

Pourquoi les méthodes de boosting risquent-elles le surapprentissage ?

- Risque de sur-ajuster et donc d'être trop spécifique au jeu de données (biais très faible) :
 - chaque arbre risque, en corrigeant successivement les erreurs de l'arbre précédent, d'apprendre les détails trop spécifiques au jeu d'apprentissage.
 - En utilisant trop d'arbres, on amplifie ce phénomène.
- On peut éviter ce phénomène en réduisant **n.trees** et **eta**. / en utilisant le **subsample** / en programmant des **arbres moins profonds** / en introduisant une méthode de **VC** / utilisation du **early-stopping**.

Pourquoi les méthodes de bagging sont moins propices au sur-apprentissage ?

- Le mécanisme anti-surapprentissage des forêts aléatoires repose sur deux principes :
 - Bootstrap-aggregating : chaque arbre est entraîné sur un sous-échantillon aléatoire des données, tiré avec remise
 - La sélection aléatoire des variables : à chaque split d'un nœud, seul un sous-ensemble aléatoire des variables explicatives est considéré.
- Ainsi, les arbres individuels capturent différents aspects des données y compris les patterns complexes. Prendre la moyenne des prédictions de nombreux arbres permet d'annuler les erreurs aléatoires en gardant les patterns réels que les arbres auront capté

Qu'est ce que la descente de gradient

C'est un algorithme d'optimisation qui en ML permet de trouver le minimum d'une fonction de coût. Il part d'un point initial et se déplace itérativement dans la direction de la plus forte pente descendante. Cette direction est donnée par le gradient négatif de la fonction de coût par rapport aux paramètres.

Faire attention à :

- bien calibrer le taux d'apprentissage : trop faible et l'algorithme converge trop lentement, trop élevé et l'algorithme risque de diverger.
- Risque de converger vers un minimum local.

Quelles sont les différentes fonctions de perte ?

- **Régression**
 - Erreur quadratique moyenne (MSE) : pénalise fortement les grandes erreurs, sensible aux outliers
 - Erreur absolue moyenne (MAE) : plus robuste aux outliers que MSE
 - Racine de l'erreur quadratique moyenne (RMSE) : c'est la racine de la MSE. S'exprime dans la même unité que la variable cible et pénalise davantage les grandes erreurs que les petites. Sensible aux valeurs aberrantes.
- **Classification**
 - Log-loss : mesure la performance d'un modèle de classification dont les sorties sont des probabilités.

Qu'est-ce qu'une matrice de confusion ?

Une matrice de confusion est un tableau utilisé pour évaluer les performances d'un modèle de classification en comparant les prédictions avec les valeurs réelles.

- Éléments réellement positifs = VP + FN
- Éléments réellement négatifs = VN + FP
- Éléments prédits positifs = VP + FP
- Éléments prédit négatifs = VN et FN
- Métriques dérivées :
 - Précision = parmi les éléments prédits positifs, combien le sont réellement ?
 - Sensibilité = parmi les éléments positifs, combien sont correctement identifiés ?

- Spécificité = parmi les éléments négatifs, combien sont correctement identifiés ?
- Exactitude = Proportion totale de prédictions correctes.
- F1-Score = métrique qui combine la précision et le rappel

Qu'est-ce qu'un arbre ?

Un arbre binaire de décision CART est un algorithme de moyennage local par partition, dont la partition est construite par divisions successives au moyen d'hyperplans orthogonaux aux axes dépendant des données $(X_i ; Y_i)$.

A chaque étape, la méthode cherche une nouvelle division qui optimise une fonction de coût comme l'entropie de Shannon ou l'indice de Gini.

Quel sera l'effet de l'agrégation sur les arbres en termes de biais et de variance ?

Agréger ne modifie pas le biais mais réduit la variance, toutefois c'est valable uniquement si les estimateurs m_1, \dots, m_B sont iid. Or ce n'est pas le cas vu qu'ils sont construits sur le même échantillon. Il faut introduire de l'aléa via le Bagging (Bootstrap Aggregating) afin de tendre vers des estimateurs indépendants. Par ailleurs on peut également forcer chaque arbre à n'apprendre qu'à partir d'un sous ensemble des p variables du dataset. Chaque arbre apprend ainsi sur une partie spécifique du dataset.