Analyse, Classification et Indexation des données (ACID)

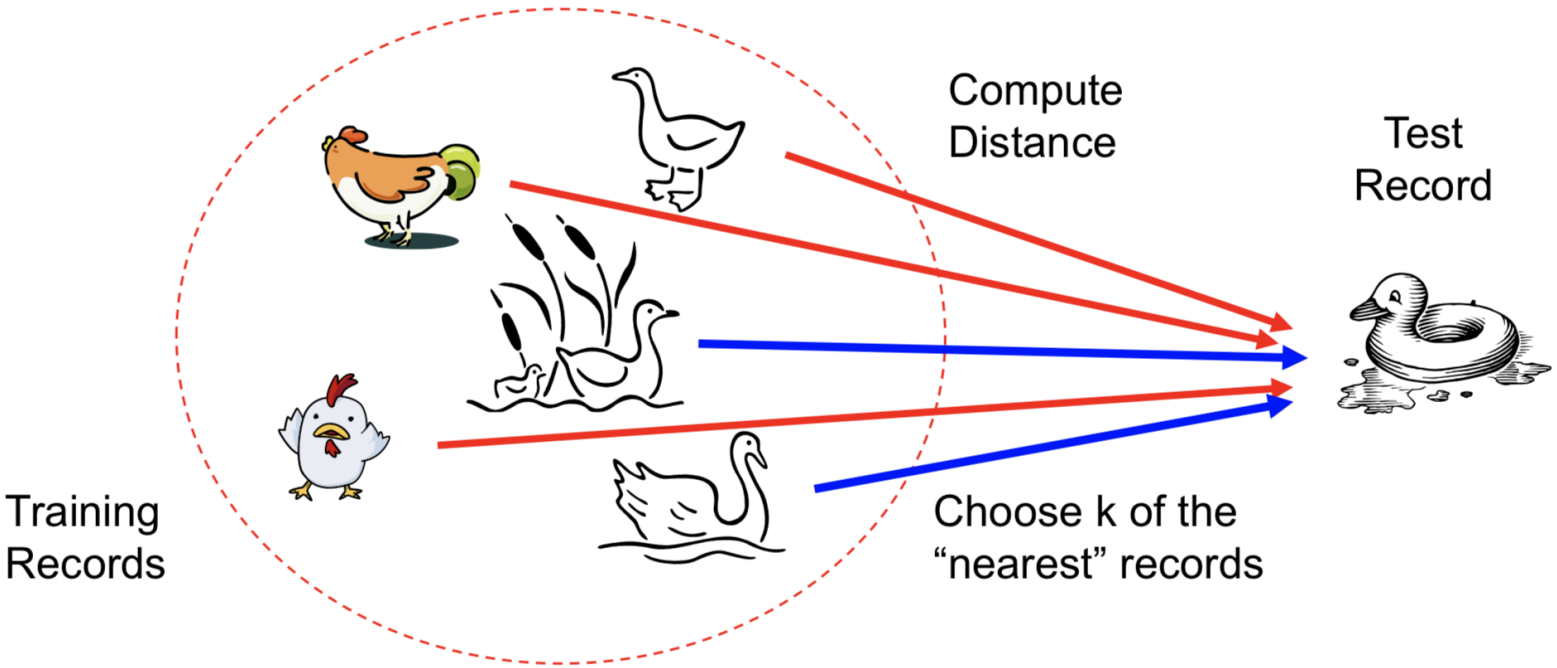
Apprentissage supervisé : k-NN ou Nearest Neighbor Classifier

Akka Zemmari

LaBRI, Université de Bordeaux

2024 - 2025

Idées de base



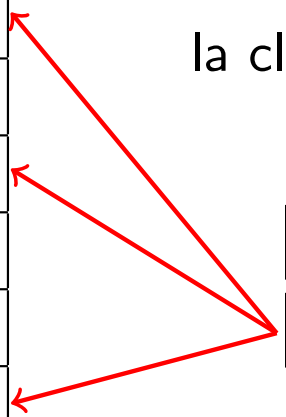
Si ça marche comme un canard, crie comme un canard, c'est que c'est probablement un canard ... [1]

[1] Schéma de B. S. Panda (IIT Delhi)

Idées de base

Données d'entrainement

A_1	A_2	$\cdots$	A_n	Classe
				A
				B
				B
				C
				A
				C
				B

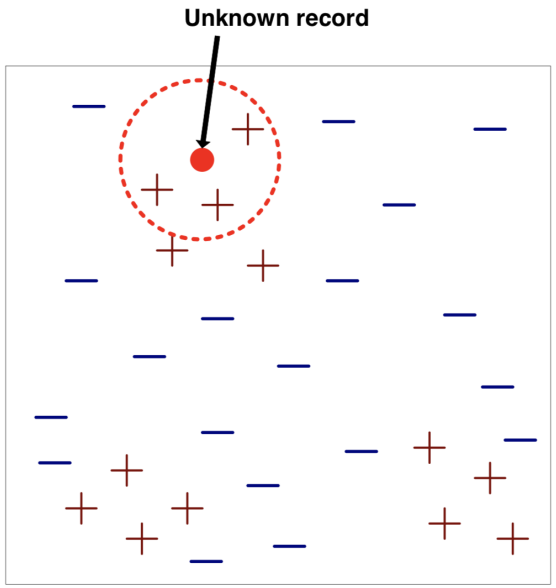


A_1	A_2	$\cdots$	A_n

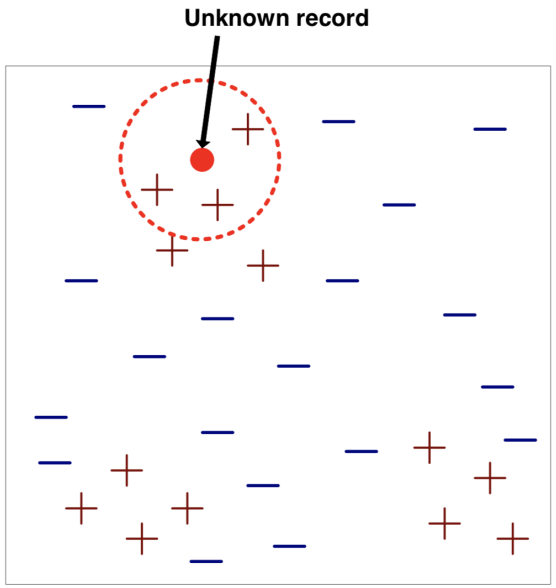
- Stocker les données d'entraînement.

- Utiliser ces données pour prédire la classe de l'ensemble de test.

Nearest-Neighbor Classifiers

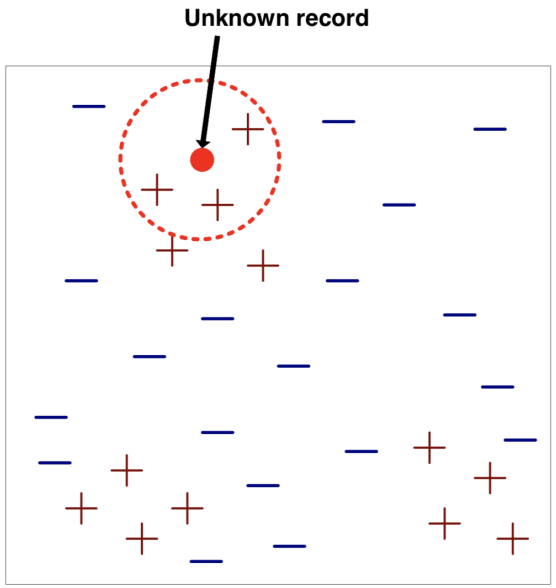


Nearest-Neighbor Classifiers



On a besoin de trois choses :

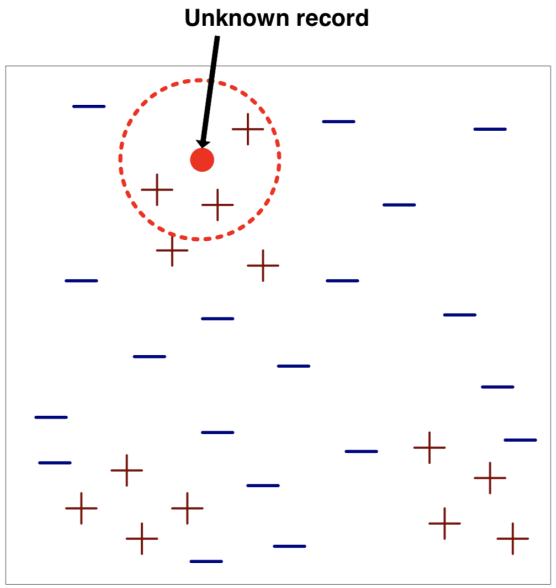
Nearest-Neighbor Classifiers



On a besoin de trois choses :

- Un ensemble d'entrainement.

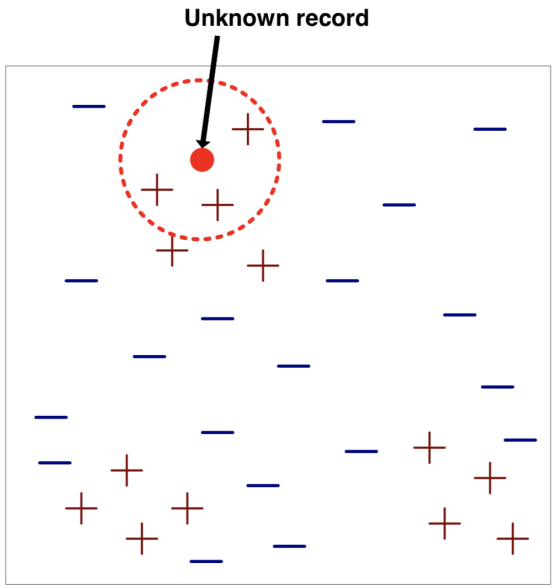
Nearest-Neighbor Classifiers



On a besoin de trois choses :

- Un ensemble d'entrainement.
- Une mesure de distance.

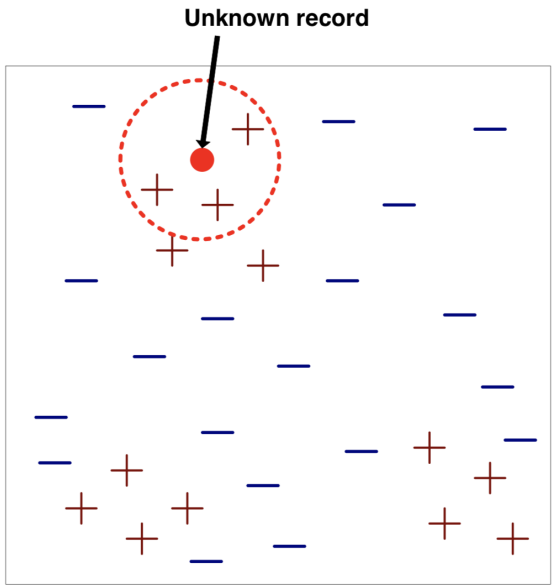
Nearest-Neighbor Classifiers



On a besoin de trois choses :

- Un ensemble d'entrainement.
- Une mesure de distance.
- La valeur de k, le nombre de voisins à interroger.

Nearest-Neighbor Classifiers

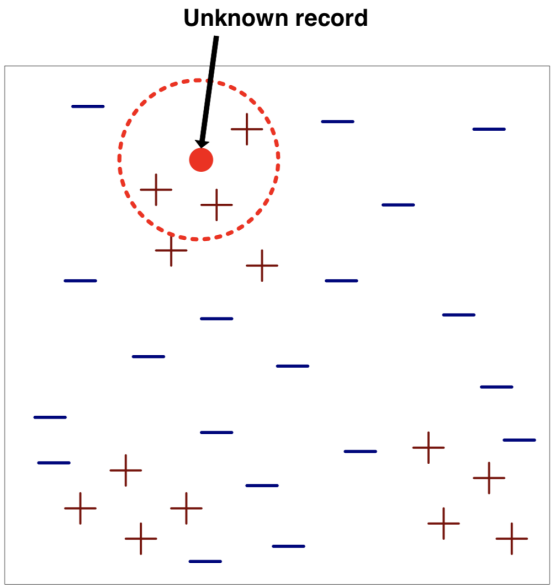


On a besoin de trois choses :

- Un ensemble d'entrainement.
- Une mesure de distance.
- La valeur de k, le nombre de voisins à interroger.

Pour classifier un nouvel enregistrement :

Nearest-Neighbor Classifiers



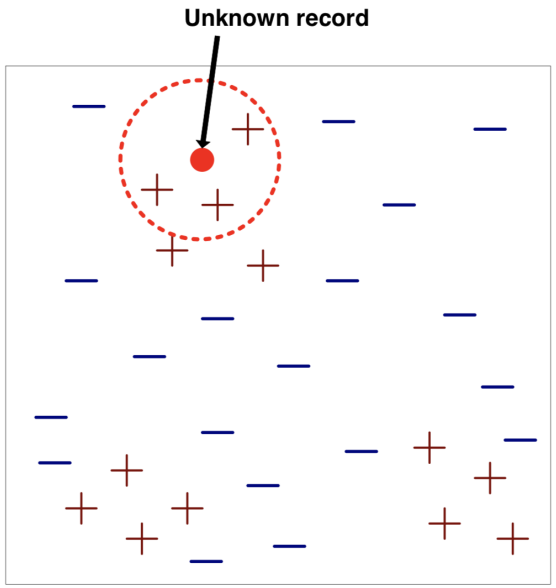
On a besoin de trois choses :

- Un ensemble d'entrainement.
- Une mesure de distance.
- La valeur de k, le nombre de voisins à interroger.

Pour classifier un nouvel enregistrement :

- Calculer la distance vers les autres enregistrements (de l'ensemble d'entraînement).

Nearest-Neighbor Classifiers



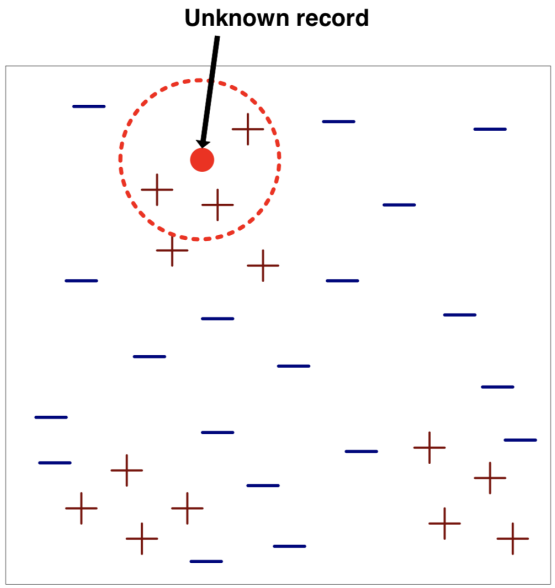
On a besoin de trois choses :

- Un ensemble d'entrainement.
- Une mesure de distance.
- La valeur de k, le nombre de voisins à interroger.

Pour classifier un nouvel enregistrement :

- Calculer la distance vers les autres enregistrements (de l'ensemble d'entraînement).
- Identifier k plus proches voisins.

Nearest-Neighbor Classifiers



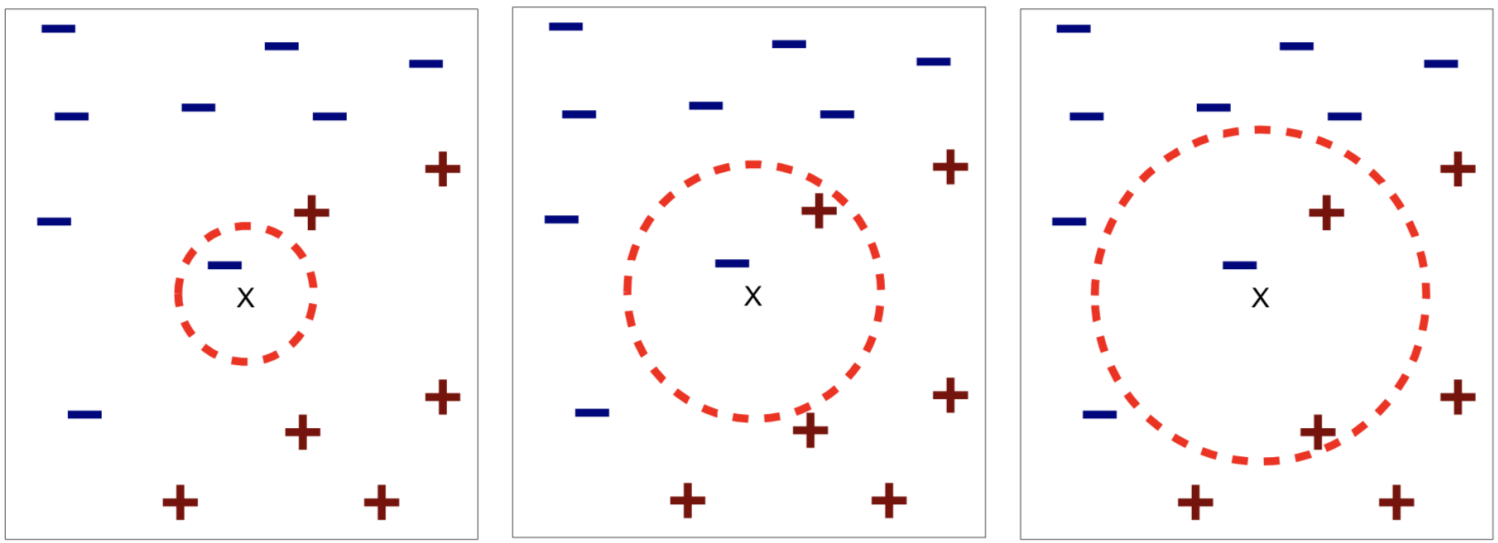
On a besoin de trois choses :

- Un ensemble d'entrainement.
- Une mesure de distance.
- La valeur de k, le nombre de voisins à interroger.

Pour classifier un nouvel enregistrement :

- Calculer la distance vers les autres enregistrements (de l'ensemble d'entraînement).
- Identifier k plus proches voisins.
- Utiliser la classe des k voisins les plus proches pour déterminer la classe du nouvel enregistrement (par un vote majoritaire par exemple).

Nearest-Neighbor ?



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

Choix des métriques pour mesurer la distance

- Une distance $d(.,.)$ doit vérifier les propriétés suivantes:
  $d(i,j) > 0$
  $d(i,i) = 0$
  $d(i,j) = d(j,i)$
  $d(i,j) \leq d(i,k) + d(k,j)$

Calcul de distance

On commence par "standardiser" les données (voir le premier cours)

Exemple de distances : distance de Minkowski

$$d(i,j) = \sqrt[q]{\sum_{k=1}^{m} \mid x_{i,k} - x_{j,k} \mid^q}\ ,\ i \text{ et } j \text{ sont deux individus.}$$

Cas particuliers :

- Distance euclidienne $q = 2$

$$d(i,j) = \sqrt{\sum_{k=1}^{m} (x_{i,k} - x_{j,k})^2}$$

- Distance de Manhattan $q = 1$

$$d(i,j) = \sqrt{\sum_{k=1}^{m} \mid x_{i,k} - x_{j,k} \mid}$$

Exemple : distance de Manhattan

	Age	Salaire
Guillaume	50	11 000
Meriem	70	11 100
Youness	60	11 122
Lina	60	11 074

Exemple : distance de Manhattan

	Age	Salaire
Guillaume	50	11 000
Meriem	70	11 100
Youness	60	11 122
Lina	60	11 074

$$\xrightarrow[\overline{Sal} = 11074,\ S_{Sal} = 148]{\overline{Age} = 60,\ S_{Age} = 5}$$

Exemple : distance de Manhattan

	Age	Salaire
Guillaume	50	11 000
Meriem	70	11 100
Youness	60	11 122
Lina	60	11 074

$$\xrightarrow[\overline{Sal} = 11074,\ S_{Sal} = 148]{\overline{Age} = 60,\ S_{Age} = 5}$$

	Age	Salaire
Guillaume	-2	-0.5
Meriem	2	0.175
Youness	0	0.324
Lina	0	2

Exemple : distance de Manhattan

	Age	Salaire
Guillaume	50	11 000
Meriem	70	11 100
Youness	60	11 122
Lina	60	11 074

$$\xrightarrow[\overline{Sal} = 11074,\ S_{Sal} = 148]{\overline{Age} = 60,\ S_{Age} = 5}$$

	Age	Salaire
Guillaume	-2	-0.5
Meriem	2	0.175
Youness	0	0.324
Lina	0	2

$$\begin{cases} d(G, M) = 120 \\ d(G, Y) = 132 \end{cases} \Rightarrow \text{G. ressemble plus à M. qu'à Y.}$$

Exemple : distance de Manhattan

	Age	Salaire
Guillaume	50	11 000
Meriem	70	11 100
Youness	60	11 122
Lina	60	11 074

$$\xrightarrow[\overline{Sal} = 11074,\ S_{Sal} = 148]{\overline{Age} = 60,\ S_{Age} = 5}$$

	Age	Salaire
Guillaume	-2	-0.5
Meriem	2	0.175
Youness	0	0.324
Lina	0	2

$$\begin{cases} d(G,M) = 120 \\ d(G,Y) = 132 \end{cases} \Rightarrow \text{G. ressemble plus à M. qu'à Y.}$$

$$\begin{cases} d(G,M) = 4.675 \\ d(G,Y) = 2.324 \end{cases} \Rightarrow \text{G. ressemble plus à Y. qu'à M.}$$

Exemple : distance de Manhattan

	Age	Salaire
Guillaume	50	11 000
Meriem	70	11 100
Youness	60	11 122
Lina	60	11 074

$\overline{Age} = 60,\ S_{Age} = 5$

$\overline{Sal} = 11074,\ S_{Sal} = 148$

	Age	Salaire
Guillaume	-2	-0.5
Meriem	2	0.175
Youness	0	0.324
Lina	0	2

$$\begin{cases} d(G,M) = 120 \\ d(G,Y) = 132 \end{cases} \Rightarrow \text{G. ressemble plus à M. qu'à Y.}$$

$$\begin{cases} d(G,M) = 4.675 \\ d(G,Y) = 2.324 \end{cases} \Rightarrow \text{G. ressemble plus à Y. qu'à M.}$$

Qu'en pensez-vous ?

NN Classification

Déterminer la classe à partir de la liste de voisins (pour classifier un exemple z) :

NN Classification

Déterminer la classe à partir de la liste de voisins (pour classifier un exemple z) :

- choisir la classe majoritaire dans le k-voisinage :

$$y' = \arg\max_{v} \sum_{(x_i, y_i) \in D_z} I\left(v = y_i\right), \tag{1}$$

où D_z est l'ensemble des k exemples les plus proches de z.

k-NN en une diapo

k-NN en une diapo

Soit k le nombre de voisins les plus proches et D l'ensemble d'entraînement.

k-NN en une diapo

Soit k le nombre de voisins les plus proches et D l'ensemble d'entraînement.

1. pour chaque exemple $z = (x', ?)$ de l'ensemble de test :

k-NN en une diapo

Soit k le nombre de voisins les plus proches et D l'ensemble d'entraînement.

1. pour chaque exemple $z = (x', ?)$ de l'ensemble de test :
   1.1 Calculer $d(x, x')$, la distance de z et chaque exemple (x, y) de D;

k-NN en une diapo

Soit k le nombre de voisins les plus proches et D l'ensemble d'entraînement.

1. pour chaque exemple $z = (x', ?)$ de l'ensemble de test :
   - 1.1 Calculer $d(x, x')$, la distance de z et chaque exemple (x, y) de D;
   - 1.2 Choisir $D_z \subset D$, l'ensemble des k exemples les plus proches de z;

k-NN en une diapo

Soit k le nombre de voisins les plus proches et D l'ensemble d'entraînement.

1. pour chaque exemple $z = (x', ?)$ de l'ensemble de test :
   - 1.1 Calculer $d(x, x')$, la distance de z et chaque exemple (x, y) de D;
   - 1.2 Choisir $D_z \subset D$, l'ensemble des k exemples les plus proches de z;
   - 1.3 $y' = \arg\max_v \sum_{(x_i, y_i) \in D_z} I(v = y_i)$

k-NN en une diapo

Soit k le nombre de voisins les plus proches et D l'ensemble d'entraînement.

1. pour chaque exemple $z = (x', ?)$ de l'ensemble de test :
   - 1.1 Calculer $d(x, x')$, la distance de z et chaque exemple (x, y) de D;
   - 1.2 Choisir $D_z \subset D$, l'ensemble des k exemples les plus proches de z;
   - 1.3 $y' = \arg\max_v \sum_{(x_i, y_i) \in D_z} I(v = y_i)$
2. Fin pour

NN Classification

Choix de le valeur de k :

NN Classification

Choix de le valeur de k :

- Si k est trop petit, la classification sera trop sensible au "bruit".

NN Classification

Choix de le valeur de k :

- Si k est trop petit, la classification sera trop sensible au "bruit".
- Si k est trop grand, le voisinage peut contenir des éléments d'autres classes.

NN Classification

Quelques précautions à prendre :

NN Classification

Quelques précautions à prendre :

- Les attributs doivent être normalisés pour éviter que les distances soient faussées par des attributs à grande valeur.

NN Classification

Quelques précautions à prendre :

- Les attributs doivent être normalisés pour éviter que les distances soient faussées par des attributs à grande valeur.
- Exemple : Taille (H), Poids (W) et revenu (I) d'une personne avec :
  - $H \in [1.5m, 1.8m]$
  - $W \in [60kg, 100kg]$
  - $I \in [15k, 60k]$.

NN Classification

Quelques précautions à prendre :

NN Classification

Quelques précautions à prendre :
Attention à la distance euclidienne ...

NN Classification

Quelques précautions à prendre :
Attention à la distance euclidienne ...

- Vecteurs de features à grande dimension
  $\rightarrow$ presque tous les vecteurs sont à la même distance de l'exemple qu'on veut classifier.

NN Classification

Quelques précautions à prendre :
Attention à la distance euclidienne ...

- Vecteurs de features à grande dimension
  $\rightarrow$ presque tous les vecteurs sont à la même distance de l'exemple qu'on veut classifier.
- Solution : réduire la dimension des vecteurs (ACP par exemple)

Démo

Voyons comment ça marche en pratique