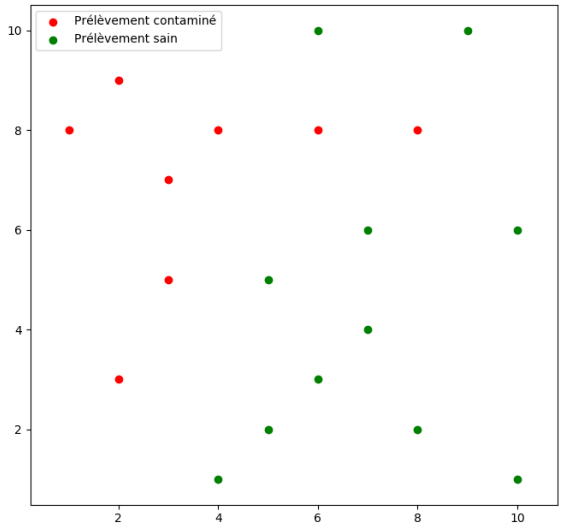
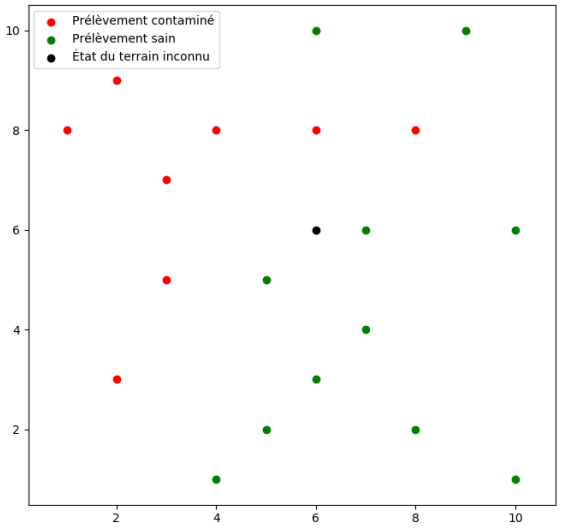
Algorithmique : algorithme des k plus proches voisins

Contexte

Imaginons la situation suivante : il y a eu une fuite de produits chimiques dans un champ, et une partie du terrain a été contaminé et est devenu impropre à la culture. On a fait venir une équipe de scientifiques qui ont fait plusieurs prélèvements sur le terrain, et pour chacun des prélèvements ils ont indiqué si la zone était contaminée (zone rouge) ou saine (zone verte). Les scientifiques fournissent des données **annotées** (qu'on pourra appeler les données **d'apprentissage**, ou les données **connues**).



(a) Prélèvements des scientifiques



(b) Point à classer

La question est la suivante : l'agriculteur souhaite utiliser la parcelle de son terrain située aux coordonnées $(6, 6)$. Faire revenir les scientifiques est très couteux et n'est pas une solution acceptable, il doit donc trouver un moyen d'exploiter les données à sa disposition.

Vocabulaire : on appelle ce type de problème un problème de **classification** : il s'agit de classer dans une catégorie (contaminé/sain) une nouvelle donnée. La méthode se base sur un **apprentissage supervisé** : on connait déjà avec certitude la classe d'un certains nombre de données, et on veut déduire à partir de ces informations la classe de la nouvelle donnée.

1 Algorithme des k plus proches voisins

Définition 1. C'est un algorithme d'apprentissage supervisé qui peut permettre la classification de données. On appelle X le point que l'on cherche à classer :

1. calculer la distance entre X et chacunes des données annotées ;
2. trouver les k données annotées dont la distance à X est la plus faible (les k plus proches voisins)
3. renvoyer la classe majoritaire parmi les k plus proches voisins.

Exemple 1. Reprennons l'exemple de la figure 1b. On a utilisé l'algorithme des 3 plus proches voisins. On rappelle la formule de la distance (euclidienne) entre deux points $A(x_A; y_A)$ et $B(x_B; y_B)$ du plan :

$$d(A, B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

Dans le cercle de centre $X(6, 6)$ se trouvent les 3 points les plus proches parmi les données d'apprentissage.

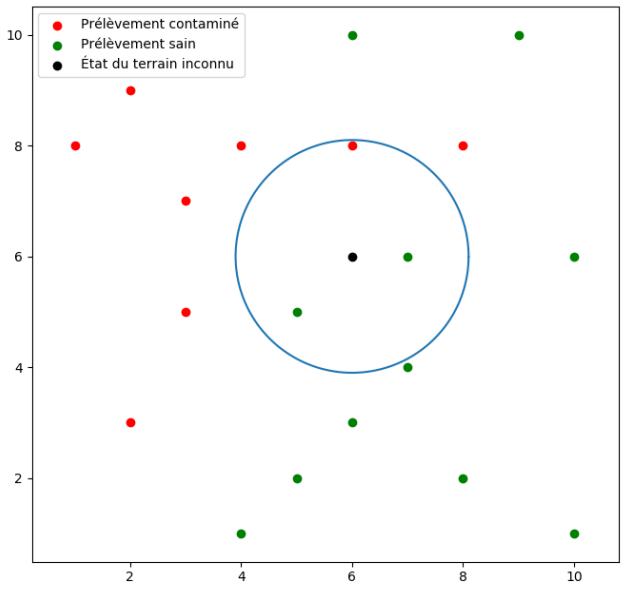


FIGURE 2 – Voisinage du point dont on cherche la classe

Parmi ces voisins, deux ont été étiquettés "sains" par les scientifiques, un seul a été étiquetté "contaminé". On peut selon toute vraisemblance penser qu'il s'agit d'une zone saine.

Exercice 1. Appliquer l'algorithme des plus proches voisins aux données suivantes. On évalue 12 élèves lors d'une épreuve sportive (le 100 mètre). On classe les élèves en deux groupes : les "sportifs" et les "non-sportifs". Ci-dessous les résultats de cette expérience.

Temps (s)	9,7	9,85	9,9	10	10,1	10,1	10,2	10,2	10,3	10,5	10,8	11
Âge	23	26	18	15	25	31	16	33	28	32	39	34
Classe	S	S	S	S	S	NS	S	NS	NS	NS	NS	NS

1. Représenter le nuage de points associé à ces données. $0,1$ sur l'axe des abscisses sera représenté par 1cm. 2 sur l'axe des ordonnées sera représenté par 1cm. On prendra $9,5$ comme abscisse minimale, 11 comme abscisse maximale, 10 comme ordonnée minimale et 40 comme ordonnée maximale. On utilisera la couleur verte pour dénoter les sportifs et la couleur bleue pour dénoter les non-sportifs.
2. Paul à 27 ans, et met 10 secondes à parcourir le 100m. Placer sur le graphique de la question précédente le point associé à Paul.
3. Si un candidat a x ans et parcourt le 100m en t secondes, la "distance" de ce candidat à Paul est donné par la formule :

$$\sqrt{(27-x)^2+(10-t)^2}.$$

Compléter le tableau suivant.

Temps (s)	$9,7$	$9,85$	$9,9$	10	$10,1$	$10,1$	$10,2$	$10,2$	$10,3$	$10,5$	$10,8$	11
Âge	23	26	18	15	25	31	16	33	28	32	39	34
Classe	S	S	S	S	S	NS	S	NS	NS	NS	NS	NS
Distance à Paul												

4. Selon l'algorithme des 3 plus proches voisins, Paul est-il plutôt sportif ou non ?

   .....

   .....

5. Reprendre les questions précédentes avec Jacques, qui a 30 ans, et court lui aussi le 100m en 10 secondes. Quelle remarque pouvez-vous faire quant aux limites de l'algorithme ?

   .....

   .....

   .....

Une généralisation. Comment faire dans le cadre de données plus complexes ? Imaginons une banque qui possède une liste de clients, répartis en 2 catégories : les clients "sûrs", et les clients "peu sûrs". Pour chacun des clients, 3 valeurs sont associées : l'âge, le salaire annuel en miliers d'euros, et le nombre de crédits. Les données possédées par la banque sont résumées dans le tableau suivant.

Âge	50	40	30	45	18	60
Salaire	48	36	96	30	24	24
Nombre de crédit	0	2	3	1	2	2
Classe	S	PS	PS	S	PS	S

Alexis, un nouveau client arrive à la banque. Il a 45 ans, gagne 33 000 euros par ans, et possède un crédit à charge. On aimerait savoir s'il peut être considéré comme sûr ou non. Si un client de la base de donnée de la banque a x ans, gagne s milliers d'euros et possède n crédits, on peut définir la distance de ce client à Alexis par la formule :

$$\sqrt{(45-x)^2+(33-s)^2+(1-n)^2}$$

Exercice 2. Appliquer l'algorithme des 3 plus proches voisins pour déterminer si Alexis peut être considéré comme un client sûr ou non.

. . . . . .
. . . . . .
. . . . . .
. . . . . .

Remarque. Si les données possèdent N caractéristiques (ou **composantes**), elles sont représentées par des "points" $A(a_1, \ldots, a_N)$ (une donnée), $B(b_1, \ldots, b_N)$ (une autre donnée), etc. La distance de A à B est alors donnée par la formule :

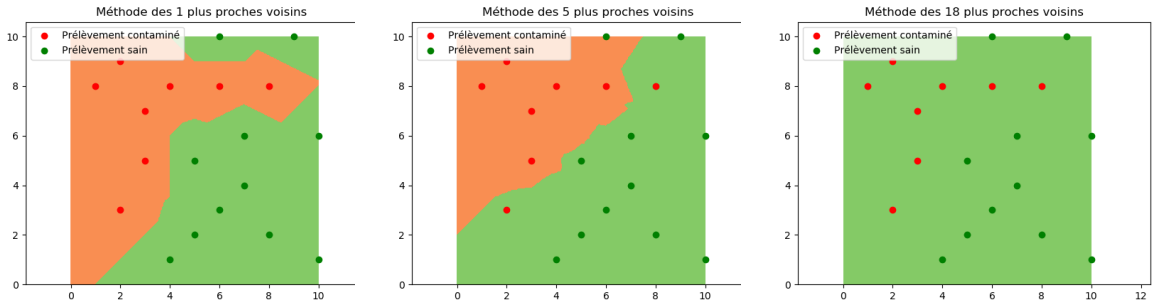
$$d(A, B) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + \ldots + (b_N - a_N)^2}.$$

2 Paramètres de l'algorithme

2.1 La valeur k

k est un paramètre de l'algorithme, définit par l'utilisateur. Différentes valeurs sont possibles.

- $k = 1$: l'algorithme "colle" au maximum au données ;
- $k = 5$: l'algorithme travaille "en moyenne" : il peut parfois être intéressant de se détacher des données pour avoir un point de vue plus "global" ;
- $k = +\infty$: pas d'apprentissage, l'algorithme renvoie toujours la même classe car toutes les données sont considérées à chaque fois dans le vote.



(a) $k = 1$ (b) $k = 5$ (c) $k = +\infty$

FIGURE 3 – Influence de k

2.2 Les données d'apprentissage

Comme on l'a vu dans l'exercice 1, il est possible que l'algorithme prenne des décisions qui ne semblent pas raisonnables si les données ont été mal échantillonnées. L'échantillon d'apprentissage doit être le plus général possible, et doit représenter au maximum les différentes possibilités pour les données.

Remarque. Dans l'exercice 1 toutes les personnes ayant moins de 26 ans étaient étiquettées "sportives" tandis que les personnes plus âgées étaient étiquettées "non sportives". Il y avait donc un biais dans notre algorithme ! Celui-ci apprennait en fait à distinguer si la personne était jeune ou non, ce qui n'était pas vraiment le but...

Pour palier à cela nous aurions dû utiliser un jeu de données plus uniforme, en prennant en compte les personnes jeunes non sportives et les personnes moins jeunes sportives.

2.3 La distance

Cet algorithme utilise la notion de "distance" pour trouver les voisins des éléments que l'on souhaite classer. Cette distance est là pour signifier à quel point les données "se ressemblent". Pour juger de la proximité de deux points $A(x_A, y_A)$ et $B(x_B, y_B)$, il n'est pas toujours judicieux d'utiliser la formule que l'on connait bien :

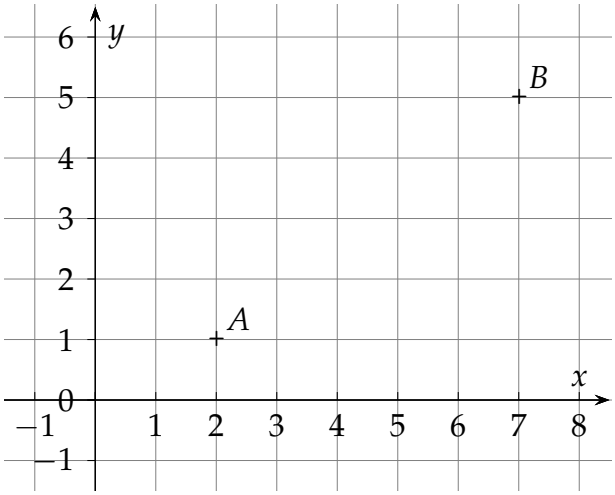
$$d(A, B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

Le mathématicien décide parfois d'utiliser une autre "règle" pour mesurer les distances entre les objets. Avec son esprit tordu il appelle aussi cela "distance", même si ce n'est pas celle que vous connaissez. Par exemple on peut décider que :

$$d(A, B) = |x_B - x_A| + |y_B - y_A|$$

On appelle cette distance la distance de Manhattan, car c'est la distance que l'on parcourt à pied pour aller d'un point A à un point B dans les rues de Manhattan.

Exercice 3. Imaginons que nous avons représenté les rues de Manhattan par le quadrillage suivant. Chaque arrête du quadrillage symbolise une rue, les piétons ne peuvent se déplacer **que** sur les arêtes. L'unité est la dizaine de mètres.



1. Un piéton part du point A et se rend au point B, quelle distance parcourt-il ? Après avoir indiqué les coordonnées respectives de A et de B, comparer avec la formule :

$$d(A, B) = |x_B - x_A| + |y_B - y_A|$$

.....
.....
.....

2. Un oiseau part du point A et se rend au point B, quelle distance parcourt-il ? Quelle formule avez-vous utilisé ?

.....
.....
.....