\hat{\theta}_n(c)) = \phi(c)(\phi(c) - 1) \qquad (n \to \infty)$$

Comme la variance ne tends pas vers 0, $\hat{\theta}_{n,B}(x)$ est instable. Il prends des valeurs 0 et 1 avec une probabilité positive meme quand n tends vers l'infini.

Regardons maintenant l'estimateur $\theta^*_{bag,B}$ du bagging :

$$\begin{aligned} \theta^*_{bag,B}(x) &= \mathbb{E}^*[\mathbb{1}_{\bar{Y_n}^* \leq x_n(c)}] = \mathbb{E}^*[\mathbb{1}_{n^{1/2}(\hat{Y}^*_n - \hat{Y}_n)/\sigma \leq n^{1/2}(x_n(x) - \hat{Y}_n)/\sigma}] \\ &= \phi(n^{1/2}(x_n(c) - \bar{Y_n})) + o(1) \\ &\approx \phi(c - Z), \quad Z \sim \mathcal{N}(0,1) \end{aligned}$$

La première approximation (deuxième ligne) vient du fait que le bagging marche pour la moyenne empirique $\bar{Y_n}$ [voir Giné et Zinn(1990)].

On en déduit que le bagging produit une fonction de décision plus douce que l'estimateur standard. (voir fig 1)

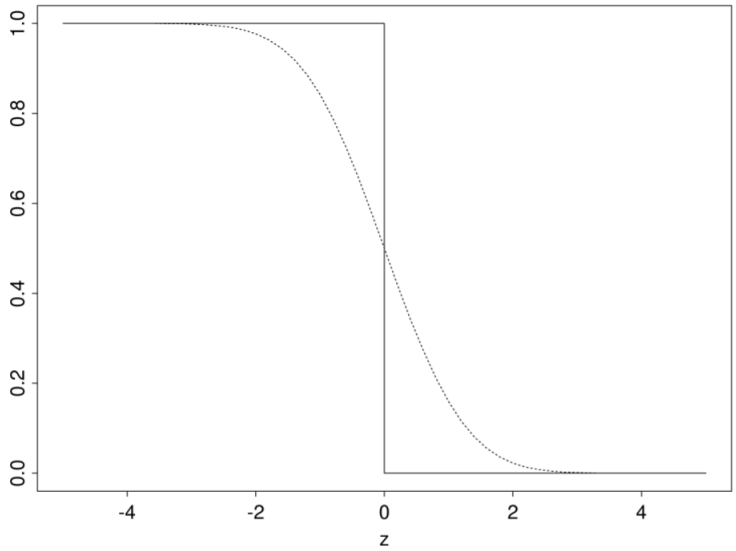

FIGURE 1 – Pour $x = x_n(0)$, estimateur standard (fermé) estimateur bagging (pointillé)

Observons le cas où $x = x_n(0) = \mu$, c'est exactement la partie la plus instable car la $\mathbb{V}(\hat{\theta}_n(x))$ est maximale. On a alors :

$$\hat{\theta}_{bag,B}(x_n(0)) \to \phi(-Z) = \phi(Z = U), \quad U \sim Uniforme[0,1]$$

Alors,

$$\mathbb{E}[\hat{\theta}_{bag,B}(x_n(0))] \to \mathbb{E}[U] = \frac{1}{2} \quad (n \to \infty)$$

$$\mathbb{V}[\hat{\theta}_{bag,B}(x_n(0))] \to \mathbb{V}[U] = \frac{1}{12} \quad (n \to \infty)$$

L'estimateur Bagging est alors asymptotiquement non biaisé ($\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n(x_n(0))]) = \phi(0) = \frac{1}{2}]$), et la variance asymptotique est réduite d'un facteur 3. Si $x \neq \mu$, le bagging réduit toujours la variance mais rajoute un petit biais.

4.2 Cas 2 : Bagging appliqué au 1-plus-proche voisin

Dans cette section, nous allons essayer de démontrer un résultat très fort. Nous montrerons que l'estimateur résultant du bagging des 1-PPV est universellement consistant.

On rappelle que pour un échantillon $\mathcal{D}_n = \{X_1, Y_1), ...(X_n, Y_n)\}$, l'estimateur 1-PPV est $r(x) = Y_{(1)}(x)$ où $Y_{(1)}(x)$ est le label correspondant au $X_{(1)}(x)$ qui est le plus proche (selon la distance euclidienne) de x parmi $\{X_1, ...X_n\}$.

On considère alors $\left\{(X_{(1)}, Y_{(1)}), ...(X_{(n)}, Y_{(n)})\right\}$ l'échantillon ordonné :

$$||X_{(1)} - x|| \leq ||X_{(2)} - x|| \leq ... \leq ||X_{(n)} - x||$$

Donc on définit $r_{k_n}(x)$ l'estimateur 1-PPV pour $\mathcal{S}_n$ un échantillon boostrap de $\mathcal{D}_n$ tel que $Card(\mathcal{S}_n) = k_n$. L'estimateur obtenu après Bagging est alors :

$$r^*{}_n(x) = \mathbb{E}^*[r_{k_n}(x)]$$

où $\mathbb{E}^*$ est l'espérance selon la loi du tirage boostrap.

Lemme 1. *Soit $r_n(x)$ l'estimateur du 1-PPV sur l'echantillon $\mathcal{D}_n$ et $r^*{}_n(x)$ son estimateur après Bagging alors $r^*{}_n(x)$ est un estimateur de type plus proche voisins pondéré donc :*

$$r^*{}_n(x) = \sum_{i=1}^{n} V_i Y_{(i)}(x) \qquad\qquad V_n \leq ... \leq V_2 \leq V_1$$

$$V_i = \mathbb{P}(\textit{i-ème voisin de x soit le 1-PPV de x dans le boostrap})$$

Démonstration. Soit x fixé, on note $r_{k_n}(x, \mathcal{S}_m)$ l'estimateur du 1-PPV sur l'échantillon boostrap $\mathcal{S}_m$, on obtient :

$$\mathbb{E}^*[r_{k_n}(x)] = \frac{1}{B} \sum_{m=1}^{B} r_{k_n}(x, \mathcal{S}_m) \tag{1}$$

De plus on remarque que :

-$r_{k_n}(x, \mathcal{S}_m) = Y_{(1)}(x)$ Si $Y_{(1)}(x) \in \mathcal{S}_m$

-$r_{k_n}(x, \mathcal{S}_m) = Y_{(2)}(x)$ Si $Y_{(1)}(x) \notin \mathcal{S}_m$ et $Y_{(2)}(x) \in \mathcal{S}_m$

-$r_{k_n}(x, \mathcal{S}_m) = Y_{(3)}(x)$ Si $Y_{(1)}(x) \notin \mathcal{S}_m$ et $Y_{(2)}(x) \notin \mathcal{S}_m$ et $Y_{(3)}(x) \in \mathcal{S}_m$ ...

Donc on a :

$$r_{k_n}(x, \mathcal{S}_m) = \sum_{i=1}^{n} Y_{(i)}(x) \mathbb{1}_{Y_{(i)}(x) \in \mathcal{S}_m} \prod_{j=1}^{i-1} \mathbb{1}_{Y_{(j)}(x) \notin \mathcal{S}_m}$$

Puis en remplaçant dans (1) on obtient :

$$\begin{aligned}
\mathbb{E}^*[r_{k_n}(x)] &= \lim_{B \to \infty} \frac{1}{B} \sum_{m=1}^{B} \sum_{i=1}^{n} Y_{(i)}(x) \mathbb{1}_{Y_{(i)}(x) \in \mathcal{S}_m} \prod_{j=1}^{i-1} \mathbb{1}_{Y_{(j)}(x) \notin \mathcal{S}_m} \\
&= \sum_{i=1}^{n} Y_{(i)}(x) \lim_{B \to \infty} \frac{1}{B} \sum_{m=1}^{B} \mathbb{1}_{Y_{(i)}(x) \in \mathcal{S}_m} \prod_{j=1}^{i-1} \mathbb{1}_{Y_{(j)}(x) \notin \mathcal{S}_m}
\end{aligned}$$

Par la loi forte des grands nombres, on a : $V_i = \mathbb{E}[\mathbb{1}_{Y_{(i)}(x) \in \mathcal{S}_m} \prod_{j=1}^{i-1} \mathbb{1}_{Y_{(j)}(x) \notin \mathcal{S}_m}]$ qui représente la probabilité que dans l'échantillon $\mathcal{S}_m$ l'estimateur 1-NN soit $Y_{(i)}(x)$. □

Lemme 2.

V_i dans le cas avec remise est :

$$V_i = (1 - \frac{i-1}{n})^{k_n} - (1 - \frac{i}{n})^{k_n}$$

V_i dans le cas sans remise est :

$$V_i = \begin{cases} \frac{\binom{n-i}{k_n-1}}{\binom{n}{k_n}} & si \qquad i \leq n - k_n + 1 \\ 0 & sinon. \end{cases}$$

Démonstration.

- La valeur des V_i s'obtient facilement en faisant un peu de combinatoire. Pour le cas sans remise, il faut juste remarquer que la probabilité que le i-ème voisin de X soit le plus proche voisins dans l'échantillon boostrap est juste la probabilité de tirer le i-ème voisins puis de tirer $k_n - 1$ autres points parmi les $n - i$ plus éloignés donc le nombre de possibilités est : $1 * \binom{n-i}{k_n-1}$.

- Pour le cas avec remise, il faut juste voir que cette probabilité est égale à la probabilité que le $X_{(i)}$ sois tiré au moins une fois et qu'aucun $X_{(j)}, j < i$ ne soient tirés. On peut voir cette probabilité comme la différence entre la probabilité de tirer avec remise k_n parmi i-1 derniers points moins la probabilité de tirer parmi k_n les i derniers points. □

Théorème 4.1. *Si $k_n \to \infty$ et $\frac{k_n}{n} \to 0$ alors l'estimateur Bagging du 1-PPV est universellement consistant.*

Démonstration. Le Lemme 1 nous assure que l'estimateur Bagging du 1-PPV est un estimateur de type moyenne locale. Ainsi, pour prouver la consistance, il nous faut juste vérifier les conditions du théorème de Stone. On considérera seulement le cas du Bagging sans remise, la preuve du second cas étant analogue.
Par soucis d'harmonisation des notations, on pose $V_i = W_{ni}$.

- La condition 1 est vérifiée puisque :

$$\begin{aligned}\max_{1\leq i\leq n} W_{ni} &= W_{n1}\\ &= 1-(1-\frac{1}{n})^{k_n}\\ &= 1-e^{k_n\ln(1-\frac{1}{n})}\\ &= 1-e^{\frac{k_n}{n}}\to 0. \quad (car\ \ \frac{k_n}{n}\to 0)\end{aligned}$$

- Pour la condition 2, [Devroye et Gjörfi Lugasi, 1996, Lemma 5-1] ont montré que si $\frac{k_n}{n}\to 0$ alors $\|X-X_{k_n}\|\leq a\ p.s$;

$$\begin{aligned}\sum_{i=1}^{n} W_{ni}\mathbb{1}_{[\|X_i-X\|>a]} &\leq \sum_{i=k_n+1}^{n} W_{ni}\\ &\leq \sum_{i=k_n+1}^{n}(1-\frac{i-1}{n})^{k_n}-(1-\frac{i}{n})^{k_n}\\ &\leq (1-\frac{k_n}{n})^{k_n}\ \textit{par téléscopage}\\ &\leq e^{k_n\ln(1-\frac{k_n}{n})}\sim e^{\frac{-k_n^2}{n}}\to 0. \quad (car\ \ \frac{k_n}{n}\to 0)\end{aligned}$$

  On a donc $\sum_{i=1}^{n} W_{ni}\mathbb{1}_{[\|X_i-X\|>a]}\to 0$. La somme étant bornée, par convergence dominée, on obtient que : $\mathbb{E}[\sum_{i=1}^{n} W_{ni}\mathbb{1}_{[\|X_i-X\|>a]}]\to 0$

- Pour ce qui est de la condition 3, on sait que W_{ni} est une fonction décroissante en i. On peut alors trouver des coefficients $(a_i)_{1\leq i\leq n}$ avec $a_i=W_{n(i+1)}-W_{ni}$ et donc on a que $\sum_{i=1}^{n} a_i=W_{ni}$ et $\sum_{j=1}^{n} ja_j=\sum_{i=1}^{n} W_{ni}=1$ Alors :

$$\begin{aligned}\mathbb{E}[\sum_{i=1}^{n} W_{ni}f(X_i)] &= \mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{n} a_j f(X_i)]\\ &= \mathbb{E}[\sum_{j=1}^{n} a_j\sum_{i=1}^{n} f(X_i)]\\ &\leq c\sum_{i=1}^{n} a_j j\mathbb{E}[f(X)] \quad \text{Voir Stone [1977]}\\ &\leq c\mathbb{E}[f(X)]\end{aligned}$$

□

5 Version améliorée du Bagging : Forêts aléatoires

En 2004, Breiman propose une amélioration du Bagging pour les arbres de décision par l'ajout d'une composante aléatoire. Le but étant de rendre les arbres encore plus indépendants, il propose de rajouter de l'aléa dans la sélection des variables qui interviennent

dans les modèles. On parle alors de forêts aléatoires. Depuis la publication, cette méthode bien que difficile à interpréter est devenu la cible à abattre en terme de performance.

Nous présentons très brièvement la famille des Random Forest, pour une étude plus approfondie [voir Gérard Biau et Erwan Scornet, A random forest tour]. Ce type de forêt est construit avec l'algorithme CART. Le principe de l'algorithme CART est de partitionner récursivement l'espace engendré par les variables explicatives de façon binaire (voir fig 2).

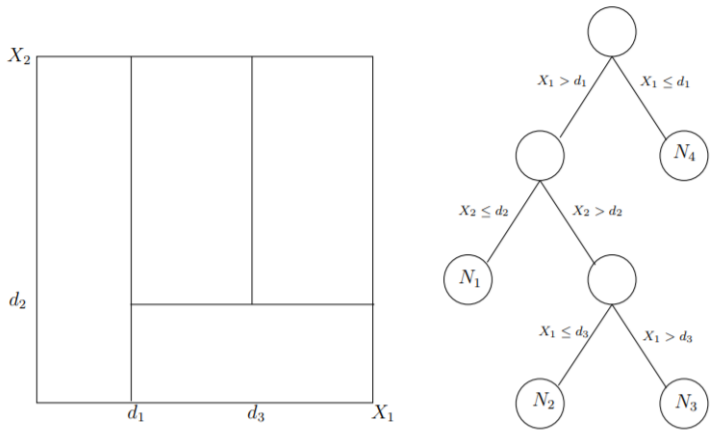

FIGURE 2 – Arbre Cart

Notons que les coupures (X_j, d) sont choisies de façon à minimiser une fonction de coût particulière (indice de Gini, ou la variance des noeuds fils dans la régression).

Nous proposons une variante des CART pour la construction de notre forêt aléatoire. A chaque étape de CART, le Bagging est appliqué en sélectionnant de manière aléatoire m variables parmi les p dont nous disposons et la meilleure coupure est choisie sur ces m variables.

Algorithm 2: Forêts aléatoires

input :
- x l'observation à prédire
- m le nombre de variables explicatives pour découper
- d_n l'échantillon
- B le nombre d'arbres

for $k \leftarrow 1$ **to** B **do**

1. tirer un échantillon boostrap d_n^k de d_n
2. Construire un arbre de CART sur cet échantillon boostrap, chaque coupure est sélectionner en minimisant la fonction de coût de CART sur l'ensemble des m variables explicatives choisies au harsard parmi p. On $h_k(.)$ l'arbre construit.

end

output: $h(x) = \frac{\sum_{k=0}^{B} h_k(x)}{B}$

Références

[1] Gérard Biau, Frédéric Cérou, et Arnaud Guyader. *On the Rate of Convergence of the Bagged Nearest Neighbor Estimate*. Journal of Machine Learning Research 11, (2010).

[2] Gérard Biau. *Cours apprentissage statistique*.
`https://www.ljll.math.upmc.fr/MathModel/enseignement/polycopies/gerardbiau.pdf`

[3] L. Devroye. *On the almost everywhere convergence of nonparametric regression function estimates*. The Annals of Statistics.

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*

[5] L. Breiman. *Bagging predictors*. Machine Learning, (1996).

[6] Brian M. Steele. *Exact bootstrap k-nearest neighbor learners*. Machine Learning, (2009).

[7] Peter Bühlmann et Bin Yu. *Analyzing Bagging* The Annals of Statistics Vol. 30, No. 4 (Aug., 2002).

[8] Gérard Biau, Luc Devroye et Gàbor Lugosi. *Consistency of Random Forests and Other Averaging Classifiers*. Journal of Machine Learning Research 9 (2008).

[9] Jerome H.Friedman, Peter Hall, *On Bagging and Nonlinear Estimation*, (2000).

[10] Gérard Biau, Luc Devroye. *On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification*. (2010).