Algorithme des k plus proches voisins

L'algorithme des **k plus proches voisins** ou **k-NN** (k Nearest Neightbours) est un algorithme d'**apprentissage automatique** (machine learning).

L'idée est d'**utiliser un grand nombre de données** afin "d'apprendre à la machine" à résoudre un certain type de problème, pour apporter **une réponse plausible**, mais pas nécessairement exacte, à un **problème de classification**.

Les données sont constituées de **descripteurs** et d'**une étiquette** (label).

KNN

Pour **prédire l'étiquette** d'une donnée x dont on ne connait que les descripteurs, on utilise un jeu de **données d'apprentissage** et on choisi une **notion de distance** adaptée à la nature des éléments observés.

Le principe de l'**algorithme des k plus proches voisins** :
- on calcule les distances entre la nouvelle donnée x et chaque donnée appartenant au jeu d'apprentissage ;
- on retient les k données du jeu d'apprentissage les plus proches de x ;
- on attribue à x l'étiquette majoritaire parmi les k données les plus proches.

Remarques :

★ L'efficacité de cet algorithme dépend :
- du choix pertinent du calcul de **distance** entre les données ;
- du **nombre k** de voisins choisis ;
- de la **qualité** et de la **quantité des données** d'apprentissage.

★ Il existe des méthodes qui permettent de déterminer le **taux d'erreur de l'algorithme** et de **choisir la valeur de k** permettant la prédiction la plus fiable.

L'**apprentissage automatique** ou machine learning, dont l'algorithme des k plus proches voisins n'est qu'un exemple, est **très utilisé depuis quelques années** car il dépend de la qualité et de la quantité des données qui permettront à la machine d'apprendre à résoudre le problème étudié.

Or, avec le développement d'internet, **on a maintenant accès à un grand nombre de données** (les "**big data**").

C'est pourquoi il est devenu important pour toutes sortes d'entreprises ou de gouvernements d'accéder à un maximum de données, notamment sur les utilisateurs d'internet, afin de faire leurs propres prédictions...

Exercice à l'école des sorciers :

À l'entrée à l'école de Poudlard, le Choixpeau magique répartit les élèves dans les différentes maisons (Gryffondor, Serpentard, Serdaigle et Poufsouffle) en fonction de leur courage, leur loyauté, leur sagesse et leur malice.

Le Choixpeau magique dispose d'un fichier CSV (**choixpeauMagique.csv**) dans lequel sont répertoriées les données d'un échantillon d'élèves.

Un nouvel élève arrive, il n'a pas encore de maison et on cherche à lui en attribuer une !

1. Créez une fonction **charger_table()** permettant de créer une liste de dictionnaires **eleves_poudlard**, qui contiendra les données du fichier **choixpeauMagique.csv**.

   Chaque dictionnaire de la liste représentera un élève de Poudlard, avec pour clés le nom des champs du fichier csv.

   Par exemple :

   {"Nom" : "Adrian", "Courage" : 9, "Loyauté" : 4, "Sagesse" : 7, "Malice" : 10, "Maison" : "Serpentar"}.

2. Créez une fonction **distance()** permettant de calculer la distance entre deux élèves de l'école, décrit par leurs dictionnaires.

   Pour ce calcul, on utilisera la **distance de Manhattan**, dont on rappelle la formule ci-dessous :

   pour deux couples de réels (w_1, x_1, y_1, z_1) et (w_2, x_2, y_2, z_2), la distance est égale à

   $|w_1 - w_2| + |x_1 - x_2| + |y_1 - y_2| + |z_1 - z_2|$ où l'on prend la valeur absolue de l'écart entre deux valeurs.

   Exemple : si $x = (9; 4; 7; 10)$ et $y = (8; 6; 6; 6)$ alors la distance entre x et y est $d(x, y) = 1 + 2 + 1 + 4 = 8$.

3. Créez la fonction **proches_voisins()** qui retourne la liste des maisons des 7 voisins les plus proches d'un élève donné.

   Remarque : Vous pouvez créer une liste de tuples comportant la distance au nouvel élève et la maison de l'élève de Poudlard, puis trier la liste dans l'ordre croissant des distances avec la fonction **sorted()**, et ne garder que les 7 premiers élèves.

4. Créez une fonction **prediction_maison()** qui pour un élève donné, indiquera la maison à laquelle il devrait être affecté.

5. Voici quatre nouveaux élèves.

   A quelles maisons devraient-ils être affecté par le Choixpeau ?

Hermione	8	8	8	6
Drago	6	6	5	8
Cho	6	6	9	6
Cédric	7	10	5	6