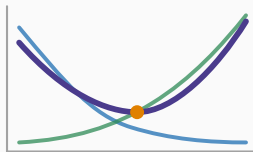


Apprentissage statistique

Chapitre 9 : Forêts aléatoires et bagging



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Moyenner pour stabiliser

La forêt aléatoire

Deux sous-produits utiles

Ce qu'il faut retenir

Moyenner pour stabiliser

Le résultat qui fonde le chapitre

La variance d'une moyenne

Pour B variables indépendantes de variance σ^2 , la moyenne a pour variance

$$\frac{\sigma^2}{B}$$

Moyenner 100 modèles indépendants divise donc la variance par 100 — **sans toucher au biais**.

Pourquoi c'est exactement ce qu'il fallait

L'arbre du chapitre 8 avait peu de biais et beaucoup de variance.

Moyenner ne change pas le biais et écrase la variance : l'opération est taillée sur mesure pour lui. C'est la raison pour laquelle les méthodes d'ensemble reposent sur des arbres et non sur des régressions.

Le bagging

Le **bagging** — agrégation bootstrap — consiste à :

1. tirer B échantillons **bootstrap** de l'échantillon d'origine ;
2. construire un arbre **profond, non élagué** sur chacun ;
3. moyenner les prédictions, ou voter à la majorité.

La forêt aléatoire

Le défaut du bagging

Le problème de la corrélation

La division par B suppose les modèles **indépendants**. Or les arbres du bagging ne le sont pas : ils voient presque les mêmes données.

Si une variable domine, **tous** les arbres la choisiront en premier, et se ressembleront. Moyenner des modèles corrélés ne réduit guère la variance.

L'idée de Breiman

À **chaque nœud**, ne considérer qu'un sous-ensemble aléatoire de m variables parmi p , tiré à nouveau à chaque fois.

Une variable dominante ne peut donc pas être choisie partout : les arbres sont forcés d'explorer d'autres découpages, et se **décorrèlent**.

Le réglage usuel de m

Tâche	Valeur courante
Classification	$m \approx \sqrt{p}$
Régression	$m \approx p/3$

Deux sous-produits utiles

L'erreur out-of-bag

Out-of-bag

Un échantillon bootstrap laisse de côté environ **un tiers** des observations. Chaque observation est donc « hors sac » pour environ un tiers des arbres.

On peut la prédire avec **ces arbres-là seulement** — qui ne l'ont jamais vue.

Le calcul de la proportion

La probabilité qu'une observation ne soit jamais tirée en n tirages avec remise :

$$\left(1 - \frac{1}{n}\right)^n \xrightarrow{n \rightarrow \infty} e^{-1} \approx 0,368$$

Environ **36,8 %** des observations sont hors sac pour un arbre donné.

Une validation croisée gratuite

L'erreur **out-of-bag** estime l'erreur de test sans découpage supplémentaire ni calcul additionnel.

C'est un avantage pratique considérable sur les méthodes qui exigent une validation croisée explicite — même si celle-ci reste préférable pour comparer des familles de modèles différentes.

L'importance des variables

Deux mesures

Par l'impureté — cumuler la réduction de Gini apportée par chaque variable sur tous les arbres. Rapide, mais **biaisée** en faveur des variables à nombreuses modalités.

Par permutation — mélanger au hasard une variable et mesurer la dégradation de l'erreur out-of-bag. Plus coûteuse, plus fiable.

Importante ne veut pas dire causale

Une variable peut être jugée importante parce qu'elle est corrélée à la vraie cause, ou parce qu'elle en est une conséquence.

De plus, entre deux variables très corrélées, la forêt **répartit** l'importance entre elles, ce qui fait paraître les deux modérément utiles alors que l'une suffirait.

L'importance mesure la contribution à la prévision, jamais un effet. C'est le même avertissement qu'au chapitre 4 sur le lasso, et il vaut pour tout ce cours.

Ce qu'il faut retenir

Cinq points

1. Moyenner B modèles indépendants divise la variance par B , **sans toucher au biais**.
2. Le **bagging** agrège des arbres profonds construits sur échantillons bootstrap.
3. La **forêt aléatoire** ajoute un tirage de m variables à chaque nœud, ce qui **décorrèle** les arbres.
4. $m \approx \sqrt{p}$ en classification ; augmenter B ne sur-ajuste jamais.
5. L'erreur **out-of-bag** — sur les 36,8 % non tirés — donne une estimation gratuite de l'erreur de test.

La suite

Le bagging construit les arbres **en parallèle**, chacun ignorant les autres. Le chapitre 10 les construit **en séquence**, chacun corrigeant les erreurs du précédent.