Arbres et Forêts aléatoires

Nicolas Verzelen, Alexis Joly

INRA,Inria

M2 MIASH

Sommaire

Méthodes basées sur des arbres

- Nous décrivons ici des méthodes *basées sur des arbres* pour la classification et la régression.
- Cela implique de *stratifier* ou *segmenter* l'espace des prédicteurs en un certain nombre de régions simples
- Comme l'ensemble des règles de partitionnement peuvent être résumées par un arbre, ce type d'approches sont connues comme des méthodes à *arbres de décision*

Pours et contres

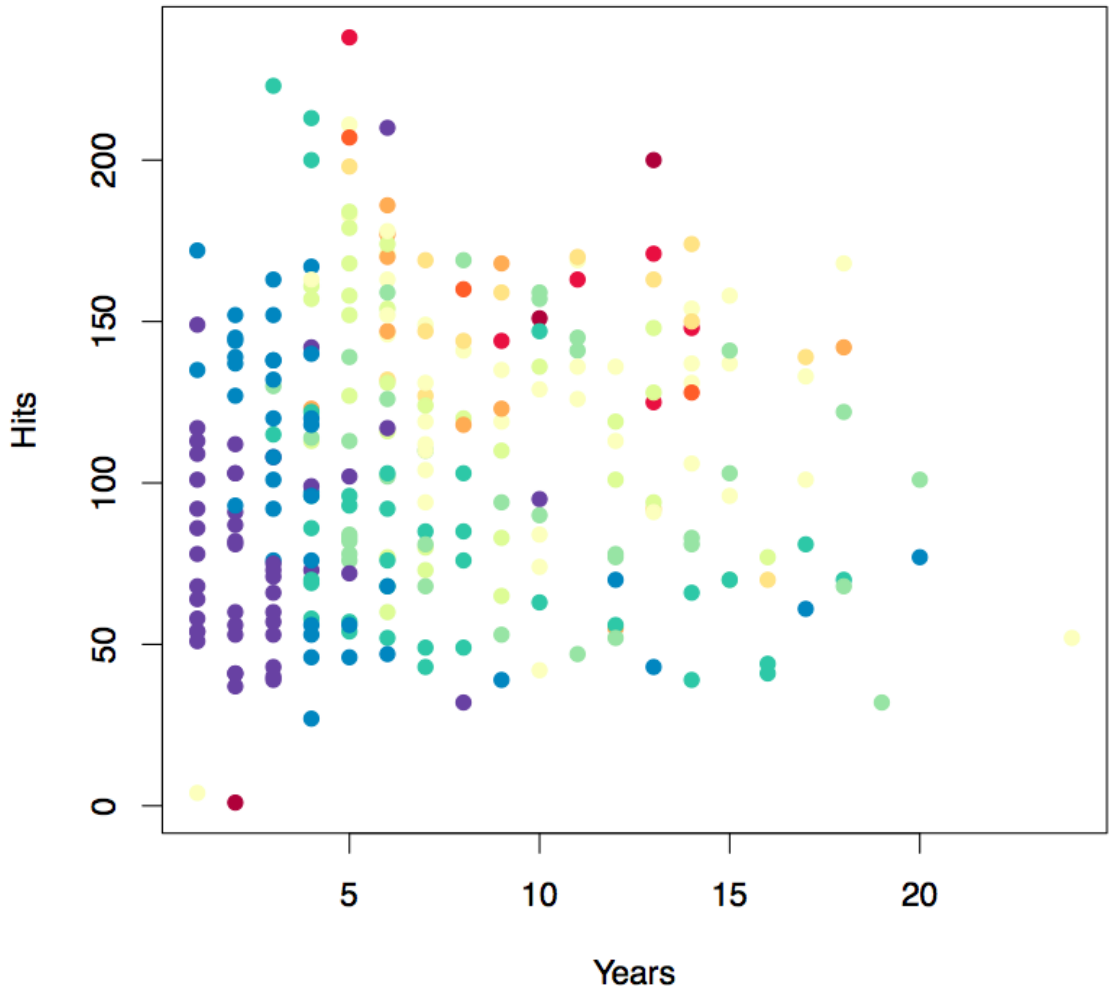
- Les méthodes basées sur des arbres sont simples et utiles pour l'interprétation.
- Cependant, elles ne sont pas capables de rivaliser avec les meilleures approches d'apprentissage supervisé en terme de qualité de prédiction
- Nous discuterons donc aussi de *bagging* et de *forêts aléatoires (random forests)*. Ces méthodes développent de nombreux arbres de décision qui sont ensuite *combinés* pour produire une réponse consensus.

Les bases des arbres de décision

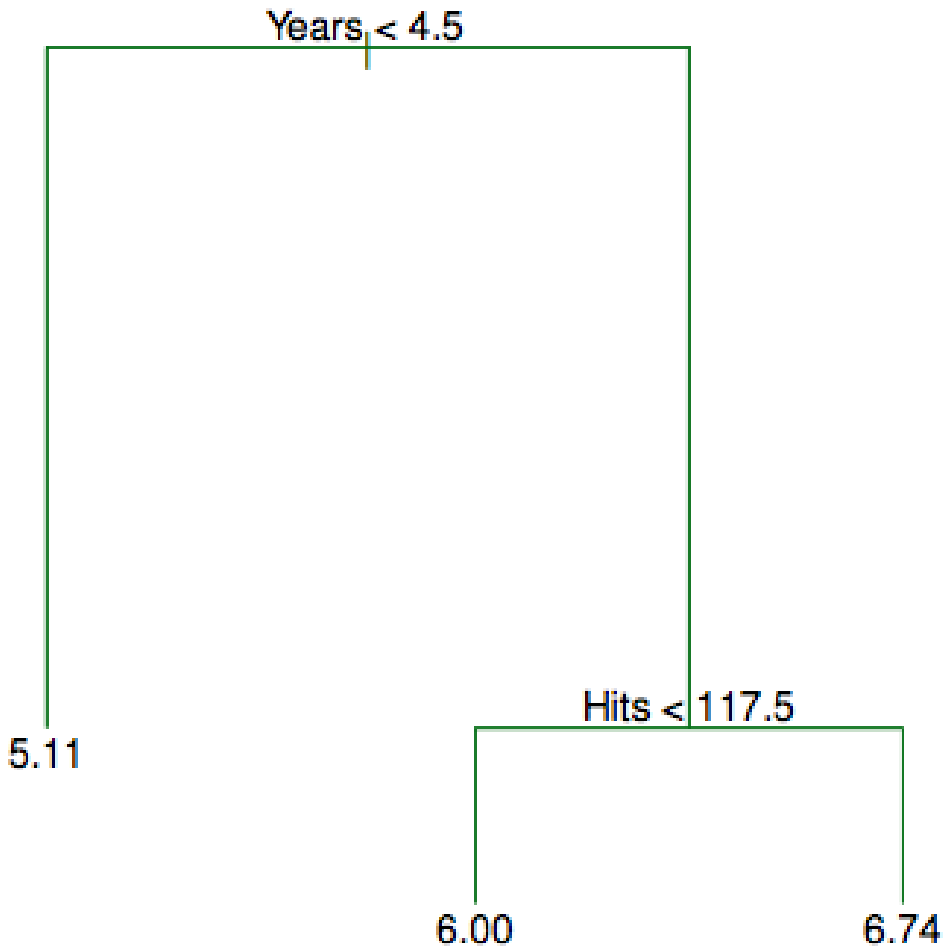
- Les arbres de décision sont utiles aussi bien pour des problèmes de régression que de classification.
- Nous commençons par présenter des problèmes de régression et nous viendrons ensuite à la classification.

Données de salaire au baseball : comment les stratifier ?

Le salaire est codé par des couleurs : les faibles valeurs sont en bleu, puis vert, les plus fortes valeurs en orange puis rouge.



Un arbre de décision sur ces données

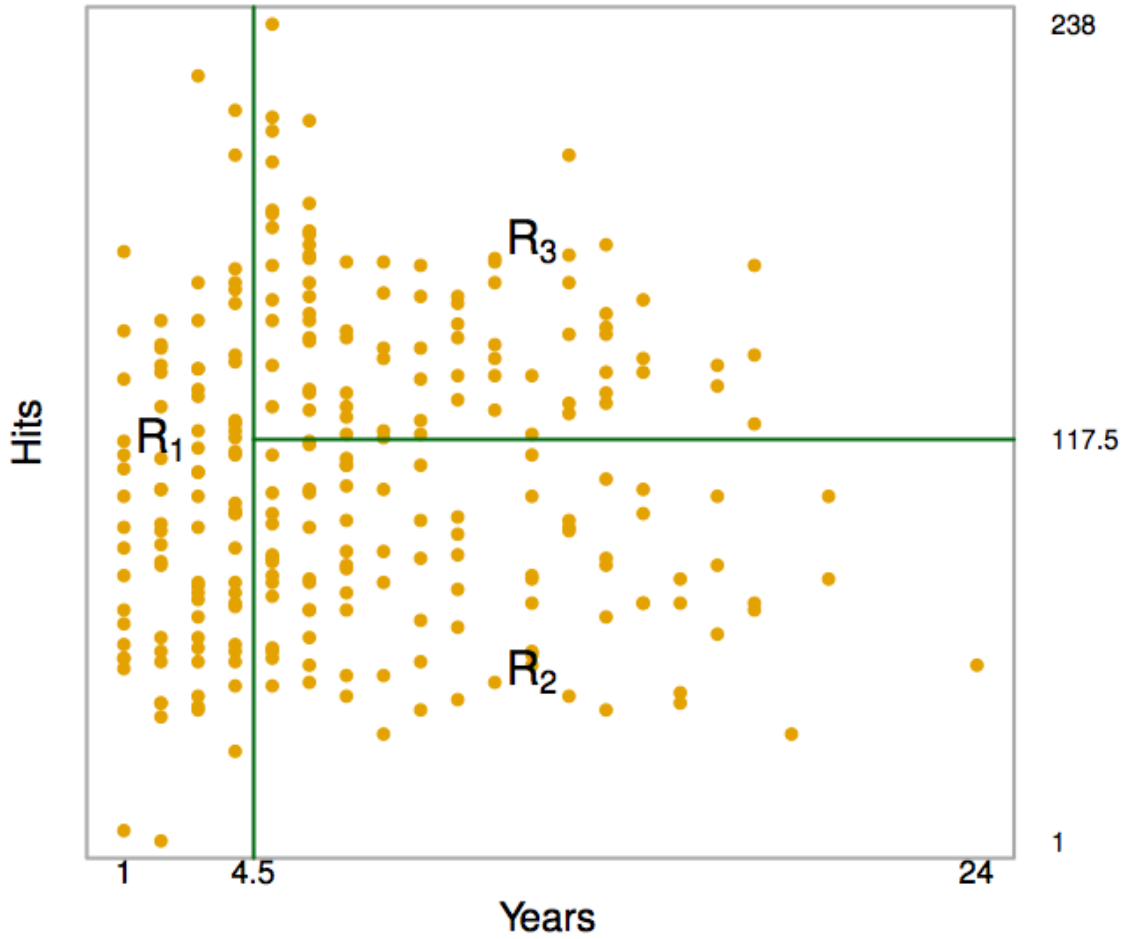


Détails sur la précédente figure

- C'est un arbre de régression pour prédire le **log** des salaires des joueurs, basé sur
  - l'expérience (`Years`)
  - le nombre de succèes (`Hits`)
- Pour chaque nœud interne, l'étiquette (de la forme $X^{(j)} < t_k$) indique la branche de gauche émanant du nœud, et la branche droite correspond à $X^{(j)} \geqslant t_k$.
- Cet arbre a deux nœuds internes et trois nœuds terminaux ou feuilles. Le nœud le plus haut dans la hiérarchie est la racine.
- L'étiquette des feuilles est la réponse moyenne des observations qui satisfont aux critères pour la rejoindre.

Résultats

- En tout, l'arbre distingue trois classes de joueurs en partitionnant l'espace des prédicteurs en trois régions : $R_1 = \{X : \texttt{Years} < 4.5\}$, $R_2 = \{X : \texttt{Years} \geqslant 4.5, \texttt{Hits} < 117.5\}$ et $R_3 = \{X : \texttt{Years} \geqslant 4.5, \texttt{Hits} \geqslant 117.5\}$.



Arbres de décision CART

Un arbre binaire de décision **CART - Classification And Regression Tree -** est un algorithme de moyennage local par partition (moyenne ou vote à la majorité sur les éléments de la partition), dont la partition est construite par divisions successives au moyen d'hyperplans orthogonaux aux axes de $\mathscr{X} = \mathbb{R}^p$, dépendant des (X_i, Y_i).

Les éléments de la partition d'un arbre sont appelés les **nœuds terminaux** ou les **feuilles** de l'arbre.

L'ensemble $\mathscr{X} = \mathbb{R}^p$ constitue le **nœud racine**. Puis chaque division définit deux nœuds, les **nœuds fils à gauche et à droite**, chacun soit terminal, soit interne, par le choix conjoint :

- d'une variable explicative $X^{(j)}$ $(j = 1 \dots p)$,
- d'une valeur seuil pour cette variable.

Ce choix se fait par maximisation du gain d'homogénéité, défini à l'aide d'une **fonction d'hétérogénéité** H, sur les observations de la variable à expliquer.

Pour un nœud k, si k_g, k_d désignent les nœuds fils à gauche et à droite issus de la division de ce nœud, on choisit la variable explicative et le seuil de la variable explicative maximisant :

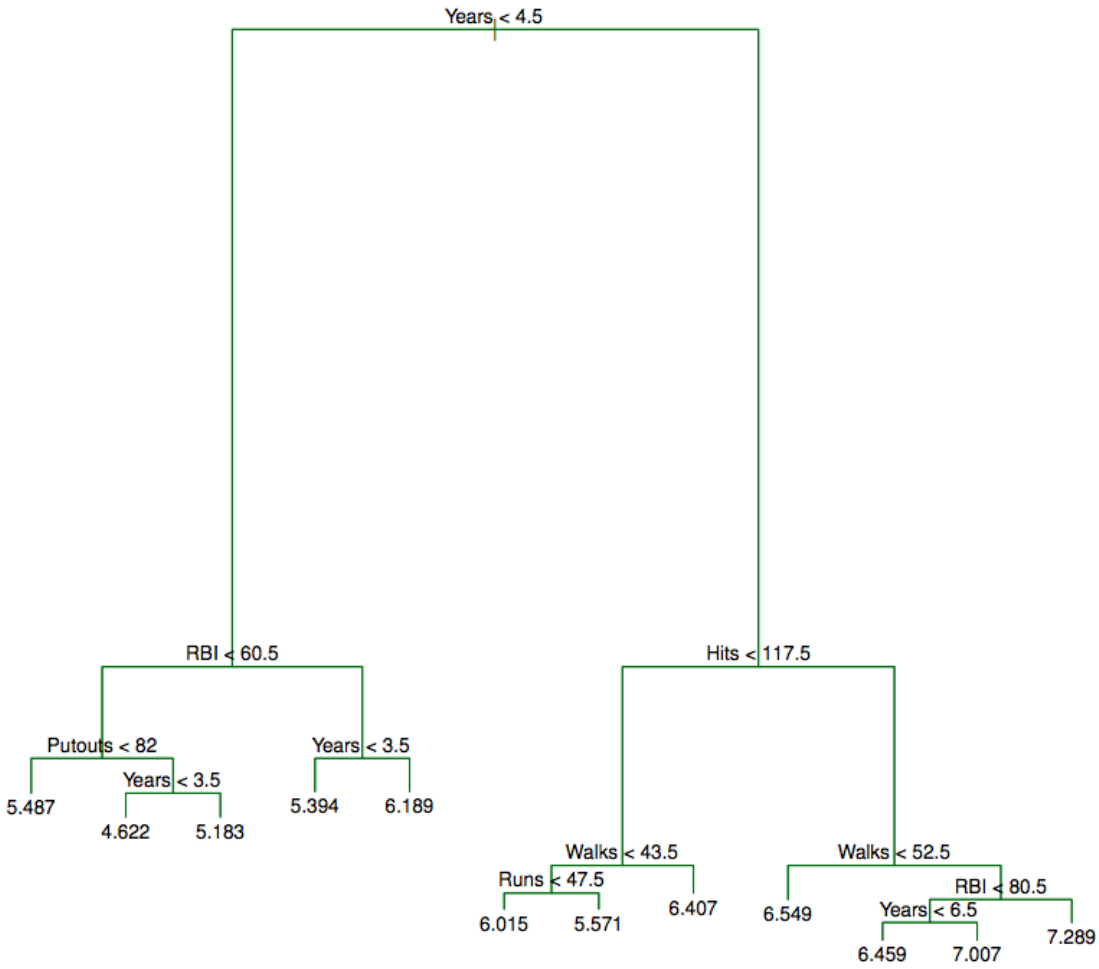
- En régression : $H_k - (H_{k_g} + H_{k_d})$, avec $H_k =$ la variance empirique des y_i du nœud k,
- En discrimination binaire : $H_k - (p_{k_g} H_{k_g} + p_{k_d} H_{k_d})$, avec p_k la proportion d'observations dans le nœud k, et $H_k = p_k^1(1 - p_k^1) + p_k^{-1}(1 - p_k^{-1}) = 1 - (p_k^1)^2 - (p_k^{-1})^2$, où p_k^δ est la proportion de y_i du nœud k égaux à $\delta \rightarrow$ Indice de Gini

Arbres de décision : sélection

Étape 1 : construction de l'arbre maximal T_{max}. T_{max} correspond à la partition qui ne peut plus être subdivisée, soit parce que ses parties contiennent moins d'observations qu'un nombre fixé au départ (entre 1 et 5), soit parce qu'elles ne contiennent que des observations de même réponse (homogènes).

Baseball : l'arbre $\mathscr{T}_0$

Avec tous les prédicteurs du jeu de données (problème de surapprentissage)



Arbres de décision : sélection

Étape 2 : **élagage**. De la suite d'arbres qui a conduit à l'arbre maximal, on extrait une sous-suite d'arbres emboîtés à l'aide du critère pénalisé suivant : pour $\alpha \geqslant 0$, $\mathrm{crit}_\alpha(T) = \widehat{\mathscr{R}}_n(\widehat{\phi}_T) + \alpha|T|/n$, où $\widehat{\mathscr{R}}_n(\widehat{\phi}_T)$ est le risque apparent de la règle de régression ou de discrimination associée à l'arbre T (taux de mal classés en discrimination, erreur quadratique moyenne empirique en régression) et $|T|$ est la taille de l'arbre c'est-à-dire son nombre de feuilles.

Théorème (Breiman *et al.*)

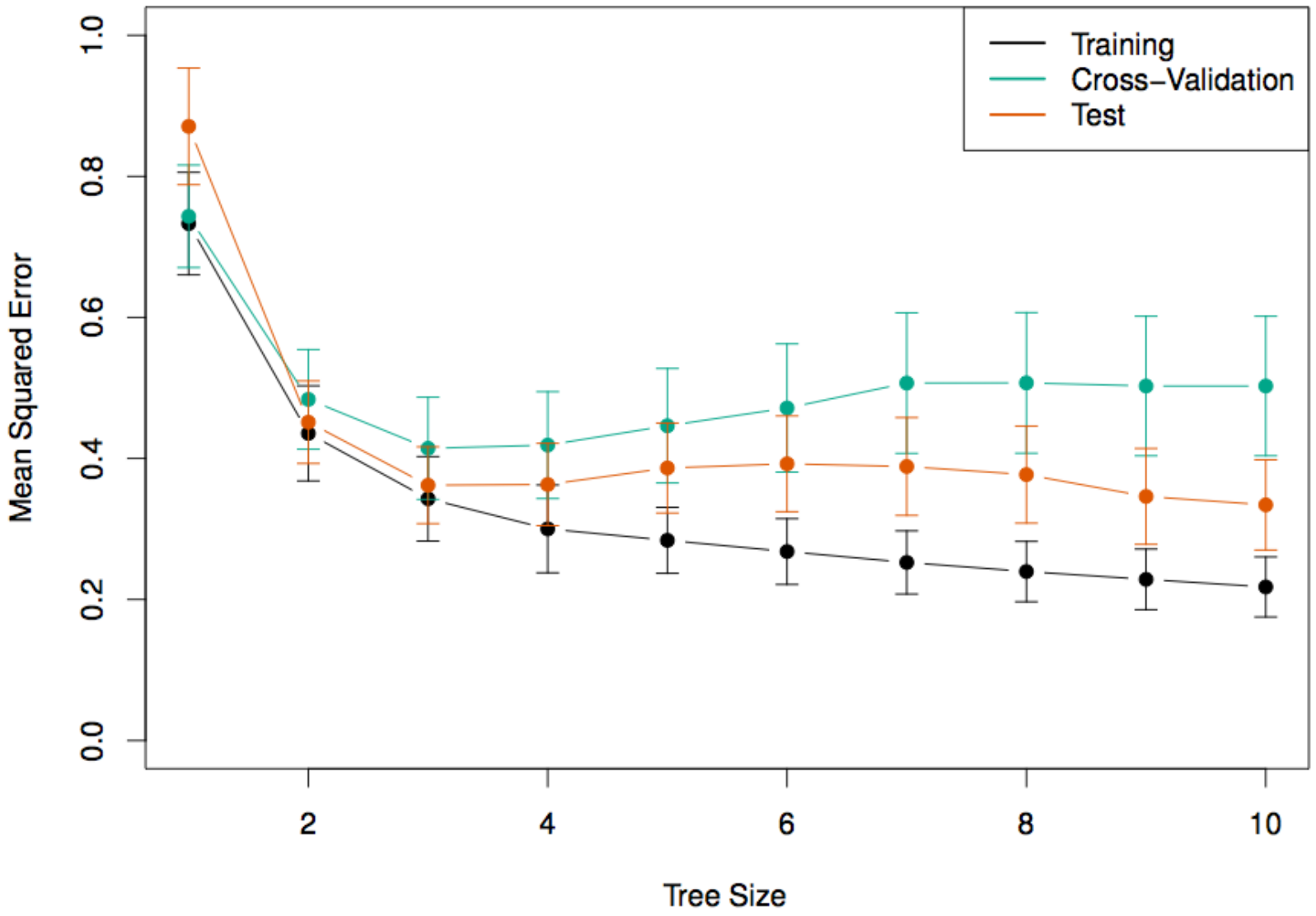
Il existe une suite finie $\alpha_0 = 0 < \alpha_1 < \ldots < \alpha_M$ et une suite associée de sous-arbres emboîtés telles que : pour $\alpha \in [\alpha_m, \alpha_{m+1}[$,

$$\text{argmin}_T \text{crit}_\alpha(T) = T_m.$$

Étape 3 : sélection du meilleur arbre dans la suite d'arbres obtenue par élagage, par estimation du risque de chaque arbre de la suite.

$\hookrightarrow$ Cette estimation se fait par validation croisée hold-out ou K blocs.

Baseball, choix de α



Avantages et inconvénients des arbres

- ▲ Les arbres sont faciles à expliquer à n'importe qui. Ils sont plus faciles à expliquer que les modèles linéaires
- ▲ Les arbres peuvent être représentés graphiquement, et sont interprétables même par des non-experts
- ▲ Ils peuvent gérer des prédicteurs discrets sans introduire des variables binaires
- ▼ Malheureusement, ils n'ont pas la même qualité prédictives que les autres approches de ce cours.

Cependant, en agrégeant plusieurs arbres de décision, les performances prédictives s'améliorent substantiellement.

Arbres de décision : interprétation

La représentation graphique de l'arbre permet une interprétation facile de l'algorithme de prédiction construit, et sa construction est algorithmiquement efficace, ce qui fait le succès de la méthode CART d'un point de vue métier.

Mises en garde

- L'arbre sélectionné ne dépend que de quelques variables explicatives, et est souvent interprété (à tort) comme une procédure de sélection de variables.
- L'arbre souffre d'une grande instabilité (fléau de la dimension, sensibilité à l'échantillon).
- La qualité de prédiction d'un arbre est souvent médiocre comparée à celle d'autres algorithmes.

$\hookrightarrow$ Agrégation d'arbres !

Agrégation d'algorithmes de prédiction

Les **méthodes d'agrégation** d'algorithmes de prédiction se décrivent de la façon suivante.

- Construction d'un grand nombre d'algorithmes de prédiction simples $\widehat{f}_b$, $b = 1 \dots B$.
- Agrégation ou combinaison de ces algorithmes sous la forme : $\widehat{f} = \sum_{b=1}^{B} w_b \widehat{f}_b$ ou $signe\left(\sum_{b=1}^{B} w_b \widehat{f}_b\right)$.

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

Le **bagging** s'applique à des algorithmes instables, de variance forte.

Le **boosting** s'applique à des algorithmes fortement biaisés, mais de faible variance.

Bagging (Breiman 1996)

Le bagging regroupe un ensemble de méthodes s'appliquant à des problèmes de régression ou de discrimination, introduites par Leo Breiman en 1996. Le terme bagging provient de la contraction de **B**ootstrap **agg**regat**ing**.

Rappels des notations :

Données observées de type entrée-sortie :
$d_1^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ avec $x_i \in \mathbb{R}^p$, $y_i \in \mathscr{Y}$ ($\mathscr{Y} = \mathbb{R}$ en régression, $\mathscr{Y} = \{-1, 1\}$ en discrimination binaire)
$D_1^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, (X_i, Y_i) i.i.d. $\sim P$ (totalement inconnue).

Objectif : prédire la sortie y associée à une nouvelle entrée x, où x est une observation de la variable X, $(X, Y) \sim P$ indépendant de D_1^n.

Bagging(regression)

On note

- η^* la fonction de régression définie par $\eta^*(x) = \mathbb{E}[Y|X = x]$ (minimisant le risque quadratique)

Considérons un algorithme de régression $\widehat{\eta}$.
Le bagging consiste à 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

Pour $x \in \mathbb{R}^p$, $\mathbb{E}\left[(\widehat{\eta}(x) - \eta^*(x))^2\right] \geqslant (\mathbb{E}\left[\widehat{\eta}(x)\right] - \eta^*(x))^2 + \mathrm{Var}\left(\widehat{\eta}(x)\right))$.

Si les algorithmes de régression $\widehat{\eta}_1, \ldots, \widehat{\eta}_B$ étaient i.i.d. on aurait :

$$\mathbb{E}\left[\widehat{\eta}(x)\right] = \mathbb{E}\left[\widehat{\eta}_b(x)\right] \text{ et } \mathrm{Var}\left(\widehat{\eta}(x)\right) = \mathrm{Var}\left(\widehat{\eta}_b(x)\right)/B$$

$\hookrightarrow$ biais identique, mais variance diminuée.

Attention : En pratique, les algorithmes ne peuvent pas être i.i.d. puisqu'ils sont construits sur le même échantillon D_1^n !

Si les algorithmes sont seulement identiquement distribués, si $\rho(x)$ est le cœfficient de corrélation entre $\widehat{\eta}_b(x)$ et $\widehat{\eta}_{b'}(x)$,

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

question : Comment construire des algorithmes peu corrélés entre eux, alors qu'ils sont a priori construits sur le même échantillon ?

- **Solution de Breiman** : construire un même algorithme de base sur des échantillons bootstrap de D_1^n.

Algorithme du bagging

Considérons :

- un algorithme de régression $D \mapsto \eta_D$, ou de discrimination $D \mapsto \phi_D$
- un nombre B (grand) d'échantillons bootstrap de D_1^n : $D_{m_n}^{*1} \dots D_{m_n}^{*B}$, de taille $m_n \leqslant n$, indépendants les uns des autres conditionnellement à D_1^n.

Pour $b = 1 \dots B$,

$$\widehat{\eta}_b = \eta_{D_{m_n}^{*b}} \quad \text{ou} \quad \widehat{\phi}_b = \phi_{D_{m_n}^{*b}}.$$

L'algorithme de bagging consiste à agréger les algorithmes $\widehat{\eta}_1, \dots, \widehat{\eta}_B$ ou $\widehat{\phi}_1, \dots, \widehat{\phi}_B$ de la façon suivante :

- $\widehat{\eta} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\eta}_b$ en régression
- $\widehat{\phi} = signe(\sum_{b=1}^{B} \widehat{\phi}_b)$ (vote à la majorité) en discrimination binaire

Forêts aléatoires (Breiman and Cutler 2005)

Le terme de forêt aléatoire se rapporte initialement à l'agrégation, au sens large, d'arbres de régression ou de discrimination.
Désormais, il désigne le plus souvent une forêt aléatoire **Random Input**, méthode introduite par Breiman et Cutler (2005)
`http://www.stat.berkeley.edu/users/breiman/RandomForests/`

$\hookrightarrow$ 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 $m < p$ de variables explicatives,
- partitionner le nœud en un nœud fils à gauche et un nœud fils à droite sur la base de la "meilleure" de ces m variables explicatives (sélectionnée par les critères de la méthode CART).

- Diminution encore plus importante de la corrélation entre les arbres.

input :

x : l'entrée dont on veut prédire la sortie;

d_1^n : l'échantillon observé;

m : le nombre de variables explicatives sélectionnées à chaque nœud;

B : le nombre d'itérations;

for $b = 1, \ldots, B$ **do**

  Tirer un échantillon bootstrap d^{*b} de d_1^n;

  Construire un arbre maximal $\widehat{\eta}_b$ ou $\widehat{\phi}_b$ sur l'échantillon bootstrap d^{*b} par la variante de CART suivante :;

  for chaque nœud *de* 1 *à* N_b **do**

    Tirer un sous-échantillon de m variables explicatives;

    Partitionner le nœud à partir de la "meilleure" de ces m variables;

  end

end

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

Paramètres de réglage

- Il faut choisir
  - le nombre de prédicteurs m (par défaut $\sqrt{\mathrm{p}}$ en régression) tirés à chaque division de nœud
  - le nombre total d'arbres
  - la taille des sous-échantillons bootstraps si gros jeu de données
- On peut s'appuyer sur l'erreur *out-of-bag*

Ajustement des paramètres / erreur Out Of Bag

On remarque que si m diminue, la variance diminue (la corrélation entre les arbres diminue), mais le biais augmente (les arbres ont une moins bonne qualité d'ajustement).

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

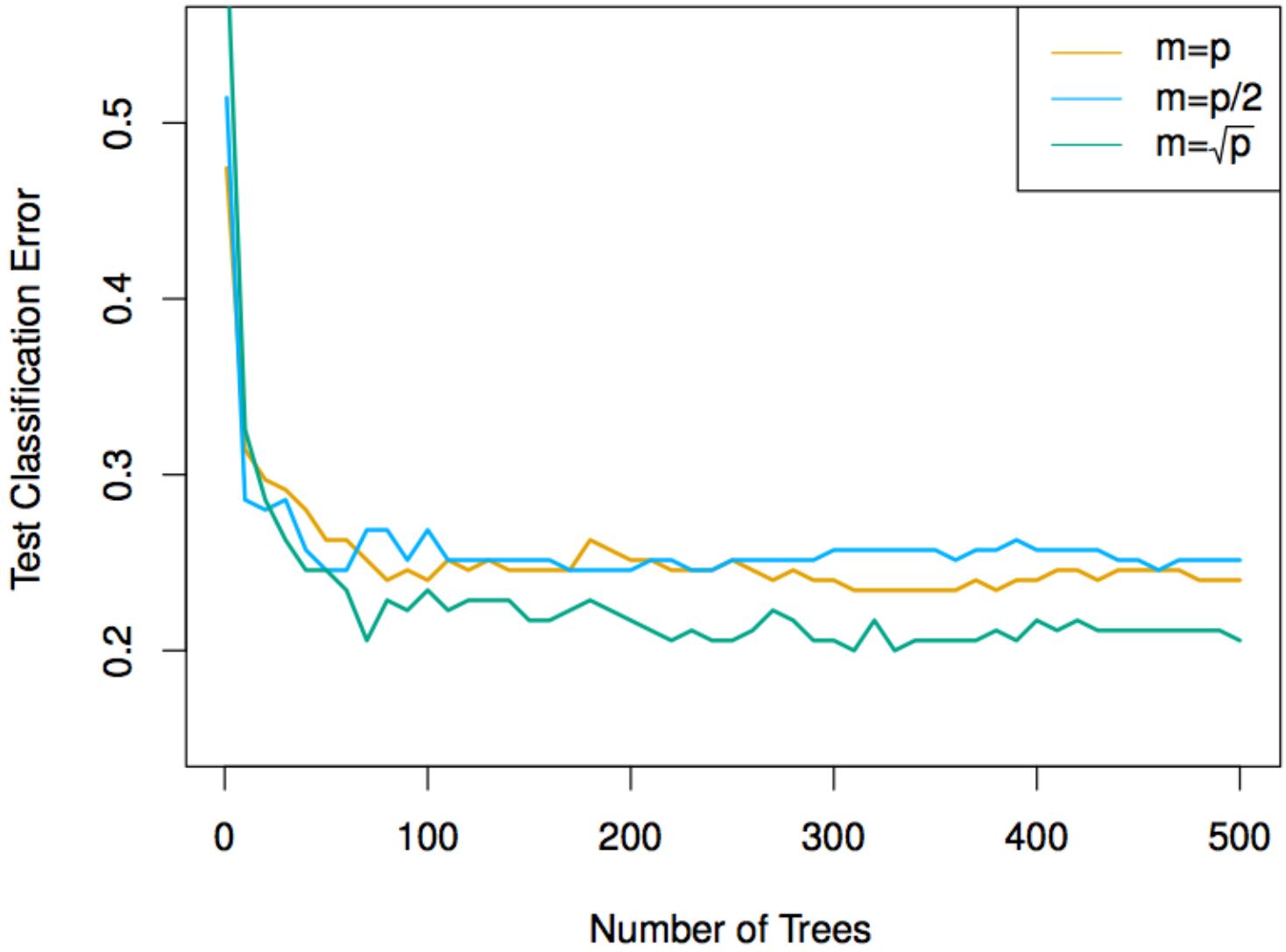
$\hookrightarrow$ Ajustement par validation croisée hold-out ou K fold ou par l'estimation Out Of Bag du risque

L'**erreur Out Of Bag** d'une forêt aléatoire est définie par

- $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{b=1}^{B} I_i^b \widehat{\eta}_b(x_i) \Big/ \sum_{b=1}^{B} I_i^b\right)^2$ en régression,
- $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{signe\left(\sum_{b=1}^{B} I_i^b \widehat{\eta}_b(x_i)\right) \neq y_i}$ en discrimination binaire,

où $I_i^b = 1$ si l'observation $i \notin d^{*b}$, 0 sinon.

Exemple sur données d'expression de 500 gènes



- Résultats de forêts aléatoires pour prédire les 15 classes à partir du niveau d'expression de 500 gènes
- L'erreur de test (évaluée par OOB) dépend du nombre d'arbres. Les différentes couleurs correspondent à différentes valeurs de m.
- Les forêts aléatoires améliorent significativement le taux d'erreur de CART (environ 45.7%)

Avantages / inconvénients

Avantages

- Bonne qualité de prédiction
- Implémentation facile
- Adaptée à la PARALLÉLISATION...

Inconvénients

- Du point de vue métier, on perd l'interprétation facile d'un arbre (effet "boîte noire")

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

Importance des variables

Méthode rudimentaire - non retenue par Breiman et Cutler - consistant à regarder la fréquence des variables explicatives sélectionnées pour découper les arbres de la forêt.

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

- Calculer l'erreur Out Of Bag de l'arbre $\widehat{\eta}_b$ ou $\widehat{\phi}_b$ (sur l'échantillon Out Of Bag correspondant) :
  $OOB_b = \frac{1}{\sum_{i=1}^n I_i^b} \sum_{i=1}^n I_i^b (\widehat{\eta}_b(x_i) - y_i)^2$ ou $\frac{1}{\sum_{i=1}^n I_i^b} \sum_{i=1}^n I_i^b \mathbf{1}_{\widehat{\phi}_b(x_i) \neq y_i}$
- 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 $\widehat{\eta}_b$ ou $\widehat{\phi}_b$ 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 méthodes pour améliorer la qualité de la prédiction des arbres de décision. Ces méthodes agrègent de nombreux arbres entrainés sur les données et ensuite combinent ces arbres pour construire la décision finale.
- Les forêts aléatoires (et le boosting que l'on verra prochainement) font parmi de l'état de l'art actuel des méthodes d'apprentissage supervisé. Cependant, le classifieur ou la fonction de régression produite peut être difficile à interpréter.