5

Modèles Arbres de décision

5. Modèles – Arbres de décision

Un **DecisionTree** est un algorithme d'apprentissage automatique supervisé.

Ce système non paramétrique, contrairement aux modèles de régression linéaire (qui supposent la linéarité), ne fait aucune hypothèse sous-jacente sur la distribution des erreurs ou des données.

Il s'agit d'une structure de type organigramme, composée de plusieurs questions (nœud) et qui, en fonction des réponses (branche) données, conduira à une étiquette de classe ou à une valeur (feuille) lorsqu'elle sera appliquée à une observation quelconque.

Les arbres de décision permettent de classer des objets en effectuant des décisions successives sur la base de leurs variables.

L'algorithme de base pour la construction d'arbres de décision utilise une recherche descendante top-down, greedy search dans l'espace des branches possibles, sans retour en arrière.

Les noeuds de l'arbre représentent ces décisions alors que les feuilles représentent les valeurs de la variable cible (à prédire).

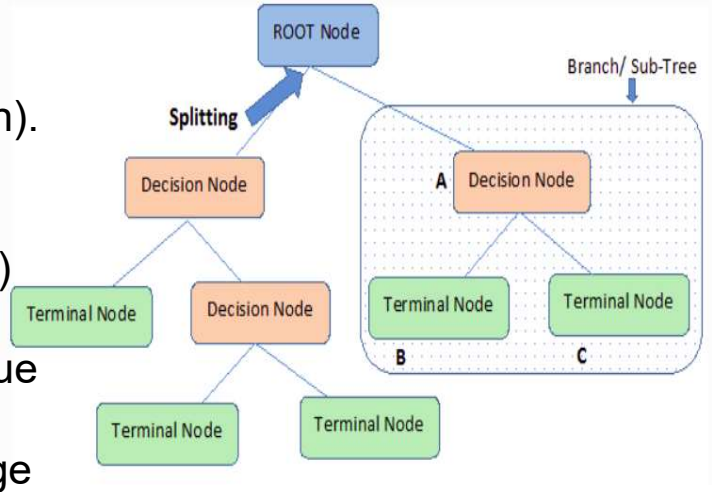
Plusieurs algorithmes existent permettant de générer des arbres de décision (*ID3*, *C4.5*, *CART*, etc.). Chaque algorithme utilise un critère pour choisir la variable à utiliser sur un noeud (gain d'information, entropie, écart-type, MSE (régression) etc.)

5. Modèles – Arbres de décision

Algorithme :
Conformément à l'analogie avec les arbres, les arbres de décision mettent en œuvre un processus de décision séquentiel.
En partant du nœud racine, une caractéristique est évaluée et l'un des deux nœuds (branches) est sélectionné.
Chaque nœud de l'arbre est fondamentalement une règle de décision.
Cette procédure est répétée jusqu'à ce qu'une feuille finale soit atteinte, qui représente normalement la cible.
Les arbres de décision sont également des modèles intéressants si l'on se soucie de l'interprétabilité.

Avantage :
- Interprétables. Pas besoin de norma/standardisation.
- Représentation graphique. Feature importances (importance dans la décison).
- Les arbres peuvent facilement traiter des prédicteurs qualitatifs sans avoir à créer des variables fictives.
- Régression : pas d'extrapolation possible (prédiction en dehors jeux de train)

Inconvénient :
- les arbres n'ont généralement pas le même niveau de précision prédictive que certaines des autres approches de régression et de classification.
- Surapprentisage si profond et feuille pure (100% apprentissage) : pré élagage : stopper en amont la création de l'arbre (max_depth, max_leaf_nodes, min_samples_leaf) ou post-élagage : supprimer ou regrouper les nœuds avec peu d'information). Pas très bon en généralisation



5. Modèles – Arbres de décision

Hyparamètres :

min_samples_split (nombre minimum d'échantillons qu'un nœud doit avoir avant d'être divisé),

min_sample_leaf (nombre minimum d'échantillons qu'une feuille doit avoir)

min_weight_fraction_leaf (identique à min_sample_leaf mais exprimé comme une fraction du nombre total d'instances pondérées)

max_leaf_nodes (nombre maximum de noeuds de feuilles)

max_features (nombre maximum de caractéristiques qui sont évaluées pour le fractionnement dans chaque nœud). L'augmentation du min ou la diminution du max régularise le modèle.