AP10 ALGORITHME K PLUS PROCHES VOISINS (KNN)

Définition 1 : Algorithme des k plus proches voisins

Abrégé en **KNN** (K-nearest neighbors), cet algorithme cherche à prédire la classe (ou propriété) d'un élément en fonction de la classe majoritaire de ses k plus proches voisins.

Exemple

Par exemple, imaginons qu'un pixel soit manquant dans une image. L'algorithme va prédire la couleur de ce pixel en fonction de la couleur des pixels voisins.

L'algorithme KNN est un algorithme d'**apprentissage** (Machine Learning) supervisé. C'est-à-dire qu'à partir d'un ensemble des données étiquetées, il va pouvoir prédire l'étiquette d'une donnée inconnue.

Par exemple, à partir d'un ensemble d'images de chiens et de chats, ce type d'algorithme permet de déterminer si une photo inconnue est celle d'un chien ou d'un chat.

Un autre exemple sont les systèmes de recommandation. Un site marchand peut chercher vos `k` plus proches voisins en termes d'âge, de lieu, etc., pour vous proposer des articles susceptibles de vous plaire.

I. Principe

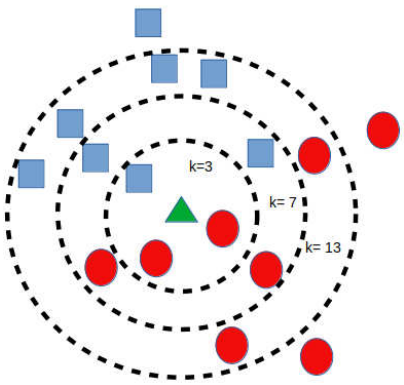
Tout d'abord, il nous faut avoir un ensemble de données d'apprentissage. Chaque donnée :

- appartient à une **classe**.
- est caractérisée par un certain nombre d'**attributs**.

Nous avons un individu non identifié, c'est à dire de **classe inconnue**, dont on possède cependant les caractéristiques. L'objectif est de classifier cet individu en regardant ses k plus proches voisins.

II. Exemple

Voici la représentation de deux classes : les carrés et les disques. En représentant leurs caractéristiques graphiquement, on obtient la répartition ci-dessous. Un individus de classe inconnue possédant ces mêmes caractéristiques, est représenté par le triangle.



- En prenant k=3, les 3 plus proches voisins sont 2 disques et un carré : on décidera que l'inconnu est un disque.
- En prenant k=7, les 7 plus proches voisins sont 4 disques et 3 carrés : on décidera encore que l'inconnu est un disque.
- En prenant k=13, les 13 plus proches voisins sont 6 disques et 7 carrés : on décidera alors que l'inconnu est un carré.

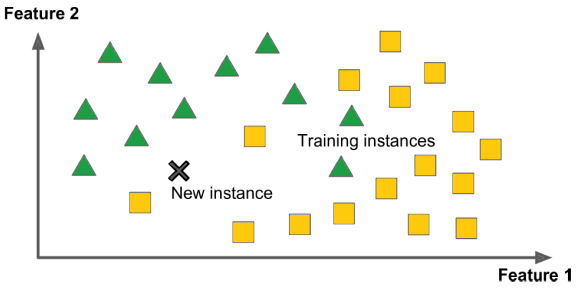
Remarque

L' algorithme k-NN est un algorithme de décision qui pourrait être résumé par la phrase « dis moi qui sont des voisins et je te dirai qui tu es ».

- En prenant une décision, on fait un choix éclairé mais on ne peut pas être à 100% certain d'avoir fait le bon :
- La définition de la **distance** entre deux classes est particulièrement importante.
- Le choix du k, c'est à dire du nombre de voisins à étudier, est également primordial.

III. Activité 1

On considère un ensemble d'objets dans le plan de deux types différents, des triangles verts et des carrés jaunes.
On ajoute un nouvel objet (la croix que l'on appelera *N*) et on souhaite déterminer à quel type d'objet il appartient.



Plus proche voisin k = 3

1. sur la figure précédente, sélectionnez les 3 plus proches voisins du point N. .....
2. quel type choisir alors pour le point N? .....

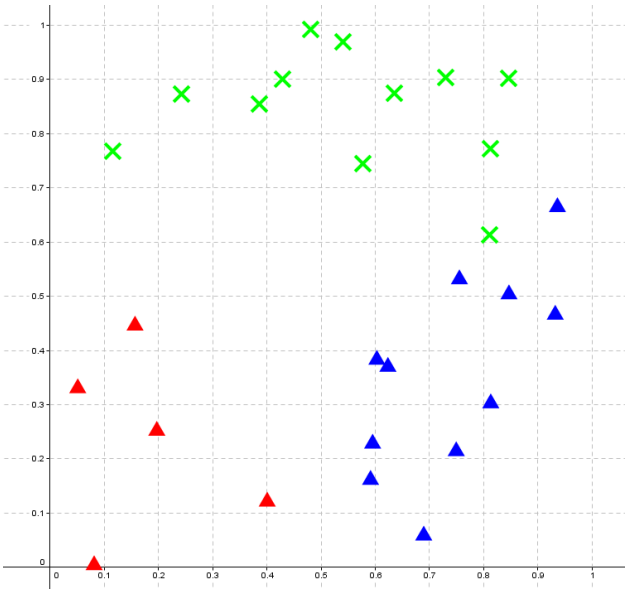
Plus proche voisin k = 5

1. sur la figure précédente, sélectionnez les 5 plus proches voisins du point N. .....
2. quel type choisir alors pour le point N? .....

IV. Activité 2

On considère la figure ci-dessous avec un nouvel ensemble de points dans le plan de trois classes différentes, et N un nouveau point que l'on souhaite colorier.

1. Choisir un point aléatoirement et le placer dans le graphique
2. Déterminer, à l'aide de l'algorithme des k plus proches voisins, la classification de votre nouveau point N pour k=1, k=3 et k=5.



Choix du k

Le choix du k est important. En effet, si k est trop petit, on risque de faire du bruit dans les données (on peut être influencé par un voisin qui n'est pas représentatif). Si k est trop grand, on risque d'inclure des voisins qui ne sont pas pertinents pour la classification de l'individu.