Chapitre 1

Introduction

1.1 Préambule

Les deux algorithmes dont il est question dans ce livre ont été proposés par Leo Breiman : les arbres CART (pour Classification And Regression Trees) introduits au milieu des années 1980 (Breiman *et al.*, 1984) et les forêts aléatoires (Breiman, 2001), émergeant un peu moins de vingt plus tard, au début des années 2000. À la confluence de la statistique et de l'apprentissage statistique, ce raccourci des multiples contributions de Leo Breiman, dont la biographie scientifique est décrite dans Olshen & Breiman (2001) et Cutler (2010), constitue un trait remarquable de ces deux disciplines.

Les arbres de décision sont la brique de base de l'édifice des méthodes d'arbres. Bien que connus depuis des décennies et très attractifs grâce à leur simplicité et leur interprétabilité, leur utilisation souffrait, jusque dans les années 1980, de sérieuses objections justifiées. De ce point de vue, CART offre aux arbres de décision un cadre conceptuel de type sélection de modèles automatique, qui leur confère ainsi à la fois une large applicabilité, une facilité d'interprétation et des garanties théoriques.

Mais l'un des défauts majeurs, l'instabilité, reste. L'idée des forêts aléatoires consiste à exploiter la variabilité naturelle des arbres. Plus précisément, il s'agit de perturber la construction en choisissant de façon aléatoire à la fois les individus et les variables. Les arbres ainsi obtenus sont ensuite combinés pour construire la prédiction finale, plutôt que de choisir l'un d'entre eux. Plusieurs algorithmes basés sur des principes semblables ont ainsi été développés, pour beaucoup d'ailleurs, par Breiman lui-même : le Bagging (Breiman, 1996), plusieurs variantes du Arcing (Breiman, 1998) mais aussi Adaboost (Freund & Schapire, 1997).

Les forêts aléatoires (RF dans la suite) sont donc une méthode non-paramétrique d'apprentissage statistique massivement utilisée dans de nombreux domaines d'application, citons par exemple l'étude des biopuces (Díaz-Uriarte & Alvarez

Les forêts aléatoires avec R – Robin Genuer et Jean-Michel Poggi
ISBN 978-2-7535-7710-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

De Andres, 2006), l'écologie (Prasad *et al.*, 2006), la prévision de la pollution (Ghattas, 1999) ou encore la génomique (Goldstein *et al.*, 2010 ; Boulesteix *et al.*, 2012), et pour une revue plus large, voir Verikas *et al.* (2011). Cette universalité est d'abord liée aux excellentes performances en prédiction. On peut le voir dans Fernández-Delgado *et al.* (2014) qui couronne les RF dans le cadre d'une récente évaluation comparative à grande échelle, alors que, moins d'une dizaine d'années auparavant, l'article de Wu *et al.* (2008) aux objectifs semblables, mentionne CART mais pas encore les forêts aléatoires. En outre, elles sont applicables à de nombreux types de données. En effet, il est possible de gérer des données de grande dimension pour lesquelles le nombre de variables dépasse largement le nombre d'observations. De plus, elles sont adaptées aussi bien à des problèmes de classification (variable réponse catégorielle) qu'à des problèmes de régression (variable réponse continue). Elles permettent également de prendre en compte un mélange de variables explicatives qualitatives et quantitatives. Enfin, elles sont bien entendu capables de traiter des données standards pour lesquelles le nombre d'observations est plus grand que le nombre de variables.

Au-delà des performances et du caractère automatique de la méthode avec très peu de paramètres à régler, l'un des aspects les plus importants sur le plan appliqué est la quantification de l'importance des variables. Cette notion, peu examinée par les statisticiens (voir par exemple Grömping, 2015, en régression), trouve dans le cadre des forêts aléatoires une définition commode, facile à évaluer et s'étendant naturellement au cas des groupes de variables (Gregorutti *et al.*, 2015).

Par conséquent, et nous insisterons beaucoup sur cet aspect, on peut utiliser les RF pour effectuer une sélection de variables. Ainsi, en plus d'avoir un outil performant en prédiction, on peut également les utiliser pour sélectionner les variables les plus intéressantes pour expliquer la variable réponse, parmi un nombre de variables potentiellement très grand. Ceci est très attrayant en pratique car cela permet à la fois d'aider à interpréter plus aisément les résultats mais aussi et surtout à déterminer des facteurs influents pour le problème étudié. Enfin, cela peut également être bénéfique pour la prédiction, car éliminer de nombreuses variables « de bruit » permet de faciliter la tâche d'apprentissage.

1.2 Notations

Dans tout le livre nous adopterons les notations suivantes. On suppose que l'on dispose d'un échantillon d'apprentissage :

$$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

constitué de n couples d'observations indépendantes et identiquement distribuées, de même loi qu'un couple noté (X, Y). Cette loi est bien entendu inconnue en pratique et le but est justement de l'estimer, ou plus particulièrement d'estimer le lien qui existe entre X et Y.

Les forêts aléatoires avec R – Robin Genuer et Jean-Michel Poggi
ISBN 978-2-7535-7710-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

Nous appelons les coordonnées de X les « variables d'entrées » (ou « variables explicatives » ou encore « variables »), que nous notons X^j pour la j-ième coordonnée, et nous supposons que $X \in \mathcal{X}$, un certain espace que nous préciserons plus tard. Cependant, nous supposons que cet espace est de dimension p, où p est le nombre (total) de variables.
Y désigne la « variable réponse » (ou « variable à expliquer » ou « variable dépendante ») et $Y \in \mathcal{Y}$. La nature du problème de régression ou de classification dépend de celle de l'espace $\mathcal{Y}$:

— si $\mathcal{Y} = \mathbb{R}$, nous avons un problème de régression ;
— si $\mathcal{Y} = \{1, \ldots, C\}$, nous avons un problème de classification à C classes.

1.3 Objectifs statistiques

Prédiction

Le premier objectif en apprentissage est la prédiction. On cherche, à l'aide de l'échantillon d'apprentissage $\mathcal{L}_n$, à construire un prédicteur :

$$\widehat{h} : \mathcal{X} \to \mathcal{Y}$$

qui à toute observation d'entrée $x \in \mathcal{X}$ est capable d'associer une prédiction $\widehat{y}$ de la variable réponse correspondant à cette observation.
Le « chapeau » de $\widehat{h}$ est une notation pour préciser le fait que ce prédicteur est construit en utilisant $\mathcal{L}_n$ (nous n'ajoutons pas la dépendance en n pour le prédicteur pour alléger les notations, mais celle-ci existe bien).
Plus précisément, nous souhaitons construire un prédicteur performant au sens de l'erreur de prédiction (appelée aussi erreur de généralisation) :

— en régression, il s'agit de l'erreur quadratique en espérance : $\mathrm{E}\left[(Y - \widehat{h}(X))^2\right]$;
— en classification, de la probabilité de mauvais classement : $\mathrm{P}\left(Y \neq \widehat{h}(X)\right)$.

L'erreur de prédiction dépendant de la loi inconnue de (X, Y), il faut l'estimer. Une façon classique de le faire est, à l'aide d'un échantillon test $\mathcal{T}_m = \{(X'_1, Y'_1), \ldots, (X'_m, Y'_m)\}$, issu lui aussi, de la loi de (X, Y), de calculer une erreur test :

— en régression, il s'agit de l'erreur quadratique moyenne : $\frac{1}{m}\sum_{i=1}^{m}\left(Y'_i - \widehat{h}(X'_i)\right)^2$;
— en classification, du taux de mauvais classement : $\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{Y'_i \neq \widehat{h}(X'_i)}$.

Dans le cas où l'on ne dispose pas d'échantillon test, on peut quand même estimer l'erreur de prédiction par exemple par validation croisée. Nous verrons dans la suite une façon de le faire propre aux forêts aléatoires.

Les forêts aléatoires avec R – Robin Genuer et Jean-Michel Poggi
ISBN 978-2-7535-7710-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

Remarque
Dans ce livre, nous nous focalisons sur les cadres de la régression et de la classification supervisée. Cependant les RF ont été généralisées à d'autres cadres statistiques.

Tout d'abord, pour l'analyse de données de survie, Ishwaran *et al.* (2008) ont introduit les Random Survival Forests, en transposant les idées principales des RF au cas où la quantité à prédire est un temps d'événement. Citons également sur ce sujet les travaux de Hothorn *et al.* (2006).

Les forêts aléatoires ont également été généralisées au cas où la variable réponse est multivariée (voir la revue de la littérature de Segal & Xiao, 2011, qui donne également des références dès les années 1990).

Sélection et importance des variables

Un second objectif classique est celui de la sélection de variables. Il s'agit de déterminer un sous-ensemble des variables d'entrée effectivement utiles pour expliquer le lien entrée-sortie.

On évalue souvent la qualité d'un sous-ensemble de variables sélectionnées par la performance obtenue avec un prédicteur utilisant uniquement ces variables.

En complément, on peut s'intéresser à construire une hiérarchie des variables d'entrée fondée sur une quantification de l'importance des effets sur la variable de sortie. Un tel indice d'importance fournit donc un classement des variables, de la plus importante à la moins importante.

1.4 Packages

Nous nous concentrerons principalement sur l'utilisation de trois packages R (R Core Team, 2018) :

- **rpart** (Therneau & Atkinson, 2018) pour les méthodes d'arbres, dans le chapitre 2 ;
- **randomForest** (Liaw & Wiener, 2018) pour les forêts aléatoires, dans les chapitres 3 et 4 ;
- **VSURF** (Genuer *et al.*, 2018) pour la sélection de variables à l'aide de forêts aléatoires, dans le chapitre 5.

Remarque
Concernant les variantes des forêts aléatoires évoquées dans la section précédente, le package `randomForestSRC` (Ishwaran & Kogalur, 2017) fournit une implémentation unifiée à la fois des forêts en régression, classification supervisée, dans un cadre de survie, et également pour le cas où la réponse est multivariée.

Les forêts aléatoires avec R – Robin Genuer et Jean-Michel Poggi
ISBN 978-2-7535-7710-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

1.5 Jeux de données

1.5.1 Exemple fil rouge : détection de spams

Nous illustrerons l'application des différentes méthodes sur les très classiques données `spam` à des fins pédagogiques, en guise de fil rouge.
Ce jeu de données, bien connu et largement disponible, est dû à un ingénieur de HP (Hewlett-Packard), prénommé George, qui a analysé un échantillon de ses emails professionnels :

- les observations sont les 4 601 emails en question dont 2 788 sont des emails souhaitables et 1 813 (soit 40 %) des emails indésirables, c'est-à-dire des spams ;
- la variable réponse est donc binaire : `spam` ou `non-spam`. Nous renommerons la modalité `non-spam` en `ok` pour faciliter la lecture de certains graphiques ;
- les variables explicatives sont au nombre de $p = 57$: 54 sont des proportions d'occurrences de mots ou de caractères, comme par exemple `$` (noté `charDollar`), `!` (noté `charExclamation`), `free`, `money`, deux sont liées aux longueurs des suites de lettres majuscules (la moyenne, la plus longue) et enfin la dernière est le nombre de lettres majuscules dans le mail. Ces variables sont classiques et définies grâce à des procédures usuelles en analyse textuelle, permettant de traiter statistiquement des observations caractérisées par des textes.

Les objectifs statistiques énoncés précédemment se formulent pour cet exemple de la façon suivante : premièrement, on souhaite construire un « bon » filtre anti-spam : un nouvel email arrive, il faut réussir à prédire si c'est un spam ou non. Deuxièmement, on s'intéresse aussi à savoir quelles sont les variables sur lesquelles se base le plus le filtre anti-spam (ici des mots ou des caractères).
Pour juger de la performance d'un filtre anti-spam, on découpe le jeu de données en deux : 2 300 mails pour l'apprentissage, 2 301 mails pour tester les prédictions. Nous avons donc un problème de **classification à 2 classes** ($C = 2$) avec un nombre d'individus ($n = 2\,300$ pour l'apprentissage, la construction des modèles) largement plus important que le nombre de variables ($p = 57$). De plus, nous disposons d'un échantillon test de grande taille ($m = 2\,301$) pour évaluer une estimation de l'erreur de prédiction.
Chargeons le jeu de données dans R, disponible dans le package `kernlab` (Karatzoglou *et al.*, 2004), renommons la modalité `nonspam` en `ok` pour faciliter la lecture des graphiques et fixons les tableaux de données d'apprentissage et de test :

```
> data("spam", package = "kernlab")
> set.seed(9146301)
> levels(spam$type) <- c("ok", "spam")
> yTable <- table(spam$type)
> indApp <- c(sample(1:yTable[2], yTable[2]/2),
    sample((yTable[2] + 1):nrow(spam), yTable[1]/2))
```


Les forêts aléatoires avec R – Robin Genuer et Jean-Michel Poggi
ISBN 978-2-7535-7710-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

```
> spamApp <- spam[indApp, ]
> spamTest <- spam[-indApp, ]
```

Remarque
La commande `set.seed(9146301)` permet de fixer la graine du générateur de nombres aléatoires dans R. Ainsi, si le bloc d'instructions précédent est exécuté plusieurs fois, il n'y aura pas de variabilité des échantillons d'apprentissage et de test.

1.5.2 Pollution par l'ozone

Les données `Ozone` sont utilisées dans de très nombreux articles et constituent l'un des jeux d'essai classiques depuis l'article de Breiman & Friedman (1985).

L'objectif est ici de prédire la concentration maximale d'ozone associée à un jour de l'année 1976 dans la région de Los Angeles, en utilisant 12 variables météorologiques et calendaires. Les données consistent en 366 observations et 13 variables, chaque observation est associée à un jour. Les 13 variables sont :

- `V1` Mois : `1` = janvier, ..., `12` = décembre ;
- `V2` Jour du mois : de `1` à `31` ;
- `V3` Jour de la semaine : `1` = lundi, ..., `7` = dimanche ;
- `V4` Maximum journalier des moyennes horaires des concentrations d'ozone du jour ;
- `V5` Hauteur de pression de 500 millibars (m) mesurée à Vandenberg AFB ;
- `V6` Vitesse du vent (mph) à l'aéroport international de Los Angeles (LAX) ;
- `V7` Humidité ($\%$) à LAX ;
- `V8` Température ($degrés\ F$) mesurée à Sandburg, Californie ;
- `V9` Température ($degrés\ F$) mesurée à El Monte, Californie ;
- `V10` Hauteur d'inversion ($pieds$) à LAX ;
- `V11` Gradient de pression ($mmHg$) de LAX à Daggett, Californie ;
- `V12` Température au niveau de la hauteur d'inversion ($degrés\ F$) à LAX ;
- `V13` Visibilité ($miles$) mesurée à LAX.

Il s'agit donc d'un problème de **régression** où l'on doit prédire `V4` (la concentration maximale d'ozone associée à un jour) en utilisant les 12 autres variables, neuf météorologiques (`V5` à `V13`) et trois calendaires (`V1` à `V3`).

Bien souvent, seules les variables explicatives continues sont considérées. Ici, les méthodes d'arbres permettent de toutes les prendre en compte, même si inclure `V2` le jour du mois est a priori dépourvu d'intérêt.

Ce jeu de données est disponible dans le package `mlbench` (Leisch & Dimitriadou, 2010) et peut se charger dans R à l'aide de la commande suivante :

```
> data("Ozone", package = "mlbench")
```


Les forêts aléatoires avec R – Robin Genuer et Jean-Michel Poggi
ISBN 978-2-7535-7710-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

1.5.3 Données génomiques pour une étude vaccinale

Le jeu de données `vac18` est issu d'un essai vaccinal prophylactique pour le VIH (Thiébaut *et al.*, 2012). Les expressions d'un sous-ensemble de 1 000 gènes ont été mesurées pour 42 observations, issues de 4 stimulations différentes :

- le candidat vaccin (LIPO5) ;
- un vaccin contenant le peptide Gag (GAG+) ;
- un vaccin ne contenant pas le peptide Gag (GAG-) ;
- et une non-stimulation (NS) ;

pour 12 participants VIH négatifs.
L'objectif de prédiction revient ici à déterminer, au vu de l'expression des gènes, la stimulation qui a été utilisée. Il s'agit donc d'un problème de **classification à 4 classes en grande dimension**. Mentionnons que ce problème de prédiction est ici un problème intermédiaire pour atteindre le véritable objectif : la sélection des gènes les plus impliqués dans la discrimination entre les différents vaccins.
Nous chargeons les données `vac18`, disponibles dans le package `mixOmics` (Le Cao *et al.*, 2017) :

```
> data("vac18", package = "mixOmics")
```

1.5.4 Pollution par les poussières

Ces données sont publiées dans Jollois *et al.* (2009).
Les particules en suspension dans l'air sont d'origine diverse, naturelle ou liée à l'activité humaine, et la composition chimique de ces particules peut beaucoup varier. En 2009, Air Normand, alors observatoire de la qualité de l'air en Haute-Normandie, disposait d'une dizaine d'appareils mesurant les teneurs globales en particules PM10 dont le diamètre est inférieur à 10 μm, et exprimées en $\mu g/m^3$ moyen dans le quart d'heure écoulé. La réglementation fixe la valeur de 50 $\mu g/m^3$ (en moyenne journalière) comme limite à ne pas dépasser plus de 35 jours dans l'année pour les PM10.
On se restreint à un sous-réseau de six stations de mesure des émissions de PM10 : trois à Rouen GCM (industrielle), JUS (urbaine) et GUI (à proximité du trafic), deux au Havre REP (trafic) et HRI (urbaine) et enfin une station rurale à Dieppe AIL.
Les données considérées pour les six stations sont :

- pour la météo : la pluie PL, la vitesse du vent VV (max et moy), la direction du vent DV (max et dominant), la température T (min, max et moy), le gradient de température GT (au Havre et à Rouen), la pression atmosphérique PA et l'humidité relative HR (min, max et moy) ;
- pour les polluants : les poussières (PM10), les oxydes d'azote (NO, NO2) pour la pollution urbaine et le dioxyde de soufre (SO2) pour la pollution industrielle : en plus de ceux mesurés en chaque station on adjoint des polluants mesurés à proximité :

Les forêts aléatoires avec R – Robin Genuer et Jean-Michel Poggi
ISBN 978-2-7535-7710-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

— pour GUI : ajout du SO2 mesuré à JUS ;
— pour REP : ajout du SO2 mesuré au Havre (station MAS) ;
— pour HRI : ajout du NO et du NO2 mesurés au Havre (station MAS).

Chargeons les données de la station JUS, incluses dans le package **VSURF** :

```
> data("jus", package = "VSURF")
```

Les forêts aléatoires avec R – Robin Genuer et Jean-Michel Poggi
ISBN 978-2-7535-7710-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr