Forêts aléatoires

Véronique Tremblay

Objectifs

- Comprendre le fonctionnement des forêts aléatoire
- Comprendre pourquoi les forêts offrent de bonnes performances
- Connaître les avantages et inconvénients des forêts
- Comprendre le calcul de l'importance des variables
- Savoir qu'il est possible de calculer des intervalles de prévision

L'algorithme est le suivant:

1. Pour $b = 1, 2, ...B$
   a. Sélectionner un échantillon bootstrap Z^b
   b. Construire un arbre en utilisant la procédure suivante:
      i. À chaque noeud, sélectionner $m < M$ variables
      ii. Parmi les m variables sélectionnées, choisir la meilleur variable et le meilleur endroit où séparer selon le critère choisi (Gini, χ^2)
      iii. Séparer le noeud en deux

      Poursuivre les étapes i à iii jusqu'à l'atteinte d'un certain critère d'arrêt.
2. Aggréger les B arbres (par une moyenne ou un vote)

Pourquoi ça fonctionne

On veut toujours minimiser

$$E[(Y - \hat{f}(x))^2] = \sigma_\epsilon^2 + \mathrm{Var}[\hat{f}(x)] + \mathrm{Biais}[\hat{f}(x)]^2$$

Rappel du bagging

$$\mathrm{Var}[\hat{f}_{bag}(x)] = \rho\sigma_b^2 + \frac{1-\rho}{B}\sigma_b^2$$

Avantage

- Nécessite peut de traitement préliminaire des données
- Ne suppose pas de distribution
- Permet une certaine gestion des valeurs manquantes
- Robuste aux valeurs extrêmes

Avantage

- Nécessite peut de traitement préliminaire des données
- Ne suppose pas de distribution
- Permet une certaine gestion des valeurs manquantes
- Robuste aux valeurs extrêmes
- Stable

Avantage

- Nécessite peut de traitement préliminaire des données
- Ne suppose pas de distribution
- Permet une certaine gestion des valeurs manquantes
- Robuste aux valeurs extrêmes

- Stable
- Excellent pouvoir prédictif

Avantage

- Nécessite peut de traitement préliminaire des données
- Ne suppose pas de distribution
- Permet une certaine gestion des valeurs manquantes
- Robuste aux valeurs extrêmes

- Stable
- Excellent pouvoir prédictif
- Réduit le temps de calcul (par rapport au bagging)

Avantage

- Nécessite peut de traitement préliminaire des données
- Ne suppose pas de distribution
- Permet une certaine gestion des valeurs manquantes
- Robuste aux valeurs extrêmes

- Stable
- Excellent pouvoir prédictif
- Réduit le temps de calcul (par rapport au bagging)
- Échantillon *OOB*

Avantage

- Nécessite peut de traitement préliminaire des données
- Ne suppose pas de distribution
- Permet une certaine gestion des valeurs manquantes
- Robuste aux valeurs extrêmes

- Stable
- Excellent pouvoir prédictif
- Réduit le temps de calcul (par rapport au bagging)
- Échantillon *OOB*
- Intervalles de prédiction

Échantillon *OOB* (forêt et *bagging*)

À chaque étape ($b = 1$ à B), on dispose d'un échantillon qui n'a pas été utilisé: l'échantillon *OOB*.

Pour chaque observation, on peut faire une prévision uniquement à partir des arbres qui ont été construits sans cette observation.

$$\hat{f}_{OOB}(x_i) = \frac{1}{B_i} \sum_{i=1}^{B} \hat{f}_b(x_i) \mathbf{1}(x_i \in OOB_b)$$

Avec $\mathbf{1}$ qui est la fonction indicatrice et B_i le nombre d'arbres qui ont été construits sans l'observation i.

Intervalle de prédiction

L'idée générale est d'utiliser les «forêts de KNN» pour approximer la distribution de Y localement. Si l'échantillon et le nombre d'effectifs dans chaque noeud terminal est suffisamment grand, on peut en déduire les quantiles.

Inconvénient

- Un hyperparamètre de plus à choisir!
- Difficile à interpréter, mais...

Importance des variables

Idée: si une variable X est importante pour prédire Y, l'erreur de prédiction augmentera beaucoup si on brise le lien entre X et Y.

Pour calculer l'importance de X_1, un approche serait de:

1. Construire un modèle pour prédire $\hat{f}(x)$
2. Calculer l'erreur de ce modèle (échantillon de validation ou *OOB*)
3. Briser le lien entre X_1 et Y en effectuant une permutation aléatoire des valeurs de X_1.
4. Prédire Y avec X_1 permuté
5. Calculer l'erreur (échantillon de validation ou OOB) et la comparer à celle obtenue en (2).

Importance des variables et sélection des données

On peut utiliser l'importance des variables pour sélectionner les variables.

Extensions

- Forêts de survie
- Forêts pour données corrélées

En pratique

```r
library(ranger)
model_ranger <- ranger(Survived~.,
                data = don, # Jeu de données
                mtry = 3, # Nombre de variables évaluées à
                          #chaque noeud
                num.trees = 500, # Nombre d'arbres
                sample.fraction = 1, # Fraction de l'échantillon
                                     # pour chaque arbre
                splitrule = 'gini', # Critère de séparation
                min.node.size = 3, # Nombre d'observation minimal
                                   # dans un noeud terminal
                importance = 'permutation', # Pour obtenir l'importance
                                            # des variables
                scale.permutation.importance = TRUE,
                keep.inbag = TRUE,
                quantreg = TRUE, # Obtenir les quantiles pour
                                 # un intervalle de prévision
                oob.error = TRUE, # Conserver les OOB
                save.memory = FALSE # Si l'échantillon est très gros
                )
```