Modèles KNeighbors

5. Modèles – KNeighbors

K- Nearest Neighbors est un algorithme d'apprentissage automatique supervisé car la variable cible est connue.
Non paramétrique car il ne fait pas d'hypothèse sur le modèle de distribution des données sous-jacentes.
Algorithme paresseux car KNN n'a pas d'étape de formation.
Tous les points de données ne seront utilisés qu'au moment de la prédiction.
Sans étape de formation, l'étape de prédiction est coûteuse.
Utilise la similarité des caractéristiques pour prédire le groupe dans lequel le nouveau point sera classé.

Algorithme :
Étape 1 : Choisissez une valeur pour K. K doit être un nombre impair.
Étape 2 : Trouver la distance du nouveau point par rapport à chacune des données d'apprentissage.
Étape 3 : Trouvez les K voisins les plus proches du nouveau point de données.
Étape 4 : Pour la classification, comptez le nombre de points de données dans chaque catégorie parmi les k voisins. Le nouveau point de données appartiendra à la classe qui a le plus de voisins.
Pour la régression, la valeur du nouveau point de données sera la moyenne des k voisins.

La distance peut être calculée en utilisant

- La distance euclidienne $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
- La distance de Manhattan $|x_1 - x_2| + |y_1 - y_2|$
- La distance de Hamming
- La distance de Minkowski $(|x_1 - x_2|^p + |y_1 - y_2|^p)^{1/p}$

5. Modèles – KNeighbors

Avantages :

- Algorithme simple et donc facile à interpréter la prédiction.
- Non paramétrique, donc ne fait aucune hypothèse sur le modèle de données sous-jacentes
- Utilisé à la fois pour la classification et la régression
- L'étape de formation est beaucoup plus rapide pour les plus proches voisins que pour les autres algorithmes d'apprentissage automatique, mais prédictions parfois longues si bcp features/données.

Inconvénients :

- KNN est coûteux en calcul car il recherche les plus proches voisins pour le nouveau point à l'étape de prédiction.
- Besoin élevé en mémoire car KNN doit stocker tous les points de données (pas plus de 100 features).
- L'étape de prédiction est très coûteuse
- Sensible aux valeurs aberrantes, la précision est affectée par le bruit ou les données non pertinentes.
- Sensible aux datasets sparses (jeux de données parsemés avec beaucoup de données nulles).
- Pour les petits jeux de données.

Paramètres importants :

- K : nombre de voisins (augmenter k lisse les prédictions, mais moins bons résultats apprentissage).
- Distance : mesurée entre les points, euclidienne, manhatan