LA FORÊT ALÉATOIRE

La forêt aléatoire (*random forests*) est une méthode ensembliste puissante, largement utilisée pour les tâches de classification et de régression. Elle combine la simplicité des arbres de décision avec la puissance de l'agrégation pour améliorer les performances prédictives et réduire le risque de surapprentissage (overfitting).

La méthode s'appuie sur la technique du bagging, qui consiste à entraîner chaque arbre sur un échantillon (*bootstrap*) tiré au hasard à partir du jeu de données initial. Elle introduit un degré supplémentaire de randomisation au moment de la construction d'un arbre, puisqu'à chaque nouvelle division (*noeud*), un sous-ensemble de variables sur lequel sera fondé le critère de séparation est **sélectionné aléatoirement**. Cette randomisation supplémentaire **réduit la corrélation** entre les arbres, ce qui permet de diminuer la variance des prédiction du modèle agrégé.

Objectifs:

- comprendre les propriétés fondamentales des forêts aléatoires afin de comprendre comment elles améliorent les performances des modèles
- comprendre les étapes permettant de construire une forêt aléatoire : échantillonnage bootstrap, sélection de variables, partitions, prédiction, évaluation, interprétation
- permettre l'optimisation des performances du modèle: fournir une présentation formelle pour guider le choix des hyperparamètres, la préparation des données, etc.

1. Principe de la forêt aléatoire

Les forêts aléatoires reposent sur plusieurs éléments essentiels :

- **Les arbres CART**: Les modèles élémentaires sont des arbres CART non élagués, c'est-à-dire autorisés à pousser jusqu'à l'atteinte d'un critère d'arrêt défini en amont.
- **L'échantillonnage bootstrap**: Chaque arbre est construit à partir d'un échantillon aléatoire tiré avec remise du jeu de données d'entraînement.
- **La sélection aléatoire de caractéristiques (*variables*)** : Lors de la construction d'un arbre, à chaque nœud de celui-ci, un sous-ensemble aléatoire de

variables est sélectionné. La meilleure division est ensuite choisie parmi ces caractéristiques aléatoires.

- **L'agrégation des prédictions** : Comme pour le bagging, les prédictions de tous les arbres sont combinées On procède généralement à la moyenne (ou à la médiane) des prédictions dans le cas de la régression, et au vote majoritaire (ou à la moyenne des probabilités prédites pour chaque classe) dans le cas de la classification.

2. Pourquoi (et dans quelles situations) la random forest fonctionne: fondements théoriques des forêts aléatoires

Les propriétés théoriques des forêts aléatoires permettent de comprendre pourquoi (et dans quelles situations) elles sont particulièrement robustes et performantes.

2.1. **Réduction de la variance par agrégation.**

L'agrégation de plusieurs arbres dans une forêt aléatoire permet de réduire la variance globale du modèle, ce qui améliore la stabilité et la précision des prédictions lorsque les biais initiaux sont "faibles". La démonstration est présentée dans la section *bagging*.

2.2. **Convergence et absence de surapprentissage.**

Les forêts aléatoires sont très performantes en pratique, mais la question de leur convergence vers une solution optimale lorsque la taille de l'échantillon tend vers l'infini reste ouverte (Loupes 2014). Plusieurs travaux théoriques ont toutefois fourni des preuves de consistance pour des versions simplifiées de l'algorithme (par exemple, Biau 2012).

Même si la convergence vers le modèle optimal n'est pas encore formellement établie, il est possible de montrer, grâce à la Loi des Grands Nombres, que l'erreur de généralisation de l'ensemble diminue lorsque le **nombre d'arbres** augmente. Cela implique que la forêt aléatoire ne souffre pas d'un surapprentissage (également appelé *overfitting*) croissant avec le nombre d'arbres inclus dans le modèle, ce qui la rend particulièrement robuste.

2.3. **Facteurs déterminants la diminution de l'erreur de généralisation.**

L'erreur de généralisation des forêts aléatoires est influencée par deux facteurs principaux :

- **La force (*puissance prédictrice*) des arbres individuels** : La force d'un arbre dépend de sa capacité à faire des prédictions correctes (sans biais). Pour que l'ensemble soit performant, chaque arbre doit être suffisamment prédictif.
- **La corrélation entre les arbres** : Une corrélation faible entre les arbres améliore les performances globales. En effet, des arbres fortement corrélés auront tendance à faire des erreurs similaires. La randomisation des caractéristiques à chaque nœud contribue à réduire cette corrélation, ce qui améliore la précision globale.

Dans le cas de la régression, où l'objectif est de minimiser l'erreur quadratique moyenne, la décomposition de la variance de l'ensemble (la forêt aléatoire) permet de faire apparaître le rôle de la corrélation entre les arbres:

$$\text{Var}\left(\hat{f}(x)\right) = \rho(x)\sigma(x)^2 + \frac{1-\rho(x)}{M}\sigma(x)^2$$

Où $\rho(x)$ est le coefficient de corrélation entre les arbres individuels, $\sigma(x)^2$ est la variance d'un arbre individuel, M est le nombre d'arbres dans la forêt.

Conséquences:

- **Si $\rho(x)$ est faible** : La variance est significativement réduite avec l'augmentation du nombre d'arbres M.
- **Si $\rho(x)$ est élevée** : La réduction de variance est moindre, car les arbres sont plus corrélés entre eux.

L'objectif des forêts aléatoires est donc de minimiser la corrélation entre les arbres tout en maximisant leur capacité à prédire correctement, ce qui permet de réduire la variance globale sans augmenter excessivement le biais.

3. Les hyper-paramètres clés

- **Nombre d'arbres** : plus le nombre d'arbes est élevé, plus la variance est réduite, jusqu'à un certain point de saturation. Souvent, quelques centaines d'arbres suffisent à stabiliser les performances des modèles.
- **Nombre de variables à sélectionner à chaque noeud**: cet hyperparamètre permet de contrôler l'interdépendance entre les arbres. Avec K le nombre total

de variables, une règle empirique usuelle est de considérer $\sqrt{K}$ pour une classification et $K/3$ pour une régression.

- **Profondeur des arbres** : laisser les arbres se développer pleinement (sans élagage) pour profiter de la réduction de variance par agrégation.

4. Evaluation des performances du modèle et choix des hyper-paramètres

4.1. Estimation de l'erreur Out-of-Bag (OOB).

L'estimation **Out-of-Bag (OOB)** est une méthode particulièrement efficace pour évaluer les performances des forêts aléatoires sans nécessité une **validation croisée** ou de réserver une partie des données pour l'étape du test. Cette technique repose sur le fait que chaque arbre dans une forêt aléatoire est construit à partir d'un échantillon bootstrap du jeu de données d'origine, c'est-à-dire un échantillon tiré avec remise. Or, en moyenne, environ **36 %** des observations ne sont pas inclus dans chaque échantillon bootstrap, ce qui signifie qu'elles ne sont pas utilisées pour entraîner l'arbre correspondant. Ces observations laissées de côté forment un **échantillon out-of-bag**. Chaque arbre peut donc être évalué sur son **échantillon out-of-bag** plutôt que sur un échantillon test.

Procédure d'Estimation OOB:

1. **Construction des arbres** : Chaque arbre de la forêt est construit à partir d'un échantillon bootstrap tiré avec remise à partir du jeu de données d'origine. Cela signifie que certaines observations seront sélectionnées plusieurs fois, tandis que d'autres ne seront pas sélectionnées du tout.
2. **Prédiction OOB** : Pour chaque observation (x_i, y_i) qui n'a pas été inclus dans l'échantillon bootstrap qui a permi de construire un arbre donné, l'arbre est utilisé pour prédire la valeur de y_i. Ainsi, chaque observation est prédite par tous les arbres pour lesquels elle fait partie de l'échantillon out-of-bag.
3. **Agrégation des prédictions** : La prédiction finale pour chaque échantillon out-of-bag est obtenue en moyennant les prédictions de tous les arbres pour lesquels cet échantillon était OOB (pour la régression) ou par un vote majoritaire (pour la classification).
4. **Calcul de l'erreur OOB** : L'erreur OOB est ensuite calculée en comparant les prédictions agrégées avec les valeurs réelles des observations y_i. Cette erreur est une bonne approximation de l'erreur de généralisation du modèle.

Avantages de l'Estimation OOB:

- **Pas besoin de jeu de validation séparé** : L'un des principaux avantages de l'estimation OOB est qu'elle ne nécessite pas de réserver une partie des données pour la validation. Cela est particulièrement utile lorsque la taille du jeu de données est limitée, car toutes les données peuvent être utilisées pour l'entraînement tout en ayant une estimation fiable de la performance.
- **Estimation directe et éfficace** : Contrairement à la validation croisée qui peut être coûteuse en temps de calcul, l'estimation OOB est disponible "gratuitement" pendant la construction des arbres. Cela permet d'évaluer la performance du modèle sans avoir besoin de réentraîner plusieurs fois le modèle.
- **Approximation de l'erreur de généralisation** : L'erreur OOB est considérée comme une bonne approximation de l'erreur de généralisation, comparable à celle obtenue par une validation croisée 10-fold.

4.2. **Estimation de l'erreur par cross-validation.**

La validation croisée est une technique d'évaluation couramment utilisée en apprentissage automatique pour estimer la capacité d'un modèle à généraliser à de nouvelles données. Bien que l'estimation Out-of-Bag (OOB) soit généralement suffisante pour les forêts aléatoires, la validation croisée permet d'obtenir une évaluation plus robuste, notamment sur des jeux de données de petite taille.

L'idée derrière la validation croisée est de maximiser l'utilisation des données disponibles en réutilisant chaque observation à la fois pour l'entraînement et pour le test. Cela permet d'obtenir une estimation de la performance du modèle qui est moins sensible aux fluctuations dues à la division des données en ensembles d'entraînement et de test. La validation croisée répète cette division plusieurs fois, puis moyenne les résultats pour obtenir une estimation plus fiable.

Procédure de validation croisée:

La validation croisée la plus courante est la validation croisée en k sous-échantillons (*k-fold cross-validation*):

- **Division des données** : Le jeu de données est divisé en k sous-échantillons égaux, appelés folds. Typiquement, k est choisi entre 5 et 10, mais il peut être ajusté en fonction de la taille des données.

- **Entraînement et test** : Le modèle est entraîné sur k - 1 sous-échantillons et testé sur le sous-échantillon restant. Cette opération est répétée k fois, chaque sous-échantillon jouant à tour de rôle le rôle de jeu de test.
- **Calcul de la performance** : Les k performances obtenues (par exemple, l'erreur quadratique moyenne pour une régression, ou l'accuracy (*exactitude*) pour une classification) sont moyennées pour obtenir une estimation finale de la performance du modèle.

Avantages de la validation croisée:

- **Utilisation optimale des données** : En particulier lorsque les données sont limitées, la validation croisée maximise l'utilisation de l'ensemble des données en permettant à chaque échantillon de contribuer à la fois à l'entraînement et au test.
- **Réduction de la variance** : En utilisant plusieurs divisions des données, on obtient une estimation de la performance moins sensible aux particularités d'une seule division.

Bien que plus coûteuse en termes de calcul, la validation croisée est souvent préférée lorsque les données sont limitées ou lorsque l'on souhaite évaluer différents modèles ou hyperparamètres avec précision.

Leave-One-Out Cross-Validation (LOOCV) : Il s'agit d'un cas particulier où le nombre de sous-échantillons est égal à la taille du jeu de données. En d'autres termes, chaque échantillon est utilisé une fois comme jeu de test, et tous les autres échantillons pour l'entraînement. LOOCV fournit une estimation très précise de la performance, mais est très coûteuse en temps de calcul, surtout pour de grands jeux de données.

4.3. **Choix des hyper-paramètres du modèle.**

L'estimation Out-of-Bag (OOB) et la validation croisée sont deux méthodes clés pour optimiser les hyper-paramètres d'une forêt aléatoire. Les deux approches permettent de comparer différentes combinaisons d'hyper-paramètres et de sélectionner celles qui maximisent les performances prédictives, l'OOB étant souvent plus rapide et moins coûteuse, tandis que la validation croisée est plus fiable dans des situations où le surapprentissage est un risque important.

5. Interprétation et importance des variables

Objectif: Identifier les variables qui contribuent le plus à la prédiction/influencent le plus la variable cible, extraire des caractéristiques pertinentes pour comprendre les mécanismes de prédiction sous-jacent, établir des règles de décision simplifiées etc.

Méthodes ususelles (mais biaisées):

- **Réduction moyenne de l'impureté** (*Mean Decrease in Impurity*) : pour chaque variable, on calcule la moyenne des réductions d'impureté qu'elle a engendrées dans tous les nœuds de tous les arbres où elle est impliquée. Les variables présentant la réduction moyenne d'impureté la plus élevée sont considérées comme les prédicteurs les plus importants.
- **Permutation importance** (*Mean Decrease Accuracy*) : Pour chaque variable, les performances du modèle sont comparées avant et après la permutation de ses valeurs. La différence moyenne de performance correspond à la MDA. L'idée est que si l'on permute aléatoirement les valeurs d'une variable (cassant ainsi sa relation avec la cible), une variable importante entraînera une hausse significative de l'erreur de généralisation.

Il est essentiel de noter que ces deux mesures peuvent présenter des **biais** importants. Elles sont notamment sensibles aux variables catégorielles avec de nombreuses modalités, qui peuvent apparaître artificiellement importantes. Elles sont également fortement biaisée en présence de variables explicatives corrélées ou d'interactions complexes avec la cible.

6. Préparation des données (Feature Engineering)

- **Variables catégorielles** : Encoder correctement (one-hot encoding, ordinal encoding).
- **Valeurs manquantes** : Les forêts aléatoires peuvent gérer les données manquantes, mais une imputation préalable peut améliorer les performances.
- **Échelle des Variables** : Pas nécessaire de normaliser, les arbres sont invariants aux transformations monotones.

References