

Arbres de décisions et forêts aléatoires.

Pierre Gaillard

7 janvier 2013

Plan

- 1 Arbre de décision
- 2 Les forêts aléatoires

Introduction



Introduction

Jeu de données (ex : n emails)

- n observations
- p variables décrivant ces observations

Variable cible à prévoir, Y .

ex : spam / non spam

Variables explicatives, $X_1, \dots, X_p$.

ex : proportions de majuscules, présence de certains mots (remove, free, ...), de certains caractère (\$, !, ...)

Idées

Arbre de décisions. Découper l'espace des variables explicatives en groupes homogènes.

Forêts aléatoires. Construire aléatoirement de nombreux arbres et moyenne.

I. Les arbres de décision



Historique et références

Méthode introduite par **Leo Breiman** en 1984.

Il existe de nombreuses variantes : **CART** (Classification and regression trees), C4.5, CHAID, ID3

Package R : tree, rpart.

Références :

Breiman L. and al (1984) Classification and Regression Trees.

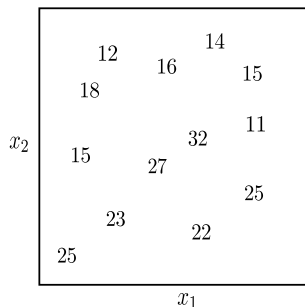


Arbre de régression

Ensemble d'entraînement $(X_t, Y_t)_{t \in S_0}$ indépendantes et identiquement distribuées.

Objectif

Expliquer Y_t en fonction des variables contextuelles X_t .

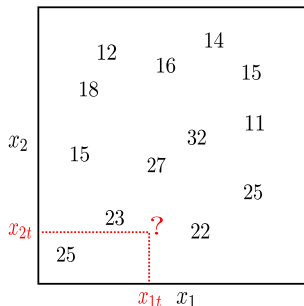


Arbre de régression

Ensemble d'entraînement $(X_t, Y_t)_{t \in S_0}$ indépendantes et identiquement distribuées.

Objectif

Expliquer Y_t en fonction des variables contextuelles X_t .

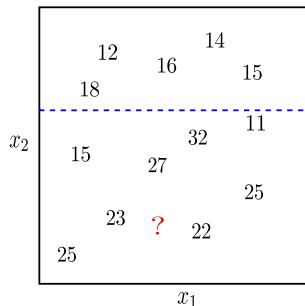


Arbre de régression

Ensemble d'entraînement $(X_t, Y_t)_{t \in S_0}$ indépendantes et identiquement distribuées.

Objectif

Expliquer Y_t en fonction des variables contextuelles X_t .

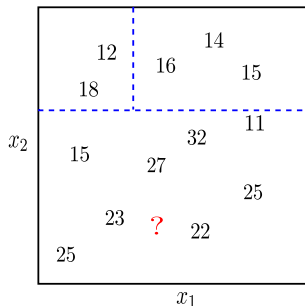


Arbre de régression

Ensemble d'entraînement $(X_t, Y_t)_{t \in S_0}$ indépendantes et identiquement distribuées.

Objectif

Expliquer Y_t en fonction des variables contextuelles X_t .

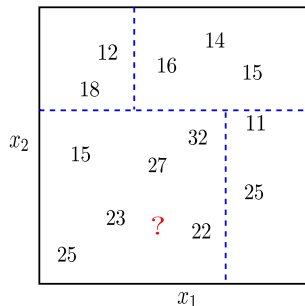


Arbre de régression

Ensemble d'entraînement $(X_t, Y_t)_{t \in S_0}$ indépendantes et identiquement distribuées.

Objectif

Expliquer Y_t en fonction des variables contextuelles X_t .

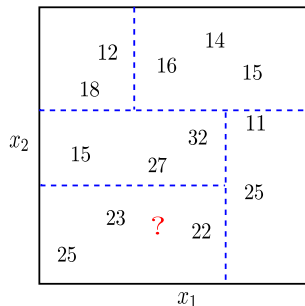


Arbre de régression

Ensemble d'entraînement $(X_t, Y_t)_{t \in S_0}$ indépendantes et identiquement distribuées.

Objectif

Expliquer Y_t en fonction des variables contextuelles X_t .

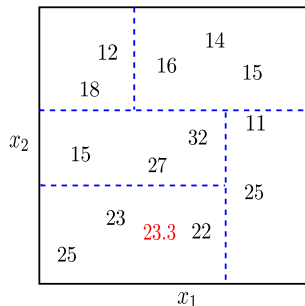


Arbre de régression

Ensemble d'entraînement $(X_t, Y_t)_{t \in S_0}$ indépendantes et identiquement distribuées.

Objectif

Expliquer Y_t en fonction des variables contextuelles X_t .



Exemple numérique : prévision de la qualité de l'air

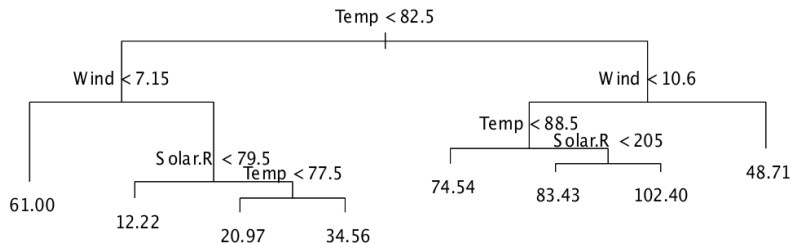
Relevé quotidien de la qualité de l'air (concentration d'ozone) à New York de mai à septembre 1973.

Jeu de données. 154 observations, 6 variables explicatives.

Variables explicatives. Vitesse moyenne du vent, température maximale, jour de la semaine, mois de l'année, ensoleillement.

Références. Sous R. Jeu de données `airquality`.

Exemple : prévision de la qualité de l'air



Construction de l'arbre

Nous devons répondre aux trois questions suivantes :

- Comment choisir la variable à segmenter parmi les variables explicatives disponibles ?
- Lorsque la variable est continue, comment déterminer le seuil de coupe ?
- Comment déterminer la bonne taille de l'arbre ?

Choisir la variable et trouver la bonne coupe

Le choix de la variable explicative j où couper, et de la coupe correspondante τ partitionne les données en deux ensembles :

$$E_{j,\tau}^- = \{i \in \{1, \dots, n\} \text{ tq } X_{i,j} \leq \tau\}$$

$$E_{j,\tau}^+ = \{i \in \{1, \dots, n\} \text{ tq } X_{i,j} > \tau\}$$

Exemple : prévision de la qualité de l'air

On peut décider de couper selon que la **température** est supérieure ou inférieure à **30°C**.

Choisir la variable et trouver la bonne coupe

Le choix de la variable explicative j où couper, et de la coupe correspondante τ partitionne les données en deux ensembles :

$$E_{j,\tau}^- = \{i \in \{1, \dots, n\} \text{ tq } X_{i,j} \leq \tau\}$$

$$E_{j,\tau}^+ = \{i \in \{1, \dots, n\} \text{ tq } X_{i,j} > \tau\}$$

Idée. Obtenir deux groupes les plus homogènes possible pour la variable à prévoir Y . On minimise leur variance :

$$(j^*, \tau^*) = \operatorname{argmin}_{j,\tau} \left\{ \operatorname{Var}(E_{j,\tau}^-) + \operatorname{Var}(E_{j,\tau}^+) \right\}$$

où $\operatorname{Var}(E) = \min_{y \in \mathbb{R}} \sum_{i \in E} (Y_i - y)^2$

Choisir la variable et trouver la bonne coupe

Le choix de la variable explicative j où couper, et de la coupe correspondante τ partitionne les données en deux ensembles :

$$E_{j,\tau}^- = \{i \in \{1, \dots, n\} \text{ tq } X_{i,j} \leq \tau\}$$

$$E_{j,\tau}^+ = \{i \in \{1, \dots, n\} \text{ tq } X_{i,j} > \tau\}$$

Exemple : prévision de la qualité de l'air

On décide de couper selon que la **température** est supérieure ou inférieure à **30°C**, si les données qui ont une température inférieure à 30°C ont une qualité de l'air similaire (de même pour celle supérieures).

Trouver la bonne coupe pour une variable catégorique

Pour des variables catégoriques non ordonnée, on ne peut pas trouver de seuil τ , pour diviser les données facilement en deux ensembles.

Exemple : prévision de la qualité de l'air

On peut avoir une variable renseignant si le temps est **ensoleillé**, **nuageux**, **pluvieux**, ou **neigeux**.

Si la variable prend ses valeurs dans un ensemble de k catégories, il y a $2^{k-1} - 1$ partitions correspondantes. C'est trop !

Solution. Ordonner les catégories suivant la moyenne des Y_i leur appartenant.

À quel point doit on laisser l'arbre grandir ?

Un arbre trop grand risque de **surprendre** les données.

Un trop petit risque de **ne pas capter leur structure**.

La taille optimal est un paramètre à **calibrer** sur les données.

Solutions :

- Descendre tant que la variance au sein des feuilles décroît plus qu'un certain seuil.
Vision à trop court terme : une mauvaise coupe peut parfois mener à de très bonnes coupes.
- **Élagage**. On fait grandir l'arbre au maximum puis on supprime les noeuds inintéressants.

Avantage-Inconvénients

Avantages

Facilement interprétable. Rapide.

Inconvénients

Instabilité !! L'élagage n'est pas toujours facile.

Exemple numérique : prévision de la qualité de l'air

Relevé quotidien de la qualité de l'air (concentration d'ozone) à New York de mai à septembre 1973.

Jeu de données. 154 observations, 6 variables explicatives.

Variables explicatives. Vitesse moyenne du vent, température maximale, jour de la semaine, mois de l'année, ensoleillement.

Références. Sous R. Jeu de données `airquality`.

Exemple numérique. Code R

Prévision de la qualité de l'air

Chargement de la library

```
library("tree")
```

Chargement des données

```
data(airquality)
```

Construction de l'arbre

```
ozone.tr <- tree(Ozone ~., data=airquality)
```

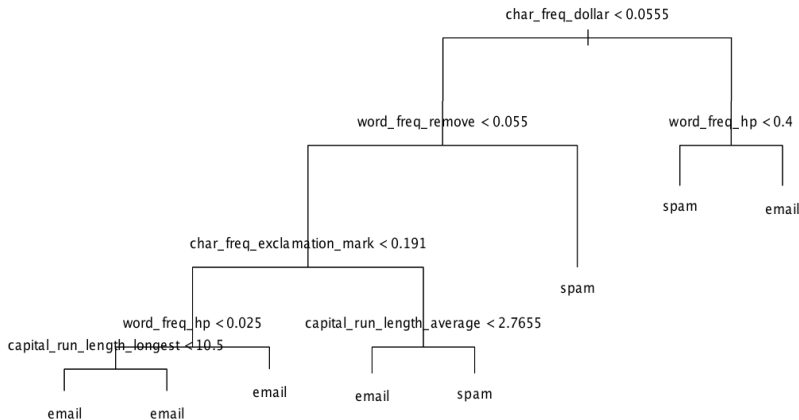
La classification

La variable à expliquer prend ses valeurs dans un ensemble de **catégories**.

Exemples.

Détection de spam. Reconnaissance d'objet, de races d'animaux, de variétés de plantes à partir de mesures,...

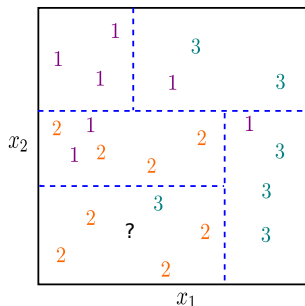
Arbre de classification. Détection de spams



Arbre de classification

Différences avec l'arbre de régression :

- au lieu de prévoir comme la moyenne des éléments de la feuille, on prévoit comme la majorité.
- le critère de variance pour sélectionner la variable et la coupe est légèrement différents.



Exemple numérique. Détection de spams

Jeu de données contenant 2788 emails et 1813 spams en anglais.

Variables contextuelles disponibles :

- `word_freq_...` : 48 réels donnant le pourcentage de certains mots dans l'email (ex : `word_freq_free`)
- `char_freq_...` : 6 réels donnant le pourcentage de certains caractères (ex : `$, !, ...`)
- `capital_run_length_average` : durée moyenne de suite de majuscules consécutives.
- `capital_run_length_longest` : durée de la plus grande suites de majuscules consécutives.
- `capital_run_length_total` : nombre total de majuscules.

II. Les forêts aléatoires



Historique et références

Méthode introduite par **Leo Breiman** en 2001.

Idées plus anciennes : **Bagging** (1996), arbres de décisions **CART** (1984)

Preuves de convergences récentes (2006,2008)

Un site web utile :

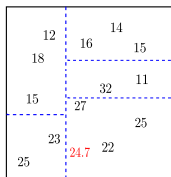
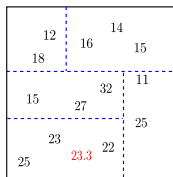
<http://www.stat.berkeley.edu/~breiman/RandomForests>



Forêts aléatoires

Les forêts aléatoires consistent à faire tourner en parallèle un **grand nombre** (≈ 400) d'arbres de décisions **construits aléatoirement**, avant de les **moyenner**.

En termes statistiques, si les arbres sont décorrélés, cela permet de réduire la variance des prévisions.

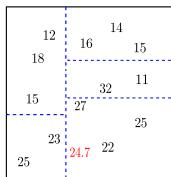
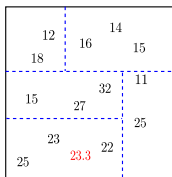


$$\text{prévoit } \frac{24.7 + 23.3}{2} = 24$$

Forêts aléatoires

Intuition : si K arbres sont identiquement distribués, de variance σ^2 , avec un coefficient de corrélation deux à deux ρ , la variance de leur moyenne est alors

$$\bar{\sigma}^2 = \frac{1}{K}(K\sigma^2 + K(K-1)\rho\sigma^2) = \rho\sigma^2 + \frac{1-\rho}{K}\sigma^2$$



prévoit $\frac{24.7+23.3}{2} = 24$

Créer des arbres peu corrélés

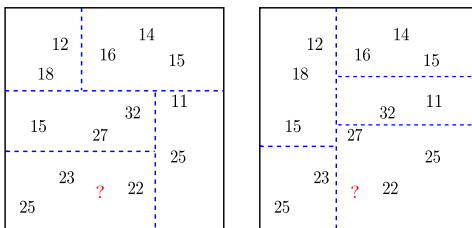
Bootstrapping. Plutôt qu'utiliser toutes les données pour construire les arbres. On choisit pour chaque arbre aléatoirement un sous-ensemble des données (ex : 1 donnée sur 5).

Choix aléatoire de la variable explicative à couper.

À la différence des arbres Cart, on ne fait pas d'élagage.

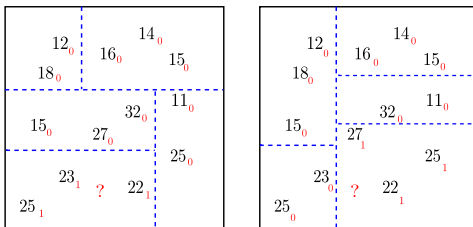
Une notion importante : la proximité

Intuition : Tomber souvent dans les mêmes feuilles des arbres $\rightarrow$ expliquer la sortie Y de façon similaire.



Une notion importante : la proximité

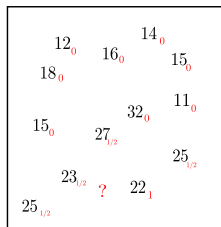
Intuition : Tomber souvent dans les même feuilles des arbres $\rightarrow$ expliquer la sortie Y de façon similaire.



$$\text{prox}(X_t, X_s) = \frac{1}{K} \sum_{k=1}^K \mathbb{1} \left\{ \begin{array}{l} X_t \text{ et } X_s \text{ tombent dans la} \\ \text{même feuille dans l'arbre } k \end{array} \right\} .$$

La notion de proximité

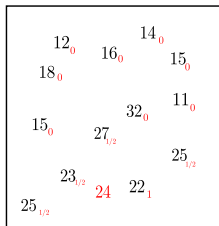
Intuition : Tomber souvent dans les même feuilles des arbres $\rightarrow$ expliquer la sortie Y de façon similaire.



$$\text{prox}(X_t, X_s) = \frac{1}{K} \sum_{k=1}^K \mathbb{1} \left\{ \begin{array}{l} X_t \text{ et } X_s \text{ tombent dans la} \\ \text{même feuille dans l'arbre } k \end{array} \right\} .$$

La notion de proximité

Intuition : Tomber souvent dans les même feuilles des arbres $\rightarrow$ expliquer la sortie Y de façon similaire.



On prédit ensuite par exemple par

$$\arg \min_{a \in \mathbb{R}} \sum_{X_s \in E_t} \text{prox}(X_t, X_s) (Y_s - a)^2.$$

Exemple numérique. Code R

Prévision de la qualité de l'air

Chargement de la library

```
library("randomForest")
```

Chargement des données

```
data(airquality)
```

Construction de l'arbre

```
ozone.rf <- randomForest(Ozone ~., data=airquality, mtry=3,  
                          importance=TRUE, na.action=na.omit)
```