RER – Random Forest

Contexte : Découvrir le modèle de machine learning et ses concepts de base des random forest.

Problématique(s) :

Comment créer et évaluer un modèle random forest

Comment optimiser le modèle

Qu'est-ce que la régularisation ? Quelles sont les techniques de régularisation ?

Mots clés :

Méthode d'ensemble : méthode ensembliste, on fait l'agrégation des prédictions par exemple par moyenne, le max, la majorité,… de plusieurs prédicteur ou classifier pour une meilleure généralisation et pour compenser les défauts éventuels de prédicteurs individuels.

La méthode d'ensemble consiste à utiliser plusieurs algorithmes d'apprentissage automatique, en les mettant en commun pour obtenir des prédictions de meilleure qualité.

RandomForest Classifier: algorithme qui appartient aux méthodes ensemblistes d'apprentissage automatique qui combine la sortie de plusieurs arbres de décision pour obtenir un résultat unique. En général on utilise le vote par majorité mais on peut faire d'autre agrégations (ex régression , max , moyenne des valeurs prédites).

Les paramètres les plus utilisés sont:

n_estimator nombre d'arbres décisionnel dans la forêt,

max_deph il s'agit de la profondeur maximale des arbres utilisés

max_feature nombre d'attribut sélectionné pour chaque arbre (par défaut racine carré du nombre d'attribut)

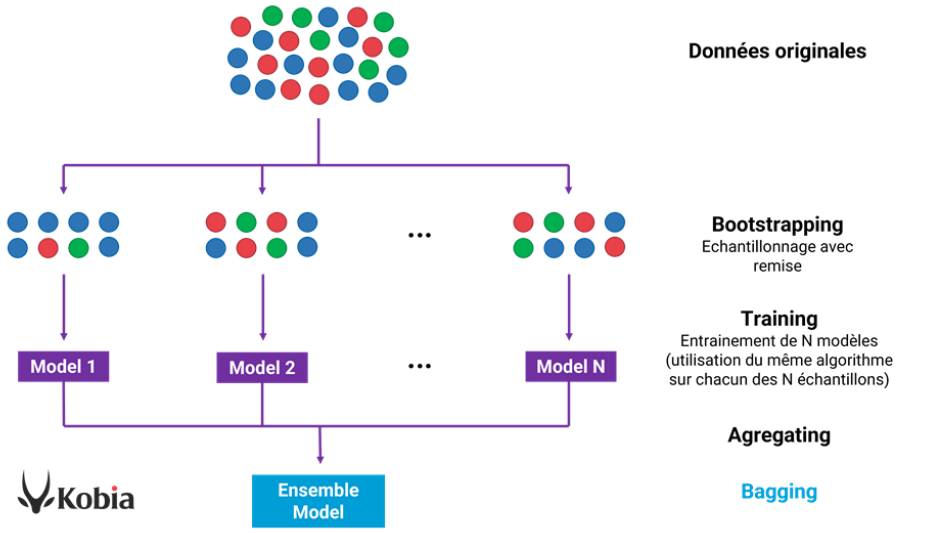
bootstrap sert à distribuer le dataset sur les arbres, False on utilise le même échantillon

feature_importances : indique quel champ à le plus grand impact sur chaque prédiction générée. On peut donner pour chaque valeur d'importance une magnitude et une direction positive ou négative qui indique comment chaque champ affecte une prédiction particulière.

Bagging ensachage

Créé plusieurs copies d'un même classifier en entrainant chaque copie sur une partie du data set .

Cela permet l'entrainement en parallèle. On l'utilise lorsqu'il y a overfitting la foule permet de réduire la variance et obtenir une meilleure généralisation. Le tirage aléatoire de sous ensemble d'entrainement se fait avec remise dans le sac.



Pasting collage même principe que le bagging sauf que Le tirage aléatoire de sous ensemble d'entrainement se fait sans remise dans le sac.

KNeighborsClassifier cas des plus proches voisins test statistique non paramétrique (suite au prochain épisode). Le principe de ce modèle consiste en effet à choisir les k données les plus proches du point étudié afin d'en prédire sa valeur.

Out-Of-Bag: c'est une métrique qui s'applique lors du bagging. Échantillon qui n'a pas été choisi. Lors de l'exécution de bagging sur un ensemble d'apprentissage seulement 63% des instance sont inclus dans le sous-ensemble d'apprentissage de ce qui signifie qu'il y a 37% des données que le classifier (ce sont des valeurs statistique) n'a pas vu avant elles peuvent être utilisées pour la validation croisée ou le test

Random Subspaces (sous-ensemble aléatoire): sous ensemble des variables d'entrées utilisé pour entrainer un prédicteur comme un arbre de décision. Le sous ensemble est constitué d'attributs choisis aléatoirement. On prend l'ensemble des enregistrements.

Random Patches Echantillonnage conjoint choix des colonnes d'entrée et des enregistrements ou observations d'entrainement de façon aléatoire.

Optimisation c'est l'amélioration des performances du modèle en machine learning, amélioration des hyperparamètres selon les contraintes de temps, de coût.

Régularisation a pour objectif :

1. minimiser la complexité du modèle
2. minimiser la fonction coût
3. parer au sur-apprentissage

Gridsearch fonction python utilisé pour l'optimisation, permettant de choisir les meilleurs hyperparamètres et d'obtenir le meilleur modèle

GridsearchCV CV = cross validation fonction scikitlearn

Elle prend en argument le classifieur, les valeur d'hyperparamètres à tester (sous forme d'un dictionnaire de listes avec comme clés le nom des paramètres), le nombre de k-fold pour l'étape de cross validation.

Bootstrapping : Technique d'échantillonnage en statistique (tirage au sort) en Machine Learning par tirage aléatoire dans une population donnée. (terme général). Elle a pour objectif d'entrainer différents modèles.

Sample (échantillon) n le nombre d'éléments pris aléatoirement dans un ensemble d'éléments noté N

Hypothèses :

Osman random forest est plus rapide et plus précis que ID3 FAUX plus précis mais pas plus rapide

Jean-Paul dans random forest il n'y a pas de racine, c'est une forêt sans racine. FAUX

Tetyana Random Forest peut être utilisé avec toutes sortes de données et dans tous les domaines de classification VRAI

Aude les racines des random forest sont choisi aléatoirement FAUX

Etienne gridsearch permet de scanner l'hyper espace en machine learning FAUX

Adeline les modèles randoms forest ne sont pas tous performant selon les domaines d'activités (données) FAUX

Briand random forest ne donne pas de sortie précise mais une approximation en fonction de la moyenne majoritaire des arbres décisionnels FAUX elle est précise

Seydou les arbres de Random Forest n'ont pas tous le même poids. VRAI

Axel

Loïc : Random forest n'est ni plus ni moins qu'un assemblage de plusieurs Decision Tree. VRAI

Nicolas le bootstrap du bagging est le même bootstrap d'internet. FAUX

Solenn le features importance dans scikitlearn permet de donner un poids dans un arbre d'une forêt FAUX mais on peut

Adrien les random forest les arbres sont choisis aléatoirement. FAUX

Adeline la prédiction finale d'un modèle random forest est la moyenne de décision aléatoire. VRAI dans la régression

Loïc On utilise le random forest pour augmenter le nombre de génération de decision tree afin d'en sortir un résultat plus précis encore que le decision tree VRAI

Briand le bagging et le pasting constitent à regrouper des arbres par correspondance FAUX

JP II il faut un minimum d'attribut pour générer une Random Forest FAUX ça n'a rien à voir avec le nombre d'attribut

Random Forrest	Decision Tree
Forces : **Permet d'obtenir de bons résultats assez rapidement** **Basée sur des principes simples** **Simplement implémentable** **Les calculs de l'apprentissage peuvent être facilement distribués**	**Forces :** Adoptée dans le cas de données peu ordonnées et claires. Nécessite très peu de préparation des données Grande interprétabilité.
Ses faiblesses : **Temps d'apprentissage peut être un peu lent** **Souvent dépassée par d'autres méthodes plus difficiles à mettre en place mais souvent plus efficaces telles que les GBM (gradient boosting machine)** **Reste l'une des plus utilisée en machine learning.**	**Faiblesses :** Extrêmement sensible à de petits changements dans le training set.

	Decision Tree	Random Forest
Interpretability	Easy to interpret	Hard to interpret
Accuracy	Accuracy can vary	Highly accurate
Overfitting	Likely to overfit data	Unlikely to overfit data
Outliers	Can be highly affected by outliers	Robust against outliers
Computation	Quick to build	Slow to build (computationally intensive)

Plan d'action :

Explorer les ressources

Définir et comprendre les mots clefs

Faire le Workshop

Tableau comparatif entre random forest et arbre de décision

Utilisation du gridsearch et validation croisée pour l'optimisation

RER