

1 Idée générale

L'**algorithme des k plus proches voisins** (ou **knn** pour k nearest neighbours) est une **méthode d'apprentissage supervisé**. Il permet de faire une **classification** de données.

On peut dire que cet algorithme repose sur le dicton populaire « qui se ressemble, s'assemble ! ».

Prenons un exemple permettant de bien sentir l'idée de cet algorithme.

On remarque que dans la cour de récréation en Primaire, les jeunes enfants ont tendance à se regrouper entre filles ou entre garçons, mais assez peu à se mélanger.

En considérant ceci, on peut essayer de prédire le sexe d'un enfant-test (de genre inconnu) simplement en observant quel est le genre majoritaire (filles ou garçons) de ses plus proches voisins.

Voir l'animation *Geogebra* en annexe (<https://youtu.be/JPPY1Lzch6A>).

De façon plus générale, **l'algorithme doit prédire la classe d'un élément en fonction de la classe majoritaire de ses k plus proches voisins**.

Remarque 1 : le nombre de classes n'est pas limité à deux comme dans l'exemple fille/garçon.

Remarque 2 : dans la pratique, la notion de proximité (proches voisins) n'est pas forcément liée à une distance physique (combien de mètres ?) entre l'élément test et les autres éléments des données de travail, mais plutôt à un « qualificatif de similarité » entre les éléments (quelques exemples possibles : couleur des yeux, âge, production de pétrole, PIB, etc.)

2 Descriptif des données

La mise en œuvre des algorithmes d'apprentissage nécessite d'avoir une banque de données de travail. Ici, les données contiennent deux types d'informations :

- des **caractéristiques numériques destinées à la comparaison** de deux éléments ;
- un **critère destiné à la classification** d'un élément.

Les caractéristiques des éléments doivent permettre de définir une « distance » séparant les éléments.

Enfin, l'algorithme repose sur un **entier k précisant le nombre de voisins à considérer** pour établir la classe de l'élément non étiqueté (de classe inconnue).

3 Fonction distance

Nous limiterons notre étude à 2 caractéristiques pour chaque élément, ce qui permettra de représenter les éléments dans un repère orthonormé du plan (abscisse et ordonnée sont les 2 caractéristiques).

Nous utiliserons la distance euclidienne entre ces points du plan pour définir « la proximité » entre deux éléments.

Formule de la distance :

Dans un repère orthonormé du plan, on définit les points $A(x_A; y_A)$ et $B(x_B; y_B)$.
La longueur AB vaut :

$$AB = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

4 Le principe de l'algorithme

Un ensemble de n données est fourni, chacune de ces données contient deux données numériques $(x_i; y_i)$ et une donnée supplémentaire correspondant à une classe c_i .

Un nouvel élément N étant donné avec ses valeurs numériques $(x_N; y_N)$, on cherche à la classe c_N inconnue à laquelle il pourrait appartenir.

Voici les différentes étapes permettant d'apporter une réponse :

- Parcourir l'ensemble des n données de travail et calculer les distances entre $(x_i; y_i)$ et $(x_N; y_N)$ pour chaque élément i ;
- Classer ces résultats par distance croissante.
- Extraire de la liste les k plus proches éléments (correspondant aux plus petites distances précédentes).
- Dans ces k éléments, compter le nombre d'occurrences de chaque classe présente.
- Attribuer au nouvel élément N la classe la plus fréquente. C'est tout !

5 Complément : Préparation des données pour l'apprentissage

Lorsqu'on travaille sur des algorithmes d'apprentissage, on doit toujours disposer d'une grande quantité de données pour « entraîner » l'algorithme.

On commence par extraire une **grande partie des données pour la phase d'apprentissage** et on conserve la **petite partie restante pour faire des tests** et valider la qualité de l'algorithme.

Ensuite on peut réellement travailler sur des données inconnues pour faire des prédictions.

Info de niveau supérieur pour les curieux : <https://realpython.com/knn-python/> : Ce site donne plus d'informations générales sur l'apprentissage machine et la mise en œuvre pratique de l'algorithme des knn, avec ses points forts et ses points faibles.