Data Mining

Mohammed Fethi KHALFI

Fethi.Khalfi@yahoo.fr

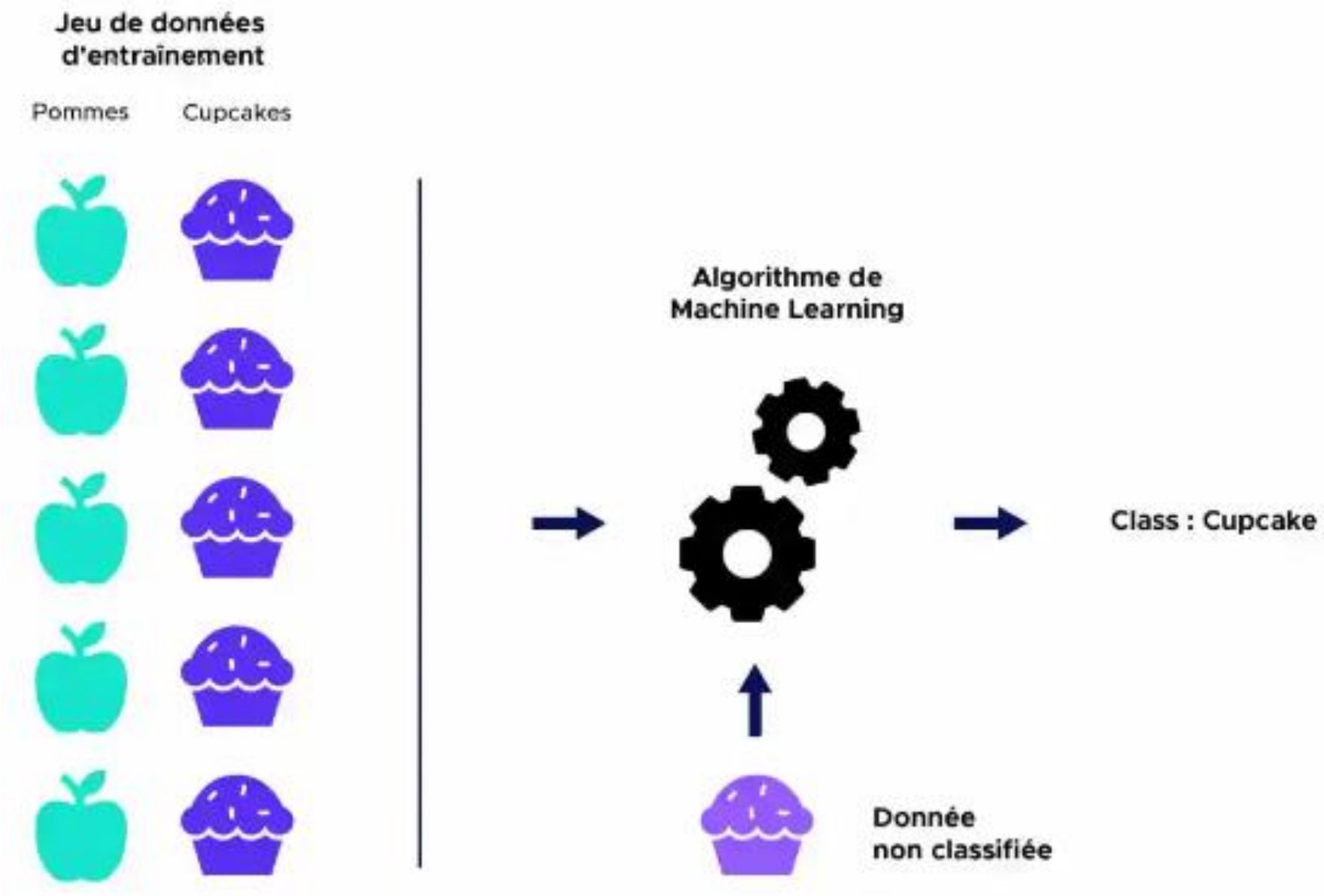
Méthode des k plus proches voisins

Université Djilali Liabés- Sidi Bel Abbes

k - Nearest Neighbour

Knn

Introduction



Définition

En anglais "**k nearest neighbors**" : knn

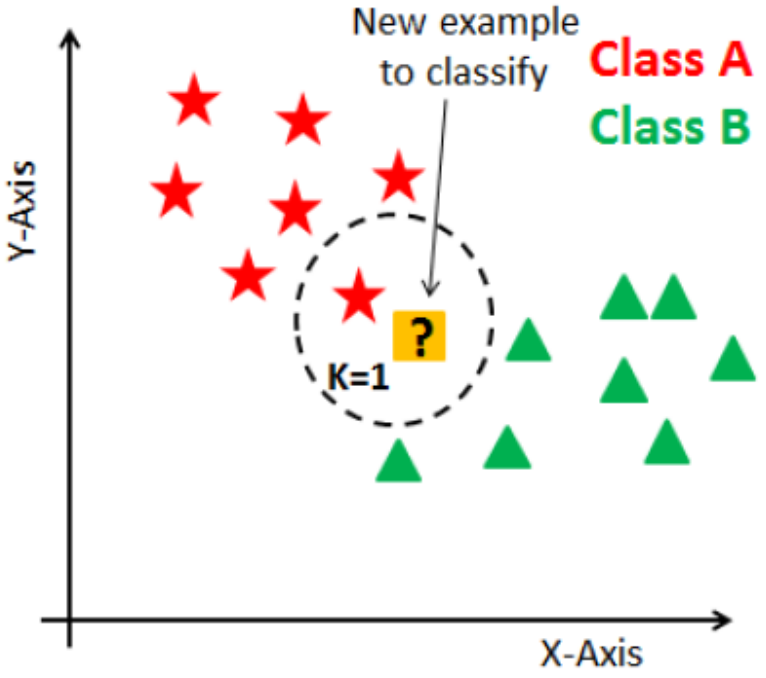
C’est un algorithme utilisé dans des structures d'apprentissage automatique supervisé simple et facile a mettre en ouvre.

Nourrie par un grand nombre d’exemples pour résoudre les problèmes de classification.

Description du problème :

Notre problème est assez simple: on relève sur des objets de différentes classes des paramètres. On sait donc que pour tel objet de telle classe, on a tels paramètres.

L'objectif est de pouvoir prévoir à quelle classe appartient un nouvel objet uniquement à l'aide de ses paramètres (apprentissage supervisé).



Principe de fonctionnement

Dis moi qui sont tes voisins, je te dirais qui tu es !

Les différents paramètres de l'algorithme, réglables par l'utilisateur, sont :

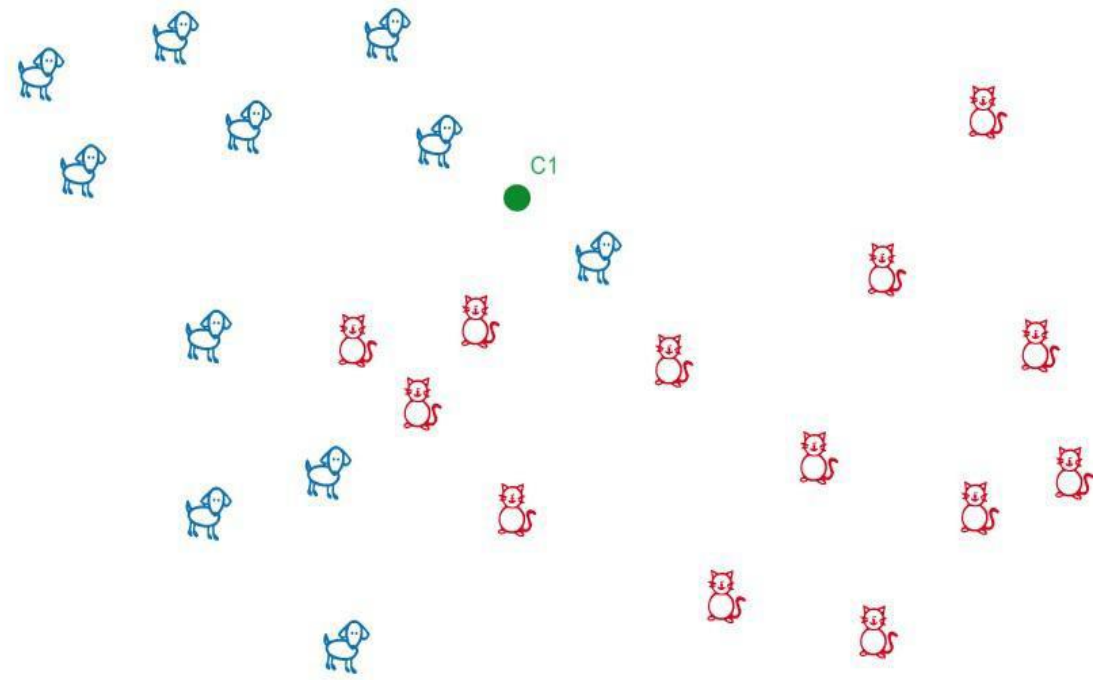
-Un entier k

-Une base d'apprentissage : le nombre d'exemples connue

-la mesure de similarité (ex. : distance euclidienne)

Description du problème :

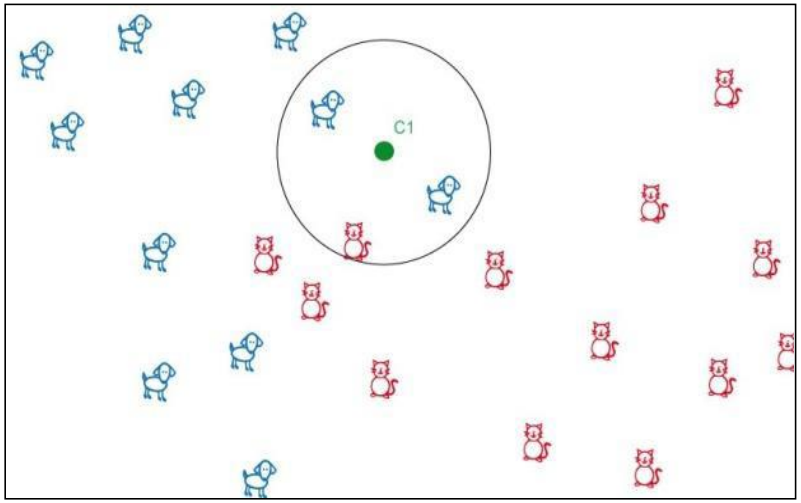
Imaginons qu'on ait quelques milliers de données disponibles de chiens et de chats et que ces données relatent la taille et le poids des animaux. Ceci donnerait le graphe ci-dessous :



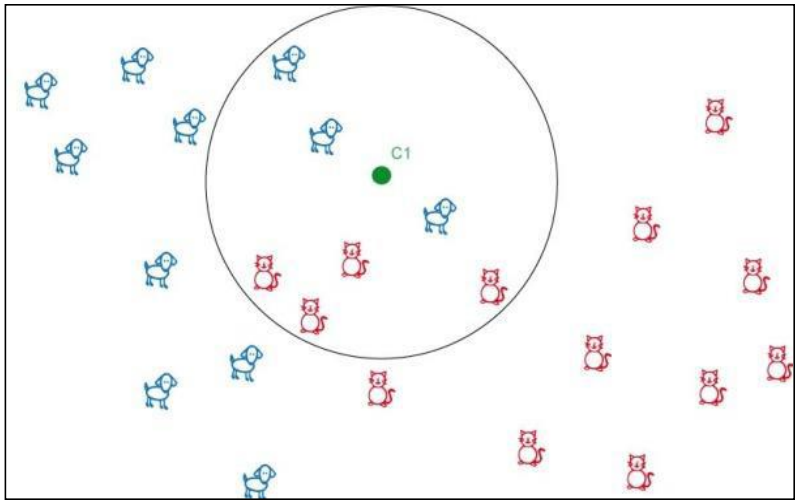
Description du problème :

On donne le poids et la taille d'un nouvel animal au programme (le point vert C1) et celui-ci doit deviner si c'est un chat ou un chien.

Pour ce, le programme va regarder ce qu'il y a de plus proche et va en faire sa déduction. La question a se poser est : dans quel rayon regarde-t-on ses voisins (que prendre comme k ?).



Pour k = 3, on en déduit que C1 est un chien



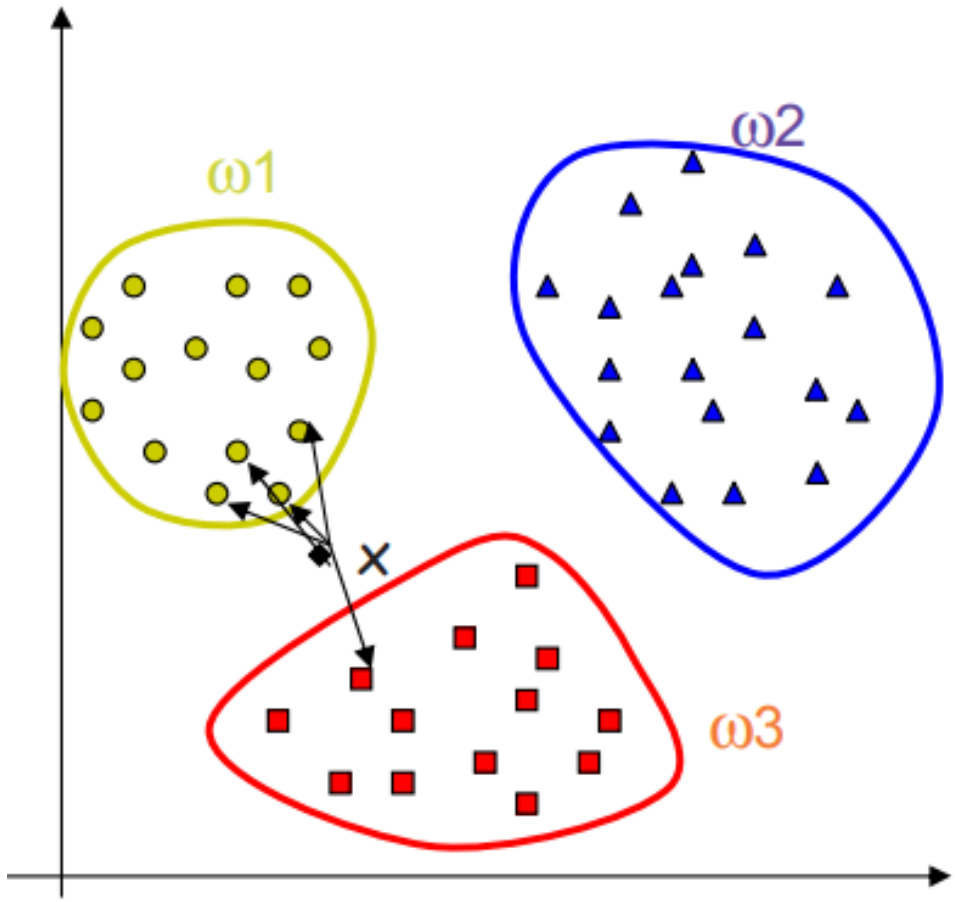
Pour k = 7, on en déduit que c'est un chat !

Principe de fonctionnement

Exemple: Dans l'exemple suivant: on a 3 classes et le but est de trouver la valeur de la classe de l'exemple inconnu x:

-On prend la distance Euclidienne **et k=5 voisins**

-Des 5 plus proches voisins, 4 appartiennent à ω1 et 1 appartient à ω3, donc x est affecté à ω1**, la classe majoritaire**



Les distance les plus usuelles

Comment définir la distance ?

Il existe plusieurs fonctions de calcul de distance on choisit la fonction de distance en fonction des types de données qu'on manipule.

Pour les données **quantitatives** (exemple : poids, salaires, taille, etc.…) et du même type, **la distance euclidienne** est un bon candidat.

Distance Measures

- There are many possible distance measures
  - Euclidean Distance:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

  - Manhattan Distance or City Block Distance :

$$\sum_{i=1}^{k}|x_i - y_i|$$

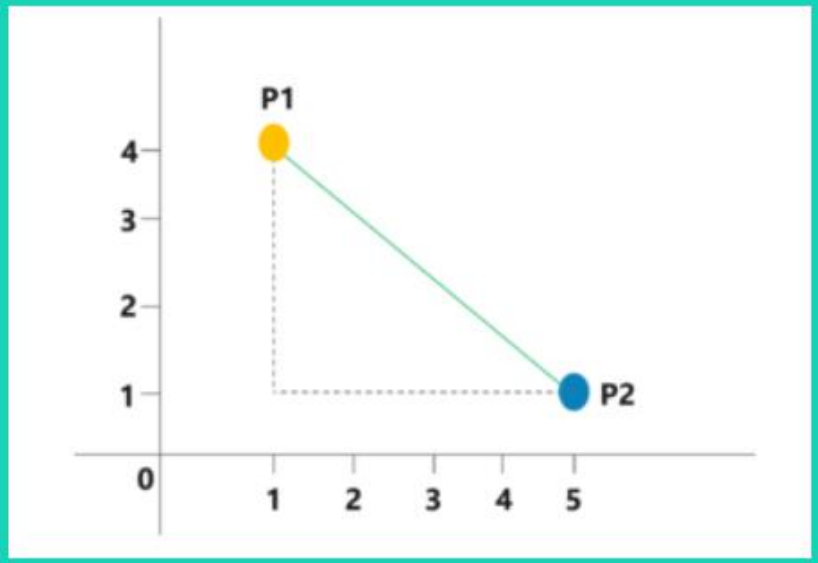
- Hamming Distance: $D_H = \sum_{i=1}^{k}|x_i - y_i|$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Comment définir la distance ?

KNN utilise la distance euclidienne comme mesure pour vérifier la distance entre un nouveau point de données et ses voisins,



Nous allons ici mesurer la distance entre P1 et P2 en utilisant la mesure de distance euclidienne.
Les coordonnées pour P1 et P2 sont respectivement (1,4) et (5,1).

La distance euclidienne peut être calculée comme suit:

Point P1 = (1,4)

Point P2 = (5,1)

Euclidian distance = $\sqrt{(5-1)^2+(4-1)^2} = 5$

Distance Measures: Euclidean Distance

- If we denote an instance in the training set by *(a1, a2)* and the unseen instance by *(b1, b2)* the length of the straight line joining the points is $\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$
- If there are two points *(a1, a2, a3)* and *(b1, b2, b3)* in a three-dimensional space the corresponding formula is

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

- The formula for Euclidean distance between points *(a1, a2, . . . , an)* and *(b1, b2, . . . , bn)* in n-dimensional space is a generalisation of these two results. The Euclidean distance is given by the formula

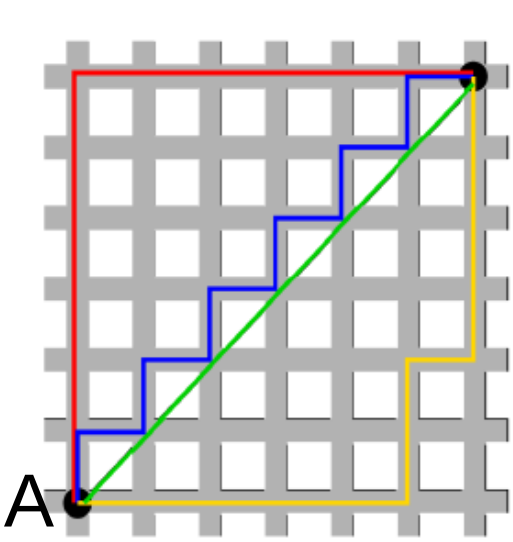
$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Distance Measures: Manhattan Distance

c- Distance de manhattan :

La distance de Manhattan a été utilisée dans une analyse de régression en 1757 par Roger Joseph Boscovich. L'interprétation géométrique remonte à la fin du XIXe siècle et au développement de géométries non euclidiennes, notamment par Hermann Minkowski et son inégalité de Minkowski, dont cette géométrie constitue un cas particulier, particulièrement utilisée dans la géométrie des nombres (Minkowski 1910).

La distance de Manhattan est appelée aussi taxi-distance, est la distance entre deux points parcourue par un taxi lorsqu'il se déplace dans une ville où les rues sont agencées selon un réseau ou quadrillage. Un taxi-chemin est le trajet fait par un taxi lorsqu'il se déplace d'un nœud du réseau à un autre en utilisant les déplacements horizontaux et verticaux du réseau.



Entre deux points *A* et *B*, de coordonnées respectives (X_A , Y_A) et (X_B , Y_B)la distance de Manhattan est définie par :

$$d(A,B) = |X_A - X_B| + |Y_A - Y_B| .$$

Les points A **et** B **sont séparés de 6 unités sur une variable et de 6 unités sur une autre variable.**
La distance de Manhattan Bleu ou jaune ou rouge est de 12.

la distance euclidienne en vert est la racine carrée de 6^2+6^2 **soit 8,48 (en bleu). Manhattan est donc supérieure.**

Distance Measures: Hamming Distance

Elle est définie par :
=0 si la valeur de l'attribut est la même,
=1 sinon.

Ex : l'attribut "couleur" ∈ {rouge, jaune, rose, vert, bleu, turquoise}.
la dissimilarité entre "bleu" et "rouge" est plus grande qu'entre "rouge" et "jaune".

Algorithme K-NN

1. Charger les données
2. Initialiser k au nombre de plus proches voisins choisi
3. Pour chaque exemple dans les données:
    3.1 Calculer la distance entre notre requête et l'observation itérative actuelle de la boucle depuis les données.
    3.2 Ajouter la distance et l'indice de l'observation concernée à une collection ordonnée de données
4. Trier cette collection ordonnée contenant distances et indices de la plus petite distance à la plus grande (dans ordre croissant).
5. Sélectionner les k premières entrées de la collection de données triées (équivalent aux k plus proches voisins)
6. Obtenir les étiquettes des k entrées sélectionnées
7. retourner la valeur la plus fréquente des k étiquettes

KNN

- A training set with 20 instances, each giving the values of two attributes and an associated classification

- How can we estimate the classification for an 'unseen' instance where the first and second attributes are 9.1 and 11.0, respectively?

Attribute 1	Attribute 2	Class
0.8	6.3	−
1.4	8.1	−
2.1	7.4	−
2.6	14.3	+
6.8	12.6	−
8.8	9.8	+
9.2	11.6	−
10.8	9.6	+
11.8	9.9	+
12.4	6.5	+
12.8	1.1	−
14.0	19.9	−
14.2	18.5	−
15.6	17.4	−
15.8	12.2	−
16.6	6.7	+
17.4	4.5	+
18.2	6.9	+
19.0	3.4	−
19.6	11.1	+

$(\underset{b_1}{9.1}, \underset{b_2}{11.0}) = ?$

for $(\underset{a_1}{0.8}, \underset{a_2}{6.3}) = \sqrt{(0.8-9.1)^2 + (6.3-11.0)^2} = 9.54$

for $(1.4, 8.1) = \sqrt{(1.4-9.1)^2 + (8.1-11)^2} = 8.23$

for $(2.1, 7.4) = \sqrt{(2.1-9.1)^2 + (7.4-11)^2} = 7.87$

for $(2.6, 14.3) = \sqrt{(2.6-9.1)^2 + (14.3-11)^2} = 7.29$

for $(6.8, 12.6) = 2.80$

for $(8.8, 9.8) = 1.24$ — (2nd)

$(9.2, 11.6) = 0.61$ — (1st)

$(10.8, 9.6) = 2.20$ — (3rd)

$(11.8, 9.9) = 2.92$

$(12.4, 6.5) = 5.58$

$(12.8, 1.1) = 10.57$

$(14.2, 18.5) = 9.07$

$(18.6, 17.4) = 9.12$

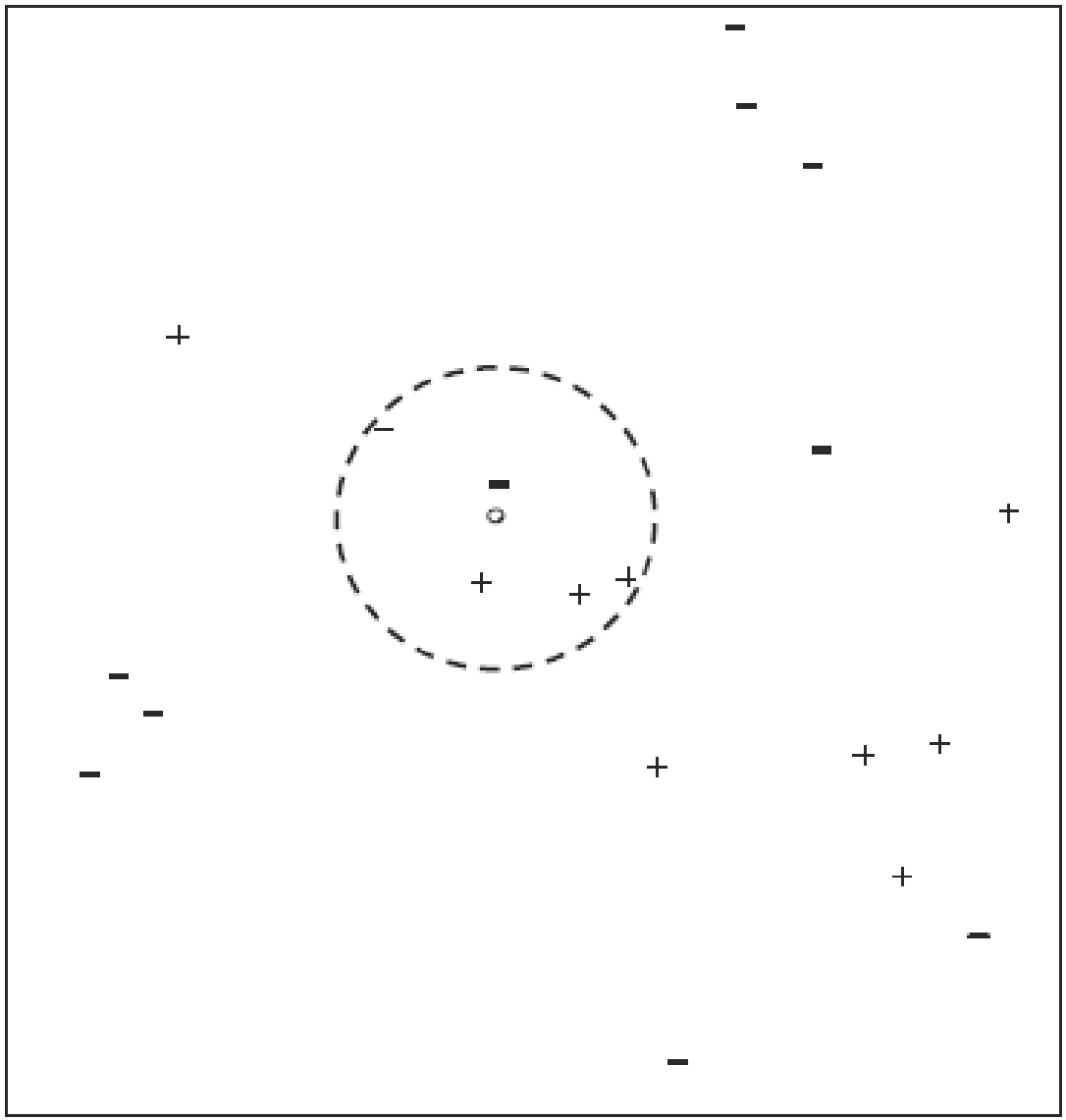
$(15.8, 12.2) =$

initially $k=3$

Attribute	class	Distance
9.2, 11.6	−	0.61
8.8, 9.8	+	1.24
11.8, 9.9	+	2.20

Since, $k=3$ the majority class is +. So, knn predicts that the unknown instance will be in positive class.

KNN



Attribute 1	Attribute 2	Class
0.8	6.3	−
1.4	8.1	−
2.1	7.4	−
2.6	14.3	+
6.8	12.6	−
8.8	9.8	+
9.2	11.6	−
10.8	9.6	+
11.8	9.9	+
12.4	6.5	+
12.8	1.1	−
14.0	19.9	−
14.2	18.5	−
15.6	17.4	−
15.8	12.2	−
16.6	6.7	+
17.4	4.5	+
18.2	6.9	+
19.0	3.4	−
19.6	11.1	+

Attribut nominal

- One of the weaknesses of the nearest neighbour approach to classification is that there is no entirely satisfactory way of dealing with categorical attributes.

- One possibility is to say that the difference between any two identical values of the attribute is zero and that the difference between any two different values is 1. (Hamming Distance)

- Effectively this amounts to saying (for a colour attribute)
- red − red = 0,
- red − blue = 1,
- blue − green = 1, etc.

Attribut nominal

a	b	c	d	e	f	Class
yes	no	no	6.4	8.3	low	negative
yes	yes	yes	18.2	4.7	high	positive

yes	no	no	6.6	8.0	low	???

- What should its classification be?

Exercise-1

Age	Loan	Default	Distance
25	40000	N	
35	60000	N	
45	80000	N	
20	20000	N	
35	120000	N	
52	18000	N	
23	95000	Y	
40	62000	Y	
60	100000	Y	
48	220000	Y	
33	150000	Y	
48	142000	??	

Exercise-1

$K=3$

$K=5$

$K=9$

Age	Loan	Default	Distance
25	40000	N	102000
35	60000	N	82000
45	80000	N	62000
20	20000	N	122000
35	120000	N	22000
52	18000	N	124000
23	95000	Y	47000
40	62000	Y	80000
60	100000	Y	42000
48	220000	Y	78000
33	150000	Y	8000
48	142000	??	

Exercise-2

N° du produit	Forme	Taille	Couleur	Prix	Classe
1	Rond	Petit	Bleu	50.3	Oui
2	Carré	Grand	Rouge	25.3	Non
4	Carré	Petit	Bleu	76.9	Oui
5	Rond	Grand	Bleu	55	Oui
6	Carré	Moyen	Blanc	92	Non
9	Carré	Petit	Rouge	66	Oui
10	Rond	Petit	Rouge	98,3	?

Comment choisir “k” ?

Il n’y a pas de méthode particulière sinon qu’il faut choisir le k qui va bien pour les données que l’on a !

- K grand:

- *moins sensible au bruit*
- *Une grande base d’apprentissage permet une plus grande valeur de k.*

- K petit :

- *Rend mieux compte de structures fines*
- *Nécessaire pour des petites bases d’apprentissage*

Conclusion

- L'algorithme KNN est l'un des algorithmes de classification les plus simples.
- K-NN stocke tout le jeu de données pour effectuer une prédiction.
- K-NN ne calcule aucun modèle prédictif et il rentre dans le cadre du Lazy Learning.