5

Modèles Ensemblistes Bagging

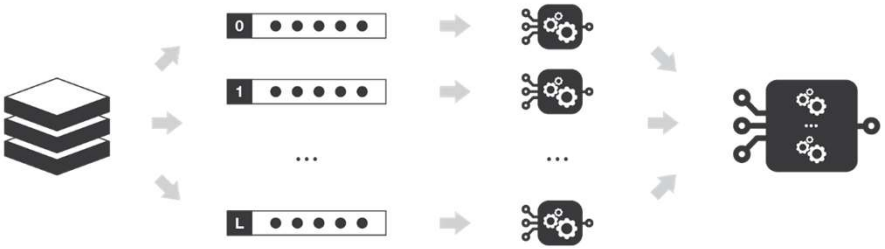
5. Modèles – Ensemblistes - Bagging

BAGGING (*bootstrap aggregation)* :

- Basée sur bootstrap (nouveaux datasets de taille identique à l'original avec remise, représentatifs et indépendants et identiquement distribués),
- entraînement de plusieurs modèles apprenants faibles presque indépendants en parallèle,
- moyenne de leurs prédictions afin d'obtenir un modèle avec une variance plus faible.

Il permet de réduire la variance des estimateurs individuels et offre une prédiction plus performante et plus stable, dépendante du nombre d'apprenants faibles utilisés.

Avantage : parallèlisation.



$\{z_1^1, z_2^1, ..., z_B^1\}, \{z_1^2, z_2^2, ..., z_B^2\}, ..., \{z_1^L, z_2^L, ..., z_B^L\}$

$z_b^l \equiv b$-th observation of the l-th bootstrap sample

$w_1(.), w_2(.), ..., w_L(.)$

$$s_L(.) = \frac{1}{L}\sum_{l=1}^{L} w_l(.)$$ (simple average, for regression problem)

$$s_L(.) = \arg\max_k [card(l|w_l(.) = k)]$$ (simple majority vote, for classification problem)

Algorithme :

- Des sous-ensembles aléatoires sont créés à partir de l'ensemble de données original (Bootstrapping). Le sous-ensemble de l'ensemble de données comprend toutes les caractéristiques.
- Un estimateur de base spécifié par l'utilisateur est ajusté sur chacun de ces petits ensembles.
- Les prédictions de chaque modèle sont combinées pour obtenir le résultat final.

5. Modèles – Ensemblistes - Bagging

Hyperparamètres :

base_estimator : Il définit l'estimateur de base à adapter sur des sous-ensembles aléatoires de l'ensemble de données.
Si rien n'est spécifié, l'estimateur de base est un arbre de décision.

n_estimators : C'est le nombre d'estimateurs de base à créer. Le nombre d'estimateurs doit être soigneusement ajusté car un grand nombre prendrait beaucoup de temps à exécuter, tandis qu'un très petit nombre pourrait ne pas fournir les meilleurs résultats.

max_samples : Ce paramètre contrôle la taille des sous-ensembles. Il s'agit du nombre maximum d'échantillons pour entraîner chaque estimateur de base.

max_features : Ce paramètre contrôle le nombre de caractéristiques à tirer de l'ensemble des données. Il définit le nombre maximum de caractéristiques requises pour entraîner chaque estimateur de base.

n_jobs : Le nombre de tâches à exécuter en parallèle. Définissez cette valeur comme étant égale au nombre de cœurs dans votre système. Si la valeur est -1, le nombre de tâches est égal au nombre de cœurs.

random_state : Indique la méthode de répartition aléatoire. Lorsque la valeur de random state est la même pour deux modèles, la sélection aléatoire est la même pour les deux modèles. Ce paramètre est utile lorsque vous souhaitez comparer différents modèles.