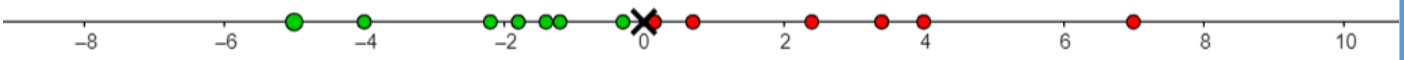
Algorithme des K plus proches voisins

I Introduction

A. Activité 1

Plusieurs points ronds verts et rouges sont situés sur un axe comme le montre la figure ci – dessous.



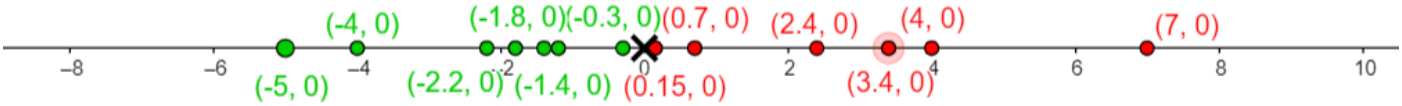
On a placé une croix noire sur ce même axe. On souhaite l'inclure dans un groupe de couleur. On se pose la question : la croix est –elle plus proche des ronds rouges ou des ronds verts ?

Pour cela, pour l'instant à vue d'œil, on va repérer la couleur prédominante parmi les 3 ronds les plus proches.
Dans ce cas, quelle catégorie classerait-on la croix ?
On classerait la croix dans la catégorie des ronds .....

On se repose la question en prenant les 5 ronds les plus proches.
Dans ce cas, quelle catégorie classerait-on la croix ?
On classerait la croix dans la catégorie des ronds .....

On se repose la question en prenant les 7 ronds les plus proches.
Dans ce cas, quelle catégorie classerait-on la croix ?
On classerait la croix dans la catégorie des ronds .....

On souhaite justifier les 3 résultats annoncés précédemment. On a gradué l'axe précédent en plaçant l'origine du repère sur la croix.



Quel outil mathématique permet de calculer la distance entre un point et la croix ?

Compléter le tableau suivant en indiquant les distance les plus courtes et la couleur des points correspondants.

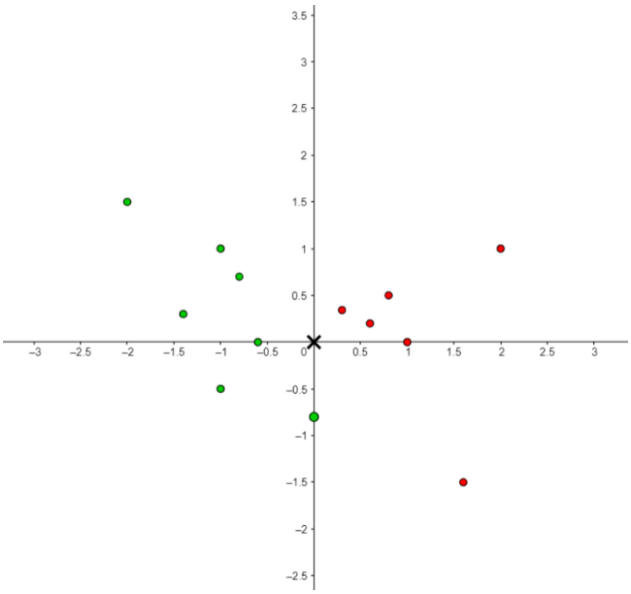
Pour k = 3	
Distances plus petites	Couleur

Pour k = 5	
Distances plus petites	Couleur

Pour k = 7	
Distances plus petites	Couleur

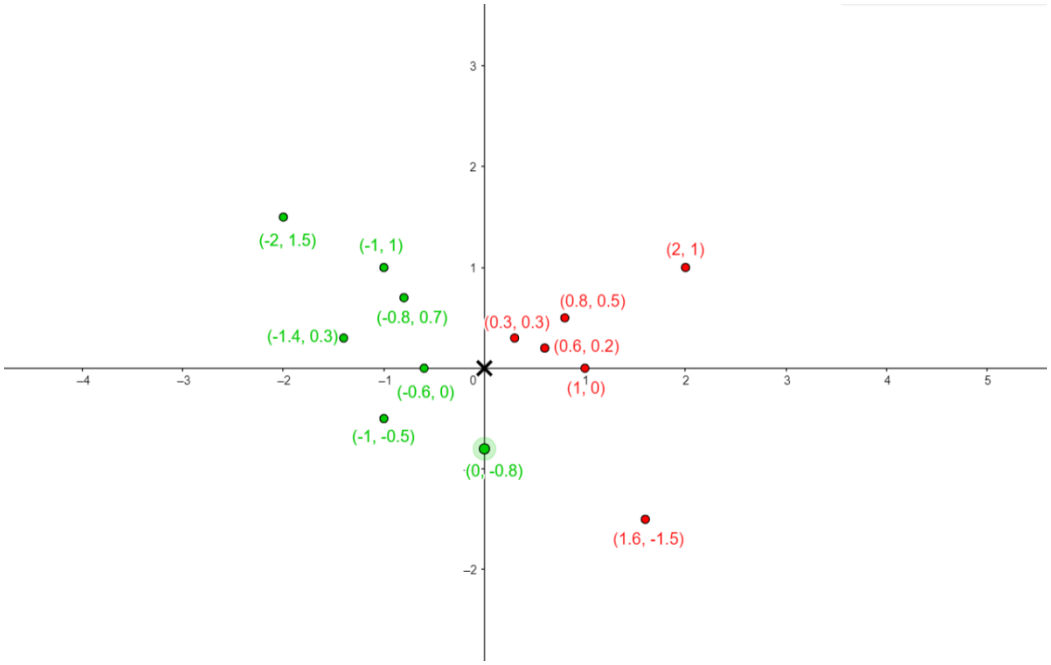
B. Activité 2

On se pose le même problème dans cette configuration de points.



Sans faire de calcul, si on sélectionne les 3 points les plus proches de la croix, dans quelle catégorie la place-ton ?

On a indiqué les coordonnées des points dans le repère du plan.



https://www.geogebra.org/classroom/byzbwtkr **ou**
https://www.geogebra.org/classic/buhuvymd

On rappelle la formule pour calculer la distance entre deux points A(x_A, y_A) et B(x_B, y_B) du plan : d = $\sqrt{(x_B - x_A)^2 + (y_B - y)^2}$

Compléter le tableau suivant en indiquant les distance les plus courtes et la couleur des points correspondants.

Pour k = 3	
Distances plus petites	*Couleur*

Pour k = 5	
Distances plus petites	*Couleur*

II Algorithme des K plus proches voisins

Regarder :

https://bscv-my.sharepoint.com/:v:/g/personal/myriam_lair_edu-baudimont_com/EaORuco3OxRDqdV5cniEDekBxGdNRrNpXvhnR6tkGsFSLQ?e=CFQecn

A. Définition

L' algorithme des K plus proches voisins noté aussi l'algorithme K-NN (K-nearest neighbors) est une méthode d'apprentissage supervisé. C'est un algorithme simple de « machine learning » un sujet très en vogue à l'heure actuelle dans le domaine de l'informatique.

Pour effectuer une prédiction, l'algorithme K-NN va se baser sur tout un jeu de données que l'on fournit à l'ordinateur pour qu'il y trouve des similarités (c'est ce que l'on appelle de l'apprentissage supervisé). C'est un algorithme dit d'apprentissage automatique qui permet à un programme d'apprendre à classer des « objets ».

Son fonctionnement peut être assimilé à l'analogie suivante "dis-moi qui sont tes voisins, je te dirais qui tu es...".

L'algorithme des K-plus proches voisins a été publié en 1951 par Evelyn Fix et Joseph L. Hodges.

B. Exemple

On peut fournir à un programme une grande quantité d'écritures de chiffres.

Le programme va lire toutes les données, et grâce à des algorithmes plus ou moins évolués, le programme va trouver les points communs entre les chiffres représentant le même nombre.

Ensuite, on peut donner au programme une image non annotée, et il nous dira s'il s'agit d'un 1, d'un 6 ou d'un 8...

C'est un système qui est utilisé depuis des années pour la lecture des codes postaux sur les lettres avec un efficacité supérieure à 99%.

C. Principe de l'algorithme en langage naturel

Début Algorithme

Données en entrée :

un ensemble de données D .

une fonction de définition distance d.

Un nombre entier K

Pour une nouvelle observation X dont on veut prédire sa variable de sortie y Faire :

1. *Calculer toutes les distances de cette observation X avec les autres observations du jeu de données*
2. *Retenir les K observations du jeu de données D les proches de X en utilisation la fonction de calcul de distance d*
3. *Prendre les valeurs de y des K observations retenues :*
   a. *Si on effectue une classification, calculer le mode de K retenues*
   b. *Si on effectue une régression, calculer la moyenne (ou la médiane) de K retenues*
4. *Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par K-NN pour l'observation .*

Fin Algorithme

Dans l'algorithme K-NN, on a besoin d'une fonction de calcul de distance entre deux observations :

Il existe plusieurs fonctions de calcul de distance **: la distance euclidienne**, *la distance de Manhattan, la distance de Hamming (utilisée par exemple en bio-informatique pour comparer des séquences génomiques)* **...etc.**

On choisit la fonction « distance » en fonction du type de données qu'on manipule.

Ainsi on utilise **la distance euclidienne** *pour les données quantitatives (exemple : poids, salaires, taille, etc...) et* **du même type.**

C. Code en Python du problème des points dans le plan

Ecrire le code en Python pour résoudre le problème de l'activité 2.

Pour utiliser la fonction sqrt(fonction racine carrée), il faut importer le module math.

III. Application de l'algorithme KNN

Dans le fichier Platanes_Remarquables_Paris.csv, sont données des informations de plusieurs platanes dans Paris.
Y sont indiqués la circonférence de l'arbre en cm, la taille en m, le stade de développement : 1 pour jeune, 2 pour adulte et 3 pour mature et la caractéristique d'arbre remarquable ou non remarquable.

En utilisant l'algorithme des k plus proches voisins, prévoir si les trois platanes suivants dont les caractéristiques sont données dans le tableau ci-dessous sont remarquables ou non remarquables.

	Circonférence (cm)	Taille (m)	Stade de développement
Platane 1	230	18	3
Platane 2	290	25	2
Platane 3	553	29	3

Le plus grand arbre de Paris, un platane d'Orient plus que centenaire, veille sur le parc de l'hôtel de Villeroy, siège du ministère de l'Agriculture et de l'Alimentation.