KNN Method

La méthode des K plus proches voisins (KPPV) ou en anglais *K-nearest neighbours* (KNN) est un algorithme d’apprentissage supervisé non-paramétrique qui permet de résoudre des problèmes de classification et de régression.

C’est un algorithme dit non-paramétrique car il se base seulement sur les données d’entraînement (à l’exception de K qui doit être fixé au préalable).

Comme évoqué précédemment, en apprentissage supervisé, le modèle est entraîné avec des données étiquetées ce qui permet à l’algorithme de classifier des données et prédire des résultats avec précision.

La méthode des K plus proches voisins se base sur une idée très simple : à partir d’un jeu de données, déterminer les plus proches voisins de l’entrée en calculant la distance de chaque point de donnée à .

Cet algorithme fait partie de la famille de modèles “d'apprentissage paresseux” car il stocke seulement un jeu de données pour effectuer ses prédictions au lieu de passer par une phase d’entraînement. Ainsi, tous les calculs sont effectués au moment de la classification ou de la prédiction. De plus, l’algorithme a fortement recours à la mémoire pour stocker les données d’entraînement.

En classification, autrement dit avec des variables discrètes, on attribue à la classe dominante parmi celles des voisins identifiés.

Quant à la régression, avec des variables continues, la valeur prédite est généralement la moyenne, la médiane ou la variance des plus proches voisins.



Pour calculer la distance, deux méthodes sont principalement utilisées :

* la distance Euclidienne
* la distance Manhattan

avec : le point observé

: le point de référence

: la dimension de l’espace

Avantages

La méthode des K plus proches voisins est très facile à mettre en œuvre car il n’est pas nécessaire de créer un modèle ou de régler des paramètres. Tout dépend du jeu de données fourni. De plus, il s’agit d’un algorithme polyvalent puisqu’il peut être utilisé pour des problèmes de classification ou de régression.

Inconvénients

Étant donné qu’il repose sur le jeu de données d’apprentissage, plus la taille du jeu de données est importante, plus l’algorithme sera coûteux en temps mais aussi en argent.

De même, le nombre de voisins est aussi un facteur à prendre en compte car il influe sur les calculs effectués par l’algorithme.

Références :

<https://www.jedha.co/formation-ia/algorithme-knn-apprentissage-supervise>

<https://ijettjournal.org/Volume-70/Issue-7/IJETT-V70I7P205.pdf>

<https://datascientest.com/knn>