K-nearest neighbors

Dr. Ilham KADI

i.kadi@emsi.ma

2023/2024

K-NN: Définition

K-NN (K-nearest neighbors) est une méthode d'apprentissage supervisé. Il peut être utilisé aussi bien pour la **régression** que pour la **classification**. Son fonctionnement peut être assimilé à l'analogie suivante:

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

K-NN: Définition

- L'algorithme K-NN figure parmi les plus simples algorithme d'apprentissage.
- **Objectif**: classer les exemples non étiquetés sur la base de leur similarité avec les exemples de la bases d'apprentissage.

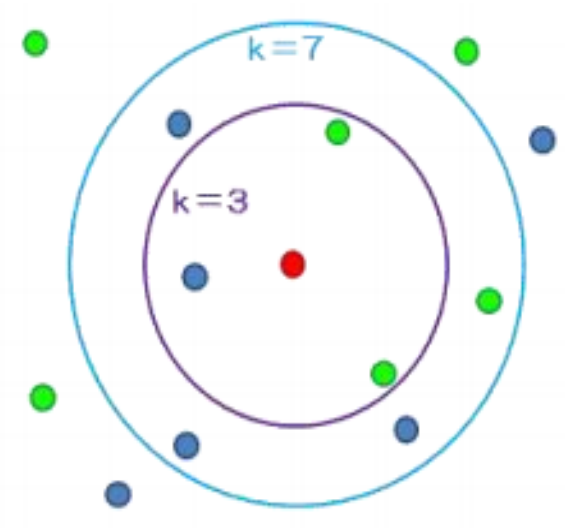
  ⇔ K-NN est parfois catégorisé dans le **Lazy Learning**.

K-NN: Définition

- Pour effectuer une prédiction, l'algorithme K-NN ne va pas calculer un modèle prédictif à partir d'un ensemble d'apprentissage.
- K-NN n'a pas besoin de construire un modèle prédictif.
- Pour pouvoir effectuer une prédiction, K-NN **se base sur le jeu de données** pour produire un résultat.

K-NN: Définition

- ¨K¨ représente le nombre d'éléments de l'ensemble de données qui sont considérés pour la classification.
- **Exemple**: l'image montre la classification pour différentes valeurs k



Comment K-NN effectue une prédiction ?

- K-NN **se base sur le jeu de données en entier**.
- Pour une observation, qui ne fait pas parti du jeu de données, l'algorithme va chercher les **K instances du jeu de données les plus proches de notre observation.**
- Ensuite pour ces **K voisins**, l'algorithme se basera sur leurs variables de sortie *y (prédiction) pour calculer la valeur de la variable y de l'observation qu'on souhaite prédire.*

Comment K-NN effectue une prédiction ?

- Si K-NN est utilisé pour la régression, c'est la moyenne (ou la médiane) des variables y des K plus proches observations qui servira pour la prédiction.
- Si K-NN est utilisé pour la classification, c'est le mode des variables y des K plus proches observations qui servira pour la prédiction.

Ecriture algorithmique

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 :

Ecriture algorithmique

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

Fin Algorithme

Calcul de similarité

- K-NN a besoin d'une fonction de calcul de distance entre deux observations.
- Plus deux points sont proches l'un de l'autre, plus ils sont similaires et vice versa.
- Il existe plusieurs fonctions de calcul de distance:
  - La distance euclidienne,
  - La distance de Manhattan,
  - La distance de Minkowski
  - La distance de Hamming

Calcul de similarité

- La fonction de distance est choisie en fonction des types de données manipulés.
- La distance commune pour des variables **continues** est la **distance euclidienne**.
- Pour des variables **discrètes**, comme en classification de texte, une autre distance peut être utilisée, telle que la **distance de Hamming**.

La distance euclidienne

- Distance qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points

$$D_e(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

- **Exemple:** X={-2, 2}, Y ={2, 5}

$$D_e(x, y) = \sqrt{(-2-2)^2+(2-5)^2} = \sqrt{16+9} = 5$$

La distance Manhattan

- la distance de Manhattan: calcule la somme des valeurs absolues des différences entre les coordonnées de deux points :

$$D_m(x, y) = \sum_{j=1}^{n} |x_j - y_j|$$

- Exemple: X={1, 2}, Y ={2, 5}

$$D_m(x, y) = |1 - 2| + |2 - 5| = 1 + 3 = 4$$

Distance Hamming

- La distance entre deux points données est la différence maximale entre leurs coordonnées sur une dimension.

$$D_h(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

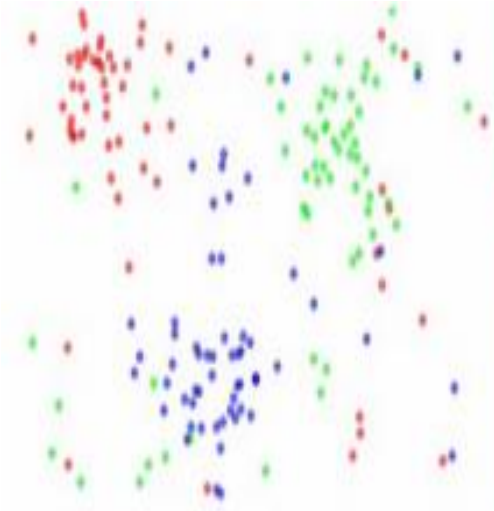
- Avec:
  - $x = y \Rightarrow D = 0$
  - $x \neq y \Rightarrow D = 1$

Comment choisir la valeur K ?

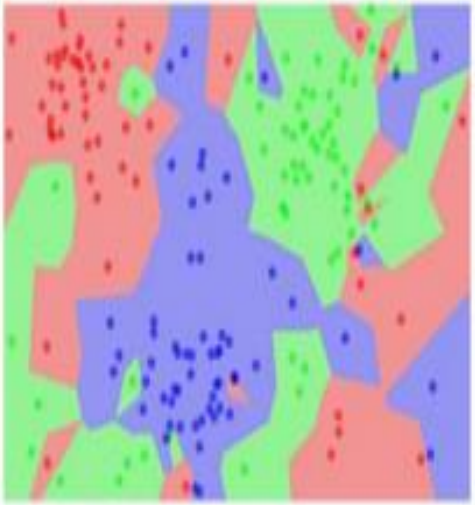
- Le choix de la valeur K varie en fonction du jeu de données.
- *Règle générale:*
  - moins on utilisera de voisins (K petit) plus on sera sujette au sous apprentissage (*underfitting*).
  - Plus on utilise de voisins (K grand), plus sera fiable dans notre prédiction.
- si $K=N$ et N étant le nombre d'observations, on risque d'avoir du *overfitting*, ie. mauvaise généralisation sur des observations qu'il n'a pas encore vu.

Comment choisir la valeur K ?

Les données



N-NN Classifier



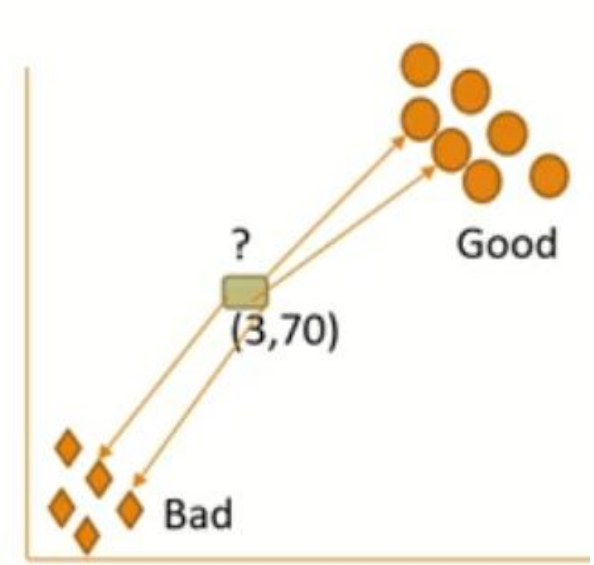
5-NN Classifier

Comment choisir la valeur K ?

- **Pour 5-NN:** les limites entre chaque région sont assez lisses et régulières.
- **Pour N-NN:** les limites sont irrégulières, car l'algorithme tente de faire rentrer tous les points bleus dans les régions bleues, les rouges avec les rouges etc… c'est un cas d'**overfitting**

⇔ Préférer le 5-NN classifier sur le NN-Classifier. Car le 5-NN classifier se généralise mieux que son opposant.

Exemple: classification



Nom	Cigarette	Poids	Crise cardiaque
A	7	70	Oui
B	7	40	Oui
C	3	40	Non
D	1	40	Non

Exemple: classification

- Classer la nouvelle observation ayant les valeurs:

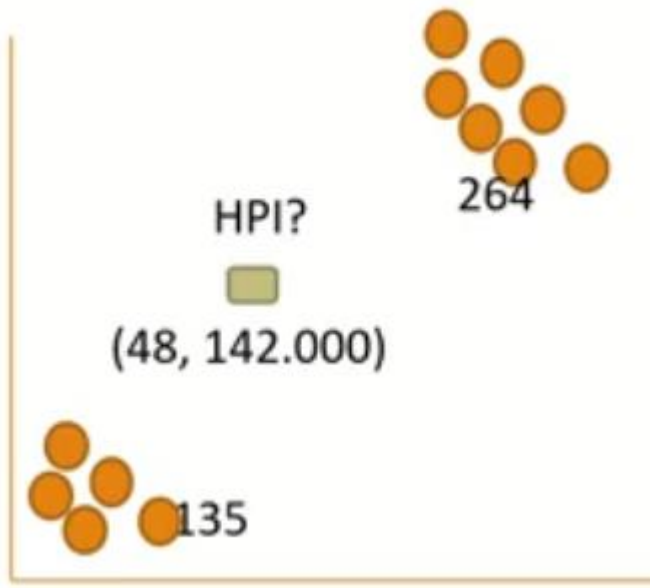
  (Cigarette=3 , Poids=70)

- Utiliser la distance euclidienne pour mesurer la distance avec chaque jeu de données.
- Considérer la valeur: k=1.

Exemple: classification

Nom	Cigarette	Poids	Crise cardiaque	Distance
A	7	70	Oui	$\sqrt{(3-7)^2+(70-70)^2}=4$
B	7	40	Oui	$\sqrt{(3-7)^2+(70-40)^2}=30{,}27$
C	3	40	Non	$\sqrt{(3-3)^2+(70-40)^2}=30$
D	1	40	Non	$\sqrt{(3-1)^2+(70-40)^2}=30{,}07$
E	3	70	**Oui**	

Exemple: régression



Age	Loan	House Price Index
25	40000	135
35	60000	256
45	80000	231
20	20000	267
35	120000	139
52	18000	150
23	95000	127
40	62000	216
60	100000	139
48	220000	250
33	150000	267

Exemple: régression

- Classer la nouvelle observation ayant les valeurs:

  (Age=48 , Loan=142000)

- Utiliser la distance euclidienne pour mesurer la distance avec chaque jeu de données.
- Considérer la valeur: k=1 et par la suite k=3.

Exemple: régression

- Utilisation d'un ensemble d'apprentissage pour classer un cas inconnu,
- Calculer la distance euclidienne entre (Age=48 , Loan=142000) et chaque jeu de données.
- **Exemple:**
  - Distance entre (48, 142000) et (33, 150000)
  - $D_e(x, y) = \sqrt{(48-33)^2+(142000-150000)^2} = 8000{,}01$

Exemple: régression

Age	Loan	House Price Index	Distance	
25	$40,000	135	102000	
35	$60,000	256	82000	
45	$80,000	231	62000	
20	$20,000	267	122000	
35	$120,000	139	22000	2
52	$18,000	150	124000	
23	$95,000	127	47000	
40	$62,000	216	80000	
60	$100,000	139	42000	3
48	$220,000	250	78000	
33	$150,000	264	8000	1
48	**$142,000**	**?**		

Exemple: régression

- K=1:
  - Le voisin le plus proche est le dernier cas de l'ensemble
  - Donc HPI=264
- K=3:
  - La prédiction de HPI est égale à la moyenne des HPIs pour les trois premiers voisins
  - Donc HPI = (264+139+139)/3
    = 180,7