ENSA
ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES
KHOURIBGA

Université Sultan Moulay Slimane

ENSA KHOURIBGA ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES

Rapport de présentation

Master Big Data et Aide à la Décision

Présenté par

Wahia Asmae
Nasser Dounia
Kassri Ilham

K-nearest neighbors (kNN)

Encadré par : Pr.OURDOU Amal

Table des matières

Introduction

L'algorithme KNN, ou k-plus proches voisins, est une méthode d'apprentissage supervisé utilisée pour la classification et la régression. Cet algorithme est basé sur la recherche des k échantillons d'entraînement les plus proches de l'échantillon de test, et sur la prise de décision en fonction de la classe majoritaire de ces k voisins. Le fonctionnement de cet algorithme est simple et intuitif, mais il est également connu pour sa robustesse et sa capacité à gérer des données non linéaires et bruyantes.
L'algorithme KNN a été introduit pour la première fois en 1951 par Fix et Hodges, mais il a été popularisé dans les années 1960 par le statisticien Edward L. Hart. Depuis lors, l'algorithme a été largement utilisé dans divers domaines, notamment la reconnaissance de caractères, la classification d'images, la prédiction de la qualité du vin, la recommandation de films, la détection de fraudes, etc.
Dans ce rapport, nous allons explorer en détail l'algorithme KNN, en expliquant son fonctionnement, ses avantages et inconvénients et ses applications. Nous conclurons en discutant des perspectives d'avenir de cet algorithme et de ses améliorations possibles.

1 Calcul des distances

1.1 La distance euclidienne

Distance qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points :

$$D_e(x, y) = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2}$$

1.2 Distance Manhattan

La distance de Manhattan : calcule la somme des valeurs absolues des différences entre les coordonnées de deux points :

$$D_m(x, y) = \sum_{i=1}^{k} \mid x_i - y_i \mid$$

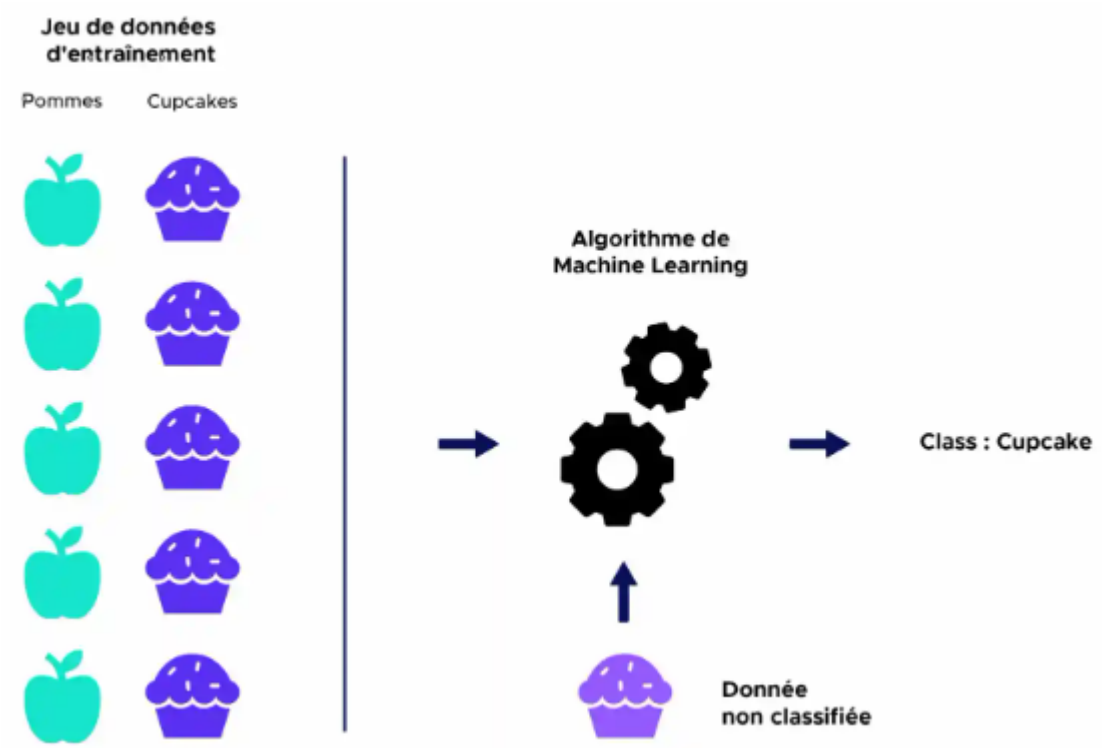
1.3 Distance Hamming

la distance entre deux points données est la différence maximale entre leurs coordonnées sur une dimension.

$$D_h(x, y) = \sum_{i=1}^{k} \mid x_i - y_i \mid$$

- $x = y \Rightarrow D = 0$
- $x \neq y \Rightarrow D = 0$

2 KNN fonctionnement de l'algorithme

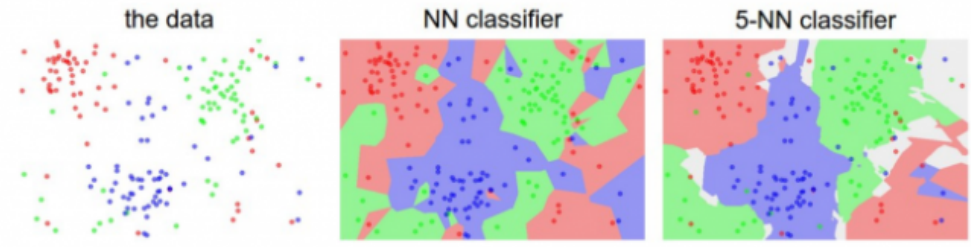


L'algorithme KNN fonctionne sur la base d'une approche d'apprentissage supervisé, où les données d'entrée sont étiquetées avec une classe connue ou une valeur cible. Son fonctionnement consiste à trouver les k points les plus proches (les k voisins) d'un point inconnu, en se basant sur une mesure de distance entre les différents points.
La distance est généralement calculée selon la distance euclidienne, qui correspond à la distance entre deux points dans un espace n-dimensionnel. Pour trouver les k voisins les plus proches, on calcule la distance entre le point inconnu et tous les autres points de l'ensemble de données, puis on sélectionne les k points ayant la distance la plus courte.
Une fois les k voisins sélectionnés, la classification ou la régression est effectuée en prenant la classe majoritaire pour la classification ou la moyenne des valeurs cibles pour la régression. Le choix de la valeur de k est un paramètre important

de l'algorithme, car elle peut affecter significativement les performances de l'algorithme.
En cas de données manquantes, plusieurs approches peuvent être utilisées pour gérer ces données manquantes, comme la suppression des données manquantes ou la substitution des valeurs manquantes par la valeur moyenne ou la valeur médiane.
Pour illustrer le fonctionnement de l'algorithme, prenons l'exemple de la classification d'images. Si nous avons une image inconnue, l'algorithme KNN va chercher les k images les plus proches dans notre ensemble de données étiquetées, et va prédire que l'image inconnue appartient à la classe majoritaire des k voisins.
En résume les etapes de KNN est :
Étape 1 : Sélectionnez le nombre K de voisins
Étape 2 : Calculer la distance euclidienne du nombre K de voisins
Étape 3 : Prenez les K voisins les plus proches selon la distance euclidienne calculée.
Étape 4 : Parmi ces k voisins, comptez le nombre de points de données dans chaque catégorie.
Étape 5 : Attribuez les nouveaux points de données à la catégorie pour laquelle le nombre de voisins est maximal.
Étape 6 : Notre modèle est prêt.

3 Choix de la valeur de K

Le choix de la valeur K à utiliser pour effectuer une prédiction avec K-NN, varie en fonction du jeu de données. En règle générale, moins on utilisera de voisins (un nombre K petit) plus on sera sujette au sous apprentissage (underfitting). Par ailleurs, plus on utilise de voisins (un nombre K grand) plus, sera fiable dans notre prédiction. Toutefois, si on utilise K nombre de voisins avec K=N et N étant le nombre d'observations, on risque d'avoir du overfitting et par conséquent un modèle qui se généralise mal sur des observations qu'il n'a pas encore vu.



L'image ci-dessus à gauche représente des points dans un plan 2D avec trois types d'étiquetages possibles (rouge, vert, bleu). Pour le 5-NN classifieur, les limites entre chaque région sont assez lisses et régulières. Quant au N-NN Classifier,

on remarque que les limites sont “chaotiques” et irrégulières. Cette dernière provient du fait que l'algorithme tente de faire rentrer tous les points bleus dans les régions bleues, les rouges avec les rouges etc... c'est un cas d'overfitting.
Pour cet exemple, on préférera le 5-NN classifier sur le NN-Classifier. Car le 5-NN classifier se généralise mieux que son opposant.

4 Les avantages et inconvénients de l'algorithme KNN

Les avantages de l'algorithme KNN comprennent :

1. Simplicité : KNN est un algorithme simple à comprendre et à mettre en œuvre, sans besoin d'une connaissance préalable de la distribution des données.
2. Flexibilité : L'algorithme KNN peut être utilisé pour des tâches de classification et de régression, et il peut fonctionner avec différents types de données tels que les données numériques, textuelles, et même les images.
3. Bonne précision : KNN a tendance à bien fonctionner lorsque les données sont bien structurées et les frontières de décision sont clairement définies.
4. Interprétabilité : Les résultats de KNN sont facilement interprétables et peuvent aider à comprendre les données.

Cependant, l'algorithme KNN présente également certains inconvénients, notamment :

1. Temps de calcul : L'algorithme KNN peut être très lent sur de grandes quantités de données, car il nécessite de calculer la distance entre chaque point de données.
2. Problème de choix de la valeur de K : Il est difficile de déterminer la meilleure valeur pour K, qui peut influencer significativement les résultats.
3. Sur-apprentissage : KNN peut être sensible au sur-apprentissage, en particulier lorsque la valeur de K est faible et que le nombre de voisins est petit.
4. Sensibilité aux données bruitées : L'algorithme KNN peut être sensible aux données bruitées ou aberrantes, car il affecte les distances entre les points de données.

5 Les applications de l'algorithme KNN

L'algorithme KNN est largement utilisé dans une variété de domaines, notamment :

- Classification d'images : KNN peut être utilisé pour classer des images en fonction de leurs caractéristiques, telles que la couleur, la texture, la forme, etc.
- Reconnaissance de caractères : KNN peut être utilisé pour reconnaître des caractères manuscrits ou imprimés en fonction de leurs caractéristiques.
- Diagnostic médical : KNN peut être utilisé pour diagnostiquer certaines maladies en fonction des symptômes et des caractéristiques des patients.
- Prédiction de la bourse : KNN peut être utilisé pour prédire les mouvements du marché boursier en fonction de l'historique des prix.
- Prédiction de la qualité de l'air : KNN peut être utilisé pour prédire la qualité de l'air en fonction des caractéristiques telles que la température, l'humidité, la pollution, etc.
- Reconnaissance de la parole : KNN peut être utilisé pour reconnaître la parole en fonction des caractéristiques du son.
- Détection de spam : KNN peut être utilisé pour détecter les emails indésirables ou de spam en fonction du contenu et des caractéristiques de l'email.

Ces exemples montrent la polyvalence de l'algorithme KNN et sa capacité à être appliqué à une grande variété de domaines.

Conclusion et perspectives

La conclusion de notre étude de l'algorithme KNN est que c'est un outil très puissant pour la classification et la régression, surtout pour les ensembles de données avec des dimensions élevées. L'un des avantages de l'algorithme KNN est qu'il ne nécessite pas d'hypothèses sur la distribution des données et peut être utilisé pour des ensembles de données non linéaires.
Cependant, l'algorithme KNN a également des inconvénients, tels que la sensibilité au bruit, la nécessité de choisir une valeur appropriée de K et le coût élevé de stockage et de calcul pour les grands ensembles de données.
Les applications de l'algorithme KNN sont nombreuses et variées, notamment la classification d'images, la reconnaissance de caractères, la prédiction de la qualité du vin, la recommandation de films, la détection de fraudes, etc.
En termes de perspectives d'avenir, il y a des améliorations possibles de l'algorithme KNN, comme la prise en compte de l'importance différente des voisins en fonction de leur distance, la combinaison avec d'autres algorithmes pour améliorer la précision, etc. De plus, l'algorithme KNN peut être utilisé dans de nouveaux domaines, tels que la médecine personnalisée, la reconnaissance de la parole et la prédiction de l'évolution des maladies.
En conclusion, l'algorithme KNN est un outil puissant pour la classification et la régression, avec des avantages et des inconvénients à prendre en compte. Ses applications sont variées et ses perspectives d'avenir sont prometteuses.

References

[1] Chat GPT
[2] https ://datascientest.com/knn
[3] https ://mrmint.fr/introduction-k-nearest-neighbors