



Introduction à l'Apprentissage Statistique

Bagging et Random Forest

M2 Actuariat – ISFA – 2021/2022

Pierrick Piette
Actuaire à Seyna
pierrick.piette@gmail.com

Seyna.

In God we trust, all others bring data
- William Edwards Deming





Ensemble learning

Problématique majeure d'un CART ?

La construction d'un arbre optimal peut varier fortement quand bien même le jeu de données initial varie peu...

Certains techniques ont été développées afin de stabiliser la prévision donnée par un arbre

- Estimer plusieurs arbres sur des sous-parties : *weak learner*
- Pour créer un estimateur sur cet ensemble : *strong learner*

Pour éviter de corrélérer les estimateurs simples qui composeront l'estimateur agrégé, deux stratégies possibles

- Tirage aléatoire de sous-jeux de données
- Choix aléatoire des covariables considérées

Mieux qu'un arbre : une forêt !



Stratégies d'agrégation

Deux stratégies s'opposent dans les méthodes d'Ensemble Learning

Stratégie d'agrégation aléatoire : **Bagging** (Bootstrap AGGRegatING)

- créer des échantillons, estimer le modèle puis combiner les modèles
- Réduction de la variance

Stratégie d'apprentissage incrémental : **Boosting**

- Apprentissage sur un paquet, puis prévision sur un deuxième paquet
- Apprendre des erreurs de prédiction, actualiser le modèle, etc.
- Réduction du biais

L'ensemble learning : une idée ancienne !

XX. — Le même hiver, les Platéens toujours assiégés par les Péloponnésiens et les Béotiens, souffrant de la disette et n'espérant plus aucun secours d'Athènes ni d'ailleurs, firent de concert avec les Athéniens enfermés avec eux dans la ville le projet suivant : Ils sortiraient tous ensemble, en franchissant de force s'ils le pouvaient les murailles de l'ennemi. C'étaient le devin Théænnétos fils de Tolmidas et Eumolpidas fils de Daïmakhos un de leurs stratèges, qui avaient conçu ce dessein. Par la suite la moitié de la garnison, effrayée des difficultés de l'entreprise, y renonça. Deux cent vingt volontaires acceptèrent les risques de la sortie. Voici comment ils s'y prirent. Ils fabriquèrent des échelles ayant la hauteur de la muraille ennemie. Ils calculèrent cette hauteur en dénombrant les rangées de briques sur la partie de la muraille qui leur faisait face et qu'on n'avait pas recouverte de crépi. Plusieurs hommes à la fois compptaient les rangées et, en admettant que quelques-uns se trompassent, la plupart devaient trouver le nombre exact; d'ailleurs ce calcul fut répété fréquemment; la distance étant peu considérable, l'on pouvait facilement apercevoir la partie du mur à examiner. C'est ainsi qu'ils déterminèrent la hauteur des échelles en la calculant d'après l'épaisseur des briques.

Siège de Platée (hiver 428 av. J-C)

Histoire de la guerre du Péloponnèse, Livre 3.

Thucydide, 411 av. J-C



Principes

Random Forests

L'objectif des forêts aléatoires est de proposer un estimateur « moyenné » afin d'améliorer la robustesse de l'estimation de la quantité d'intérêt

- On cherche à diminuer la variance de l'estimateur final

Il s'agit d'intégrer une multitude de prévisions obtenues dans une estimation finale. Deux propriétés :

- On peut dégager un classement robuste du pouvoir explicatif de chacun des facteurs de risque
- Sa consistance a été démontrée

Prévisions continues vs. discrètes

Soit \hat{Y}_i l'estimateur obtenu pour l'individu i par un **CART maximal**

On construit B arbres CART en modifiant l'échantillon à chaque fois. Pour chaque observation i , l'estimateur forêts aléatoires vaut

une **moyenne** dans le cas où Y est continue

$$\hat{Y}_i^{RF} = \frac{1}{B} \sum_{b=1}^B \hat{Y}_{i,b}$$

un **vote majoritaire** si Y est discrète

$$\hat{Y}_i^{RF} = \arg \max_{k=A,B} \{\#\hat{Y}_{i,b} = k\}$$

Stratégies d'échantillonnage

Deux stratégies différentes pour introduire de l'aléatoire dans la construction des échantillons

Bagging

- On tire aléatoirement des individus avec remise (bootstrap)
- Limite la variance de l'estimateur issue de certains individus bien spécifiques

Randomization

- On tire aléatoirement un nombre plus faible de covariables
- Limite la variance de l'estimateur, notamment en cas de corrélation entre les covariables

Construction d'une forêt

Les forêts aléatoires sont basées sur plusieurs arbres CART. Chacun de ces arbres est construit comme suit.

1. Construire un échantillon bootstrap de même taille que l'apprentissage
2. Construire l'arbre CART sur cet échantillon bootstrap : considérons qu'il y a k covariables, avec $m \ll k$
 - à chaque nœud, on tire aléatoirement m covariables parmi les k disponibles ;
 - on cherche la division optimale basée sur ces m covariables ;
3. Agréger ces arbres pour construire l'estimateur forêt

Stratégies d'élagage

Comment élagué chaque arbre ? On distingue 3 stratégies :

1. Laisser construire l'arbre maximal pour chaque échantillon
 - Bon compromis pour le volume des calculs / qualités des prévisions
 - faible biais et grande variance de chaque estimateur
2. Fixer le nombre de feuilles maximal
3. Construire l'arbre maximal puis élaguer par validations croisée
 - pénalise lourdement la quantité de calculs sans gain substantiel de prévision

A ne jamais oublier

« RF is an example of a tool that is useful in doing analyses of scientific data [...]»

But the cleverest algorithms are no substitute for human intelligence and knowledge of the data in the problem [...]»

Take the output of RF not as absolute truth, but as smart computer generated guesses that may be helpful in leading to a deeper understanding of the problem »

Leo Breiman





Force, corrélation et erreur

Erreur de la forêt

L'erreur associée à la forêt dépend de 2 paramètres

- la **corrélation** entre les arbres de la forêt : plus cette corrélation augmente, plus l'erreur est grande
- la capacité de chaque arbre dans la forêt à donner une estimation proche de la réalité (**force**) : Plus l'arbre est précis, moins l'erreur est grande

Par rapport au paramètre de tuning m on observe que

- Abaisser m réduit la corrélation et la force
- Agrandir m augmente la corrélation et la force

Arbitrage à trouver sur m pour minimiser l'erreur

Erreur out-of-bag (OOB)

Au sein de la construction de chaque arbre CART de la forêt, on ne considère qu'une portion de l'échantillon bootstrap correspondant : le reste constitue les données « out-of-bag »

C'est sur ces données out-of-bag que sont calculées

- une estimation non-biaisée de l'erreur de l'arbre
- une estimation de l'importance des facteurs de risque

Ici pas de **validation croisée** pour avoir une estimation non-biaisée de l'erreur

- on prend les observations et prévision chaque fois qu'elles sont dans l'échantillon OOB
- calcul de l'erreur individuelle
- moyenne des erreurs individuelles

Randomization

La randomization permet de diminuer la corrélation entre les arbres, et de traiter le problème de covariables corrélées qui induisent un biais.

La variance de la moyenne de B estimateurs indépendants vaut

$$Var\left(\frac{1}{B} \sum_{b=1}^B X_b\right) = \frac{1}{B^2} Var\left(\sum_{b=1}^B X_b\right) = \frac{\sigma^2}{B}$$

En revanche, si ces estimateurs sont corrélés 2 à 2, de coefficient de corrélation ρ

$$Var\left(\frac{1}{B} \sum_{b=1}^B X_b\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Randomization

Ainsi

- si $\rho \rightarrow 0$, alors on retrouve le cas initial
- si $\rho \rightarrow 1$, alors on a beau augmenter B , il restera toujours $\rho\sigma^2$
- Cela limite fortement l'avantage du bagging ...

Conclusion : lors de l'agrégation, on diminue ainsi la variance de l'estimateur tout en conservant le même ordre de grandeur pour le biais.. L'erreur globale de l'estimateur diminue donc !

Grâce à la randomization, la stratégie d'élagage peut être plus élémentaire qu'en pur bagging !



Conclusions

Pour et contre

- Simple à mettre œuvre et à comprendre
- Programmation facile, quelque soit la méthode
- Diminue la variance de l'estimateur
- Temps de calcul parfois important : nécessité d'agréger un grand nombre de modèles avant de stabiliser l'erreur OOB
- Stockage de tous les modèles demande une grande capacité en mémoire
- Perte de l'interprétabilité (on arrive dans les boîtes noires)