LE BAGGING

Le bagging, ou "bootstrap aggregating", est une méthode ensembliste qui vise à améliorer la stabilité et la précision des algorithmes d'apprentissage automatique en agrégeant plusieurs modèles (Breiman (1996)). Chaque modèle est entraîné sur un échantillon distinct généré par une technique de rééchantillonnage (*bootstrap*). Ces modèles sont ensuite combinés pour produire une prédiction agrégée, souvent plus robuste et généralisable que celle obtenue par un modèle unique.

1. Principe du bagging

Le bagging comporte trois étapes principales:

- **L'échantillonnage bootstrap** : L'échantillonnage bootstrap consiste à créer des échantillons distincts en tirant aléatoirement avec remise des observations du jeu de données initial. Chaque échantillon *bootstrap* contient le même nombre d'observations que le jeu de données initial, mais certaines observations sont répétées (car sélectionnées plusieurs fois), tandis que d'autres sont omises.
- **L'entraînement de plusieurs modèles** : Un modèle (aussi appelé *apprenant de base* ou *weak learner*) est entraîné sur chaque échantillon bootstrap. Les modèles peuvent être des arbres de décision, des régressions ou tout autre algorithme d'apprentissage. Le bagging est particulièrement efficace avec des modèles instables, tels que les arbres de décision non élagués.
- **L'agrégation des prédictions** : Les prédictions de tous les modèles sont ensuite agrégées, en procédant 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, afin d'obtenir des prédictions plus précises et généralisables.

2. Pourquoi (et dans quelles situations) le bagging fonctionne

Certains modèles sont très sensibles aux données d'entraînement, et leurs prédictions sont très instables d'un échantillon à l'autre. L'objectif du bagging est de construire un prédicteur plus précis en agrégeant les prédictions de plusieurs modèles entraînés sur des échantillons (légèrement) différents les uns des autres.

Breiman (1996) montre que cette méthode est particulièrement efficace lorsqu'elle est appliquée à des modèles très instables, dont les performances sont particulièrement sensibles aux variations du jeu de données d'entraînement, et peu biaisés.

Cette section vise à mieux comprendre comment (et sous quelles conditions) l'agrégation par bagging permet de construire un prédicteur plus performant.

Dans la suite, nous notons $\varphi(x, L)$ un prédicteur (d'une valeur numérique dans le cas de la *régression* ou d'un label dans le cas de la *classification*), entraîné sur un ensemble d'apprentissage L, et prenant en entrée un vecteur de caractéristiques x.

2.1. **La régression: réduction de l'erreur quadratique moyenne par agrégation.**

Dans le contexte de la **régression**, l'objectif est de prédire une valeur numérique Y à partir d'un vecteur de caractéristiques x. Un modèle de régression $\phi(x, L)$ est construit à partir d'un ensemble d'apprentissage L, et produit une estimation de Y pour chaque observation x.

2.1.1. *Définition du prédicteur agrégé.*

Dans le cas de la régression, le **prédicteur agrégé** est défini comme suit :

$ _A(x) = E_L[(x, L)] $

où $\phi_A(x)$ représente la prédiction agrégée, $E_L[.]$ correspond à l'espérance prise sur tous les échantillons d'apprentissage possibles L, chacun étant tiré selon la même distribution que le jeu de données initial, et $\phi(x, L)$ correspond à la prédiction du modèle construit sur l'échantillon d'apprentissage L.

2.1.2. *La décomposition biais-variance.*

Pour mieux comprendre comment l'agrégation améliore la performance globale d'un modèle individuel $\phi(x, L)$, revenons à la **décomposition biais-variance** de l'erreur quadratique moyenne (il s'agit de la mesure de performance classiquement considérée dans un problème de régression):

$$E_L\left[(Y - \phi(x, L))^2\right] = \underbrace{\left(E_L[\phi(x, L) - Y]\right)^2}_{\text{Biais}^2} + \underbrace{E_L\left[(\phi(x, L) - E_L[\phi(x, L)])^2\right]}_{\text{Variance}} \quad (1)$$

- Le **biais** est la différence entre la valeur observée Y que l'on souhaite prédire et la prédiction moyenne $E_L[\phi(x, L)]$. Si le modèle est sous-ajusté, le biais sera élevé.
- La **variance** est la variabilité des prédictions ($\phi(x, L)$) autour de leur moyenne ($E_L[\phi(x, L)]$). Un modèle avec une variance élevée est très sensible aux fluctuations au sein des données d'entraînement: ses prédictions varient beaucoup lorsque les données d'entraînement se modifient.

L'équation Equation 1 illustre l'**arbitrage biais-variance** qui est omniprésent en *machine learning*: plus la complexité d'un modèle s'accroît (exemple: la profondeur d'un arbre), plus son biais sera plus faible (car ses prédictions seront de plus en plus proches des données d'entraînement), et plus sa variance sera élevée (car ses prédictions, étant très proches des données d'entraînement, auront tendance à varier fortement d'un jeu d'entraînement à l'autre).

2.1.3. *L'inégalité de Breiman (1996).*

Breiman (1996) compare l'erreur quadratique moyenne d'un modèle individuel avec celle du modèle agrégé et démontre l'inégalité suivante :

$ (Y - _A(x))^2 E_L[(Y - (x, L))^2] $ {#eq-inegalite-breiman1996}

- Le terme $(Y - \phi_A(x))^2$ représente l'erreur quadratique du **prédicteur agrégé** $\phi_A(x)$;
- Le terme $E_L\left[(Y - \phi(x, L))^2\right]$ est l'erreur quadratique moyenne d'un **prédicteur individuel** $\phi(x, L)$ entraîné sur un échantillon aléatoire L. Cette erreur varie en fonction des données d'entraînement.

Cette inégalité montre que **l'erreur quadratique moyenne du prédicteur agrégé est toujours inférieure ou égale à la moyenne des erreurs des prédicteurs individuels**. Puisque le biais du prédicteur agrégé est identique au biais du prédicteur individuel, alors l'inégalité précédente implique que la **variance du modèle agrégé** $\phi_A(x)$ est **toujours inférieure ou égale** à la variance moyenne d'un modèle individuel :

$ (_A(x)) = (E_L[(x, L)]) E_L[((x, L))] $

Autrement dit, le processus d'agrégation réduit l'erreur de prédiction globale en réduisant la **variance** des prédictions, tout en conservant un biais constant.

Ce résultat ouvre la voie à des considérations pratiques immédiates. Lorsque le modèle individuel est instable et présente une variance élevée, l'inégalité $Var(\phi_A(x)) \leq$

$E_L[Var(\phi(x, L))]$ est forte, ce qui signifie que l'agrégation peut améliorer significativement la performance globale du modèle. En revanche, si $\phi(x, L)$ varie peu d'un ensemble d'entraînement à un autre (modèle stable avec variance faible), alors $Var(\phi_A(x))$ est proche de $E_L[Var(\phi(x, L))]$, et la réduction de variance apportée par l'agrégation est faible. Ainsi, **le bagging est particulièrement efficace pour les modèles instables**, tels que les arbres de décision, mais moins efficace pour les modèles stables tels que les méthodes des k plus proches voisins.

2.2. **La classification: vers un classificateur presque optimal par agrégation.**

Dans le cas de la classification, le mécanisme de réduction de la variance par le bagging permet, sous une certaine condition, d'atteindre un **classificateur presque optimal** (*nearly optimal classifier*). Ce concept a été introduit par Breiman (1996) pour décrire un modèle qui tend à classer une observation dans la classe la plus probable, avec une performance approchant celle du classificateur Bayésien optimal (la meilleure performance théorique qu'un modèle de classification puisse atteindre).

Pour comprendre ce résutlat, introduisons $Q(j \mid x) = E_L\left(1_{\varphi(x,L)=j}\right) = P(\varphi(x, L) = j)$, la probabilité qu'un modèle $\varphi(x, L)$ prédise la classe j pour l'observation x, et $P(j \mid x)$, la probabilité réelle (conditionnelle) que x appartienne à la classe j.

2.2.1. *Définition : classificateur order-correct.*

Un classificateur $\varphi(x, L)$ est dit **order-correct** pour une observation x si, en espérance, il identifie **correctement la classe la plus probable**, même s'il ne prédit pas toujours avec exactitude les probabilités associées à chaque classe $Q(j \mid x)$.

Cela signifie que si l'on considérait tous les ensemble de données possibles, et que l'on évaluait les prédictions du modèle en x, la majorité des prédictions correspondraient à la classe à laquelle il a la plus grande probabilité vraie d'appartenir $P(j \mid x)$.

Formellement, un prédicteur est dit "order-correct" pour une entrée x si :

$ argmax_j Q(j|x) = argmax_j P(j|x) $

où $P(j \mid x)$ est la vraie probabilité que l'observation x appartienne à la classe j, et $Q(j \mid x)$ est la probabilité que x appartienne à la classe j prédite par le modèle $\varphi(x, L)$.

Un classificateur est **order-correct** si, pour **chaque** observation x, la classe qu'il prédit correspond à celle qui a la probabilité maximale $P(j \mid x)$ dans la distribution vraie.

2.2.2. *Prédicteur agrégé en classification: le vote majoritaire.*

Dans le cas de la classification, le prédicteur agrégé est défini par le **vote majoritaire**. Cela signifie que si K classificateurs sont entraînés sur K échantillons distincts, la classe prédite pour x est celle qui reçoit **le plus de votes** de la part des modèles individuels.

Formellement, le classificateur agrégé $\varphi A(x)$ est défini par :

$ \varphiA(x) = j \{L\} I((x, L) = j) = argmax_j Q(j|x) $

2.2.3. *Performance globale: convergence vers un classificateur presque optimal.*

Breiman (1996) montre que si chaque prédicteur individuel $\varphi(x, L)$ est order-correct pour une observation x, alors le prédicteur agrégé $\varphi A(x)$, obtenu par **vote majoritaire**, atteint la performance optimale pour cette observation, c'est-à-dire qu'il converge vers la classe ayant la probabilité maximale $P(j \mid x)$ pour l'observation x lorsque le nombre de prédicteurs individuels augmente. Le vote majoritaire permet ainsi de **réduire les erreurs aléatoires** des classificateurs individuels.

Le classificateur agrégé ϕA est optimal s'il prédit systématiquement la classe la plus probable pour l'observation x dans toutes les régions de l'espace.

Cependant, dans les régions de l'espace où les classificateurs individuels ne sont pas order-corrects (c'est-à-dire qu'ils se trompent majoritairement sur la classe d'appartenance), l'agrégation par vote majoritaire n'améliore pas les performances. Elles peuvent même se détériorer par rapport aux modèles individuels si l'agrégation conduit à amplifier des erreurs systématiques (biais).

3. L'échantillage par bootstrap peut détériorer les performances théoriques du modèle agrégé

En pratique, au lieu d'utiliser tous les ensembles d'entraînement possibles L, le bagging repose sur un nombre limité d'échantillons bootstrap tirés avec remise à partir d'un même jeu de données initial, ce qui peut introduire des biais par rapport au prédicteur agrégé théorique.

Les échantillons bootstrap présentent les limites suivantes :

- Une **taille effective réduite par rapport au jeu de données initial**: Bien que chaque échantillon bootstrap présente le même nombre d'observations que

le jeu de données initial, environ 1/3 des observations (uniques) du jeu initial sont absentes de chaque échantillon bootstrap (du fait du tirage avec remise). Cela peut limiter la capacité des modèles à capturer des relations complexes au sein des données (et aboutir à des modèles individuels sous-ajustés par rapport à ce qui serait attendu théoriquement), en particulier lorsque l'échantillon initial est de taille modeste.

- Une **dépendance entre échantillons** : Les échantillons bootstrap sont tirés dans le même jeu de données, ce qui génère une dépendance entre eux, qui réduit la diversité des modèles. Cela peut limiter l'efficacité de la réduction de variance dans le cas de la régression, voire acroître le biais dans le cas de la classification.
- Une **couverture incomplète de l'ensemble des échantillons possibles**: Les échantillons bootstrap ne couvrent pas l'ensemble des échantillons d'entraînement possibles, ce qui peut introduire un biais supplémentaire par rapport au prédicteur agrégé théorique.

4. Le bagging en pratique

4.1. **Quand utiliser le bagging en pratique.**

Le bagging est particulièrement utile lorsque les modèles individuels présentent une variance élevée et sont instables. Dans de tels cas, l'agrégation des prédictions peut réduire significativement la variance globale, améliorant ainsi la performance du modèle agrégé. Les situations où le bagging est recommandé incluent typiquement:

- Les modèles instables : Les modèles tels que les arbres de décision non élagués, qui sont sensibles aux variations des données d'entraînement, bénéficient grandement du bagging. L'agrégation atténue les fluctuations des prédictions dues aux différents échantillons.
- Les modèles avec biais faibles: En classification, si les modèles individuels sont order-corrects pour la majorité des observations, le bagging peut améliorer la précision en renforçant les prédictions correctes et en réduisant les erreurs aléatoires.

Inversement, le bagging peut être moins efficace ou même néfaste dans certaines situations :

- Les modèles stables avec variance faible : Si les modèles individuels sont déjà stables et présentent une faible variance (par exemple, la régression linéaire), le bagging n'apporte que peu d'amélioration, car la réduction de variance supplémentaire est minimale.
- La présence de biais élevée : Si les modèles individuels sont biaisés, entraînant des erreurs systématiques, le bagging peut amplifier ces erreurs plutôt que de les corriger. Dans de tels cas, il est préférable de s'attaquer d'abord au biais des modèles avant de considérer l'agrégation.
- Les échantillons de petite taille : Avec des ensembles de données limités, les échantillons bootstrap peuvent ne pas être suffisamment diversifiés ou représentatifs, ce qui réduit l'efficacité du bagging et peut augmenter le biais des modèles.

Ce qui qu'il faut retenir: le bagging peut améliorer substantiellement la performance des modèles d'apprentissage automatique lorsqu'il est appliqué dans des conditions appropriées. Il est essentiel d'évaluer la variance et le biais des modèles individuels, ainsi que la taille et la représentativité du jeu de données, pour déterminer si le bagging est une stratégie adaptée. Lorsqu'il est utilisé judicieusement, le bagging peut conduire à des modèles plus robustes et précis, exploitant efficacement la puissance de l'agrégation pour améliorer la performance des modèles individuels.

4.2. **Comment utiliser le bagging en pratique.**

4.2.1. *Combien de modèles agréger?.*

"Optimal performance is often found by bagging 50–500 trees. Data sets that have a few strong predictors typically require less trees; whereas data sets with lots of noise or multiple strong predictors may need more. Using too many trees will not lead to overfitting. However, it's important to realize that since multiple models are being run, the more iterations you perform the more computational and time requirements you will have. As these demands increase, performing k-fold CV can become computationally burdensome."

4.2.2. *Evaluation du modèle: cross validation et échantillon Out-of-bag (OOB).*

"A benefit to creating ensembles via bagging, which is based on resampling with replacement, is that it can provide its own internal estimate of predictive performance

with the out-of-bag (OOB) sample (see Section 2.4.2). The OOB sample can be used to test predictive performance and the results usually compare well compared to k-fold CV assuming your data set is sufficiently large (say n≥1,000). Consequently, as your data sets become larger and your bagging iterations increase, it is common to use the OOB error estimate as a proxy for predictive performance."

5. Mise en pratique (exemple avec code)

Ou bien ne commencer les mises en pratique qu'avec les random forest ?

6. Interprétation

References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.