Article KNN

Antoine SERREAU / Corentin BRETONNIERE / Benjamin GUIGON

Janvier 2020

Ce travail est réalisé dans le cadre d'un projet collaboratif comprenant 3 étudiants du MSc Data Management de PSB :
Antoine SERREAU / Corentin BRETONNIERE / Benjamin GUIGON.
Nous avons réalisés 2 autres travaux portant sur :
ACP et Les Arbres de Décisions

1 Introduction

Vous vous aventurez dans l'apprentissage du machine learning ? Voici une rapide introduction à l'un des algorithmes du Machine Learning - KNN - qui vous aidera à en saisir les principales dynamiques et notions mathématiques.

L'algorithme K-Nearest Neighbors (ou KNN) est l'un des algorithmes d'apprentissage les plus utilisés en raison de sa simplicité. Alors, qu'est-ce que c'est ?

KNN est un algorithme d'apprentissage non paramétrique. Il utilise des données avec plusieurs classes pour prédire la classification du nouveau point d'échantillonnage. KNN est non-paramétrique car il ne fait aucune hypothèse sur les données étudiées, c'est-à-dire que le modèle est distribué à partir des données.

KNN n'utilise pas les points de données d'entraînement pour faire des généralisations.Il y a donc peu ou pas de phase d'apprentissage explicite, la phase d'apprentissage est assez rapide, KNN conserve toutes les données relatives à l'apprentissage puisqu'elles sont nécessaires pendant la phase de test. Etant donné que la plupart des données n'obéissent pas aux hypothèses théoriques typiques, comme lorsque nous considérons un modèle de régression linéaire, ce qui rend KNN crucial lorsque l'on étudie des données avec peu ou pas de connaissances préalables.

2 Quand et Pourquoi utiliser KNN ?

KNN peut être utilisé dans les problèmes de régression et de classification prédictive. Cependant, en ce qui concerne les problèmes industriels, il est surtout utilisé dans la classification car il se retrouve dans tous les paramètres évalués lors de la détermination de l'utilisabilité d'une technique.

- Pouvoir de prédiction
- Temps de calcul
- Facilité d'interprétation des résultats

L'algorithme KNN est utilisé en raison de sa facilité d'interprétation et de son faible temps de calcul.

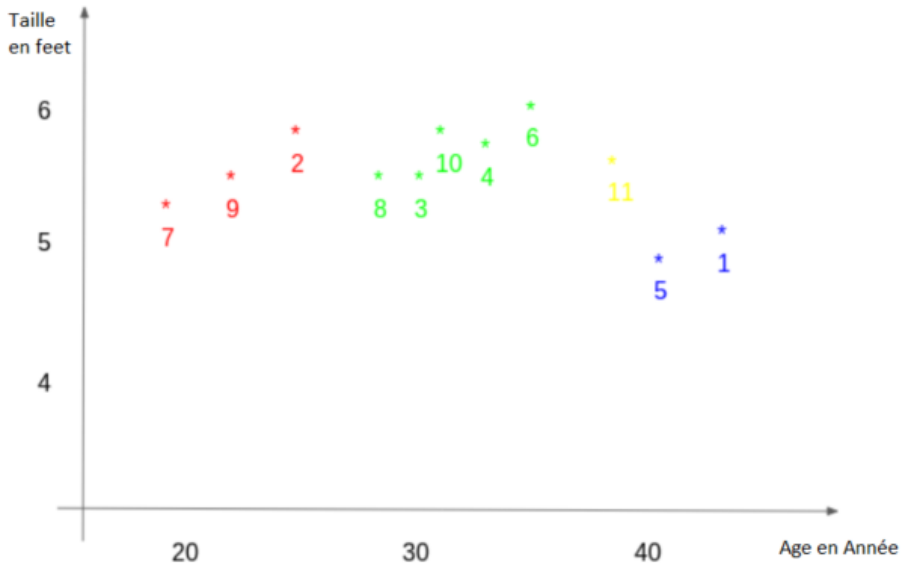
Par exemple, KNN peut être utilisé dans un système bancaire pour prédire si une personne est apte à obtenir un prêt. Ou si elle présente des caractéristiques similaires à celles d'un emprunteur défaillant. Calcul de la cote de crédit - KNN peut aider à calculer la cote de crédit d'une personne en la comparant avec celle de personnes ayant des caractéristiques similaires. L'algorithme KNN est également utilisé pour la reconnaissance vidéo, la reconnaissance d'images, la détection de l'écriture manuscrite et la reconnaissance vocale.

3 Principe de "base" de l'algorithme

Le principe de cet algorithme est de choisir les k données les plus proches du point étudié afin d'en prédire sa valeur. Commençons par un exemple simple. Considérons le tableau suivant - il se compose de la taille, de l'âge et du poids (valeur cible) pour 10 personnes. Comme vous pouvez le voir, la valeur du poids de ID : 11 est manquante. Nous devons prévoir le poids de cette personne en fonction de sa taille et de son âge.

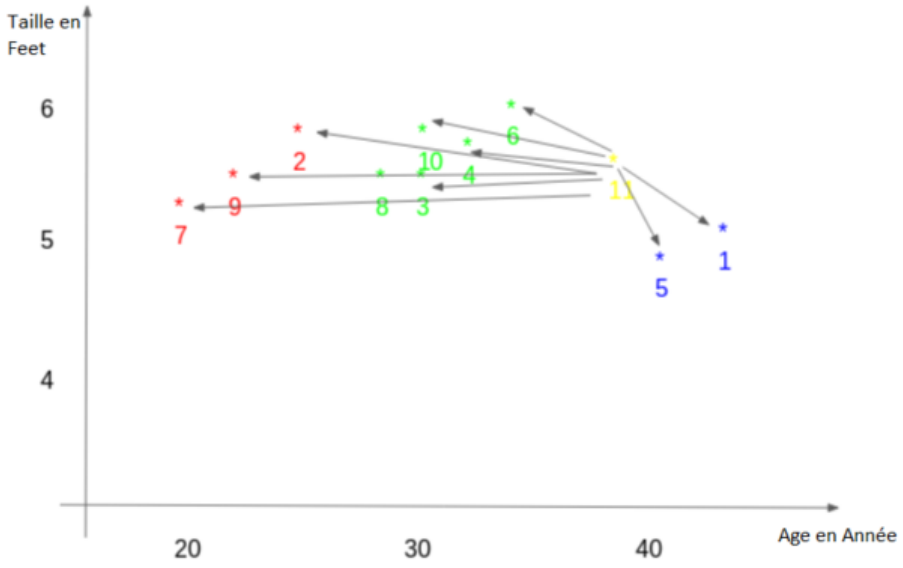
ID	Taille (feet)	Age	Poids (kg)
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	?

Pour une meilleure compréhension, voici le graphique de la taille en fonction de l'âge du tableau ci-dessus :



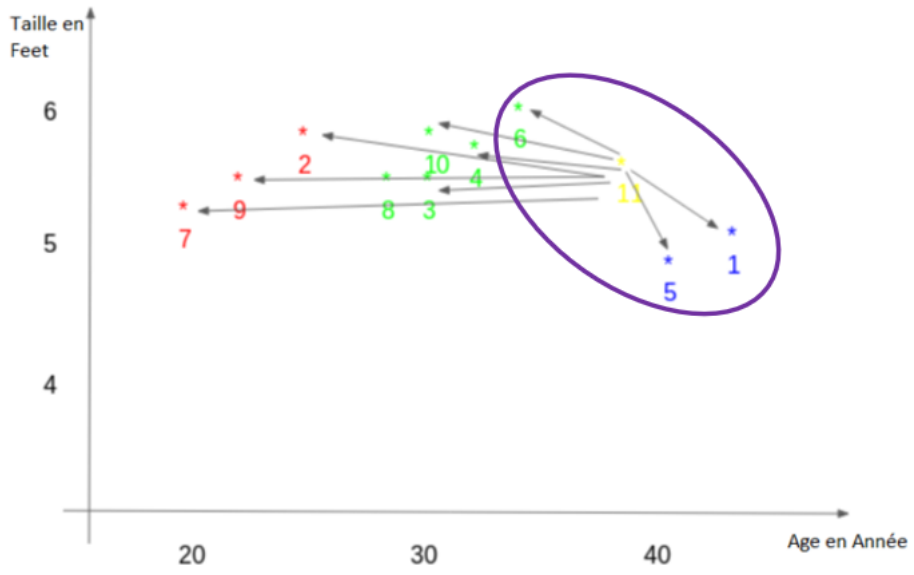
Dans le graphique ci-dessus, l'axe des y représente la taille d'une personne (en feet) et l'axe des x représente l'âge (en années). Les points sont numérotés en fonction des valeurs d'identification. Le point jaune (ID 11) est notre point de test.
Si je vous demande d'identifier le poids de ID11 en fonction du tracé, quelle serait votre réponse ? Vous répondriez probablement que puisque ID : 11 est plus proche des points 5 et 1 (ici en bleu), il doit donc y avoir un poids

similaire à ces ID, probablement entre 72 et 77 kg (poids de ID : 1 et ID : 5 du tableau). C'est logique, mais comment pensez-vous que l'algorithme prédit les valeurs ? Nous le découvrirons dans cet article.
Comme nous l'avons vu ci-dessus, l'algorithme KNN peut être utilisé pour les problèmes de classification et de régression. L'algorithme KNN utilise la "similarité des caractéristiques" pour prédire les valeurs de tout nouveau point de données. Cela signifie qu'une valeur est attribuée au nouveau point en fonction de sa ressemblance avec les points de l'ensemble de formation. D'après notre exemple, nous savons que le point ID : 11 a une taille et un âge similaires à ceux des points ID : 1 et ID : 5, donc le poids serait aussi approximativement le même.
S'il s'agissait d'un problème de classification, nous aurions pris le mode comme prédiction finale. Dans ce cas, nous avons deux valeurs de poids : 72 et 77. Avez-vous une idée de la manière dont la valeur finale sera calculée ? C'est la moyenne des valeurs est considérée comme la prédiction finale.
Vous trouverez ci-dessous une explication par étapes :
Tout d'abord, la distance entre le nouveau point et chaque point d'apprentissage est calculée. Nous verrons par la suite qu'il est possible de calculer la distance de différentes façon.



Dans un second temps, les k points de données les plus proches sont sélectionnés (en fonction de la distance). Dans cet exemple, les points 1, 5, 6 seront sélectionnés si la valeur de k est de 3. Nous explorerons plus loin la méthode

pour sélectionner la bonne valeur de k plus loin dans cet article.



La moyenne de ces points de données est la prévision finale pour le nouveau point. Ici, nous avons le poids de ID : 11 = (77+72+60)/3 = 69,66 kg.

Voici d'une manière très "généralisée", le fonctionnement de l'algorithhme KNN.

4 Principe de "mathématique" de l'algorithme