Algorithme des k plus proches voisins
Correction des exercices

Exercice 1

On considère le fichier csv dont le contenu est représenté ci-contre. Il contient les données de 10 points de l'espace colorés selon une certaine logique spatiale.

```
x,y,z,couleur
3,7,5,noir
4,6,2,noir
3,7,8,blanc
0,1,2,noir
1,0,7,blanc
5,4,4,blanc
9,1,2,noir
5,3,3,noir
1,1,4,blanc
3,3,7,blanc
```

On considère un onzième point A de coordonnées x=4, y=4 et z=5 dont on cherche à prédire la couleur en appliquant l'algorithme des k plus proches voisins.

1. Combien de descripteurs vont être utilisés pour effectuer la prédiction ? Quels sont-ils ? 3 descripteurs : x, y et z

2. Combien y a-t-il d'étiquettes différentes ? Quelles sont-elles ? il y a deux étiquettes différentes : blanc et noir

3. Après un import de ce fichier dans Python sous forme de table, quel serait le dictionnaire associé à la ligne 4 du fichier (3,7,8,blanc) ?
   `{'x':3, 'y':7, 'z':8, 'couleur':'blanc'}` ou
   `{'x':'3', 'y':'7', 'z':'7', 'couleur':'blanc'}`

4. On a obtenu le tableau incomplet des distances ci-contre qui donne les distances entre le point A(4;4;5) et les points de la table. Calculer les deux distances manquantes.
   $\sqrt{(4-0)^2} + (4-1)^2 + (5-2)^2 \approx 5{,}83$
   $\sqrt{(4-9)^2} + (4-1)^2 + (5-2)^2 \approx 6{,}56$

point de la table	distance au point A
3,7,5,noir	3.16
4,6,2,noir	3.60
3,7,8,blanc	4.36
0,1,2,noir	5.83
1,0,7,blanc	5.39
5,4,4,blanc	1.41
9,1,2,noir	6.56
5,3,3,noir	2.45
1,1,4,blanc	4.36
3,3,7,blanc	2.45

5. On applique l'algorithme des k plus proches voisins pour prédire la couleur au point A.

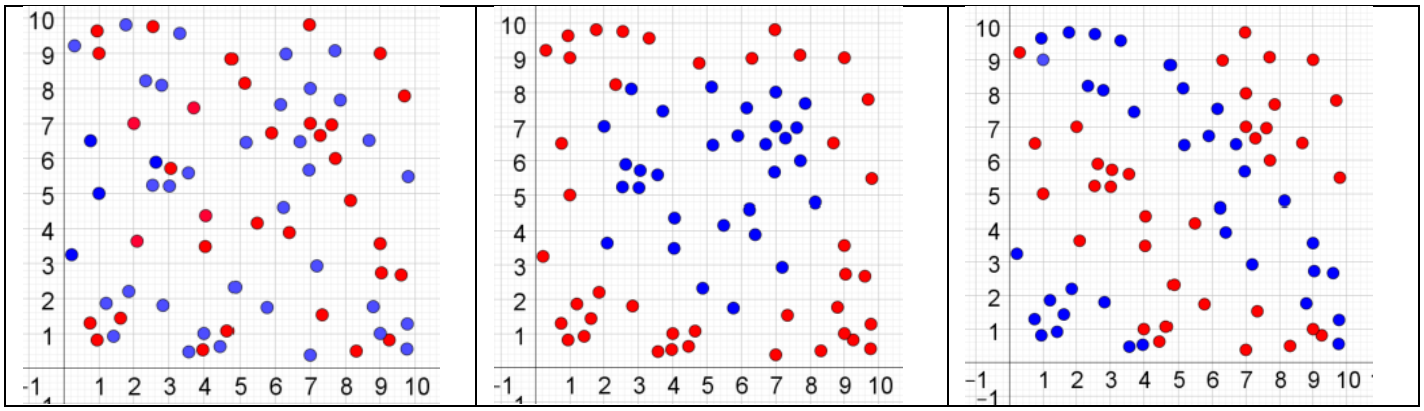
   - Si k = 1, quelle prédiction obtient-on ? blanc (distance 1.41)
   - Si k = 3, quelle prédiction obtient-on ? blanc (distances 1.41, 2.45 et 2.45)
   - Si k = 5, quelle prédiction obtient-on ? noir (distances 1.41, 2.45, 2.45, 3.16 et 3.60)

6. Pourquoi ne prend-on pas une valeur paire pour k ? pour éviter les ex-aequo

Exercice 2

On considère trois jeux de données différents pour lesquels on a deux descripteurs sur lesquels baser la prédiction. La prédiction porte sur une couleur (étiquette rouge ou étiquette bleue). On a représenté les 3 jeux de données ci-dessous.
Lequel ou lesquels de ces 3 jeux de données ne vont pas permettre de réaliser des prédictions fiables ?
Celui le plus à gauche car il ne semble pas y avoir de zone 'réservée aux rouges' et de zone 'réservée aux bleus'. Tout est mélangé. Il n'y a donc pas l'air d'y avoir de lien entre l'étiquette d'un point et l'étiquette de ses proches voisins. Il parait donc peu pertinent d'appliquer l'algorithme des k plus

proches voisins sur un tel jeu de données (pour en être certain il faudrait 10 ou 20 fois plus de points dans le jeu de données : peut-être verrait on des zones se former ?).



Exercice 3

On applique l'algorithme des k-plus proches voisins sur une table table_donnees dont deux enregistrements sont par exemple :

```
#1 {'nb_clics': 0.489,  'duree': 0.517, 'panier': 0.854, 'satisfait': 'O'}
#2 {'nb_clics': 0.828,  'duree': 0.865, 'panier': 0.142, 'satisfait': 'N'}
```

Les valeurs associées aux différents descripteurs ont toutes été ramenées entre 0 et 1 afin de leur donner la même importance lors du calcul de distance.

1) D'après vous, de quel type d'étude provient cette table de données :
   - ~~Etude de satisfaction sur les clients d'un fabricant de paniers en plastique,~~
   - Etude de satisfaction sur les clients d'un site web commercial.

2) Pour un nouveau client, on dispose de valeurs associées aux trois descripteurs dans un dictionnaire : `{'nb_clics' : 0.632, 'duree':0.321, 'panier': 0.242}`.
   On souhaite prédire la valeur d'un champ `'satisfait'` de ce client.

   a) En termes de vocabulaire, quelles sont les deux affirmations correctes ?
   - On cherche à effectuer une classification dans deux classes : 'O' et 'N'
   - On cherche à étiqueter le client en utilisant deux étiquettes : 'O' et 'N'
   - ~~On cherche à distancer le client en utilisant trois descripteurs.~~

   b) Calculer la distance entre ce nouveau client et l'enregistrement #1.

$$\sqrt{(0.632-0.489)^2+(0{,}321-0{,}517)^2+(0{,}242-0{,}854)^2} \approx 0.658$$

   c) Ecrire le code d'une fonction python qui prend en paramètres :
   - un enregistrement de la table client_table,
   - un enregistrement client_nouveau,

   et renvoie la distance entre ces deux enregistrements (on utilisera `math.sqrt` du module `math` pour calculer la racine carrée).

```
def distance_clients(client_a, client_b) :
      d = ( ( client_b['nb_clics'] - client_a['nb_clics'] )**2 +
            ( client_b['duree'] - client_a['duree'] )**2 +
            ( client_b['panier'] - client_a['panier'] )**2 )
      return math.sqrt(d)
```