esprit
Se former autrement
HONORIS UNITED UNIVERSITIES

K-plus proches voisins

UP: GL/BD

Réalisé par : Equipe ML Appliqué

United Nations Educational, Scientific and Cultural Organization
UNESCO Chair "Project-based learning" ESPRIT School of engineering, Tunisia
EUR-ACE® Délivrée par la Commission des Titres d'Ingénieur
CONFERENCE DES GRANDES ÉCOLES
cdio CONCEIVE DESIGN IMPLEMENT OPERATE

Classification: Définition

Définition:

La classification permet de prédire si un élément est membre d'un groupe ou d'une catégorie donnée.

Classes:

Identification de groupes avec des profils particuliers.

Possibilité de décider de l'appartenance d'une entité a une classe.

Caractéristiques de classification:

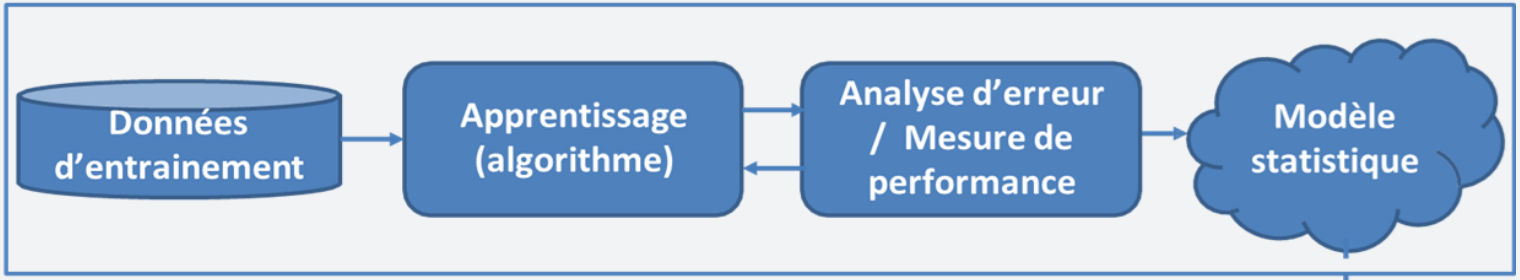
Apprentissage supervisé: classes connues a l'avance.

Qualité de la classification (taux d'erreur).

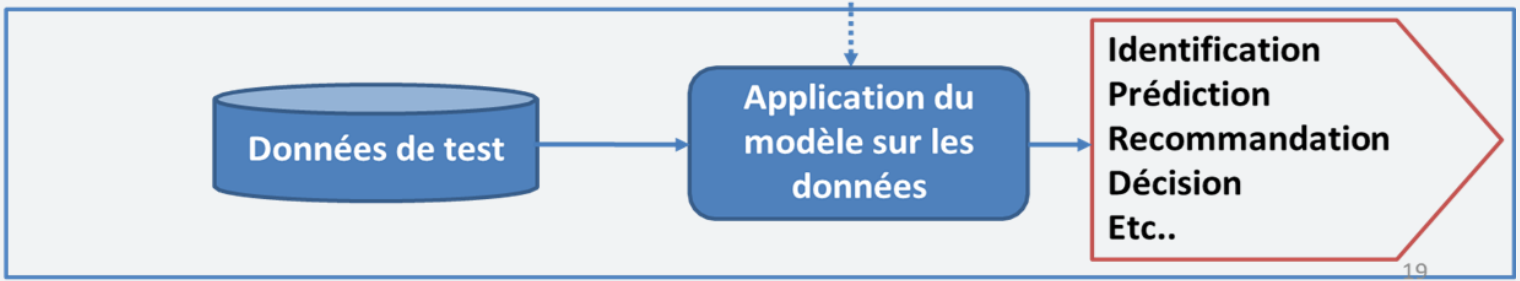
Classification: Procéssus

Le processus d'apprentissage

Phase 1 : Apprentissage



Phase 2 : Inférence/Réalisation de tâche



K- plus proches voisins (KNN)

Problématique: Prédire la catégorie d'un client pour une société de télécommunication.

X: Independent variable

Y: Dependent variable

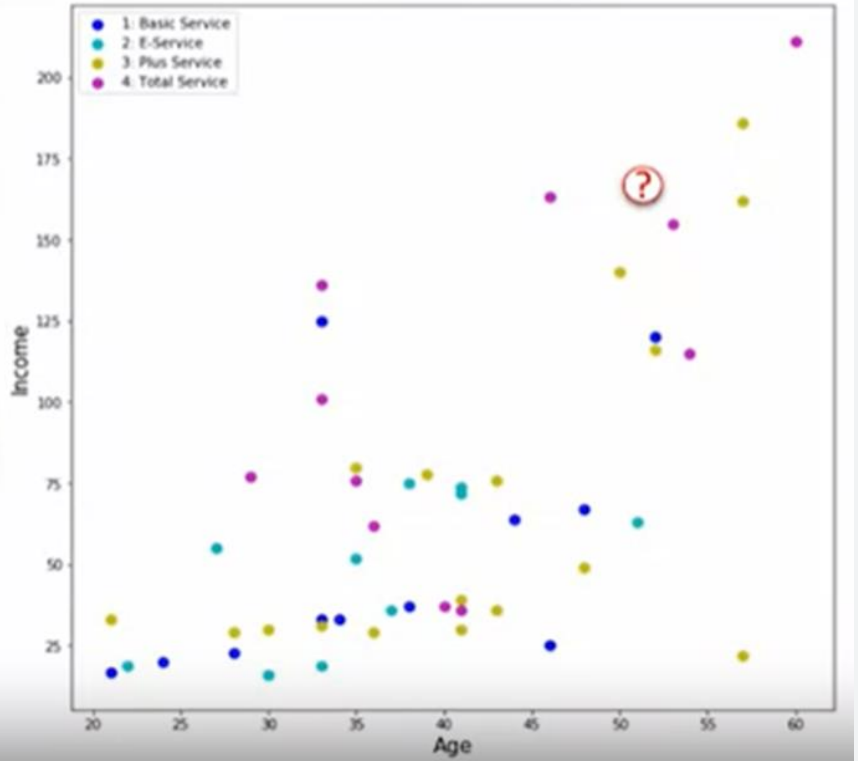
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Intuition

Pour simplifier la compréhension de l'algorithme, nous ne considérerons que deux attributs :
« Age » et “Income”.
Étant donné le graphique ci-dessous, comment pouvons-nous procéder pour identifier la catégorie d’un nouveau client ?

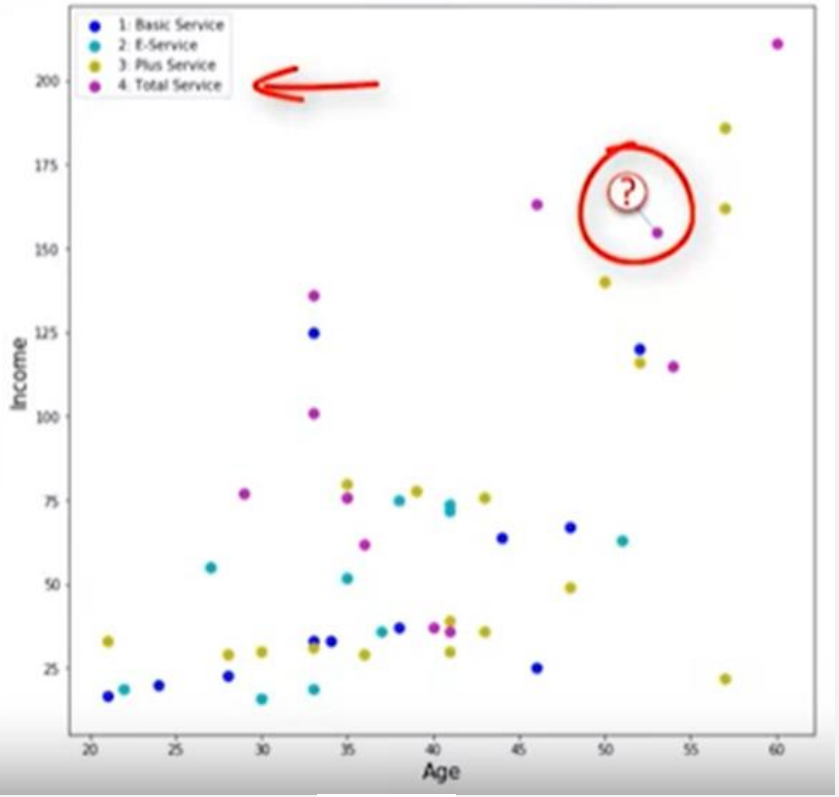
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



Intuition

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

1-NN → 4: Total Service

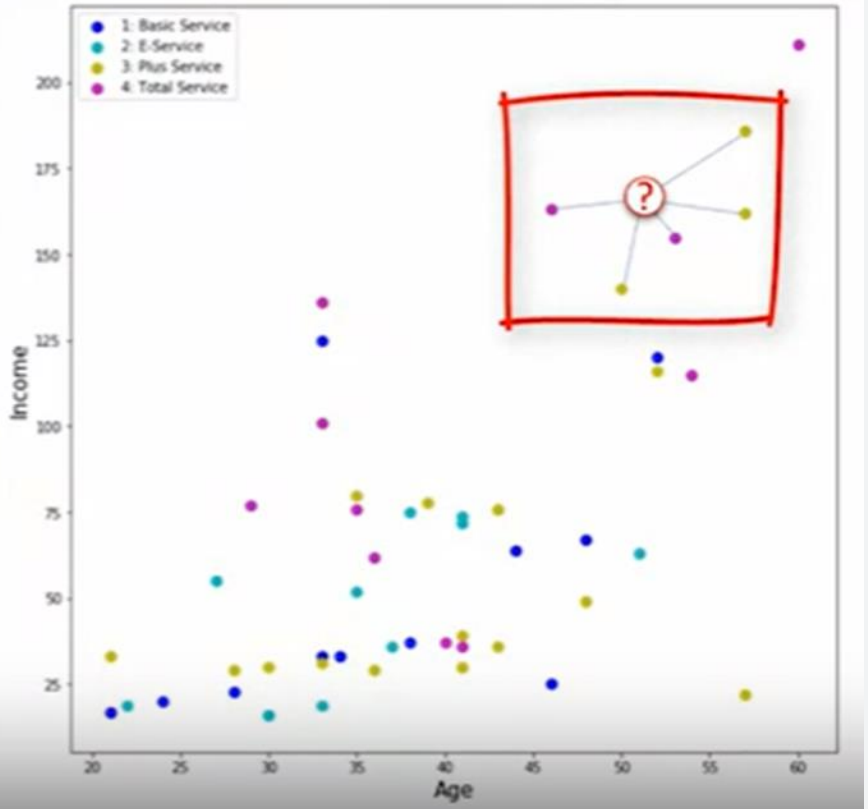


A quel point peut-on peut croire notre jugement basé sur le 1-NN?

Intuition

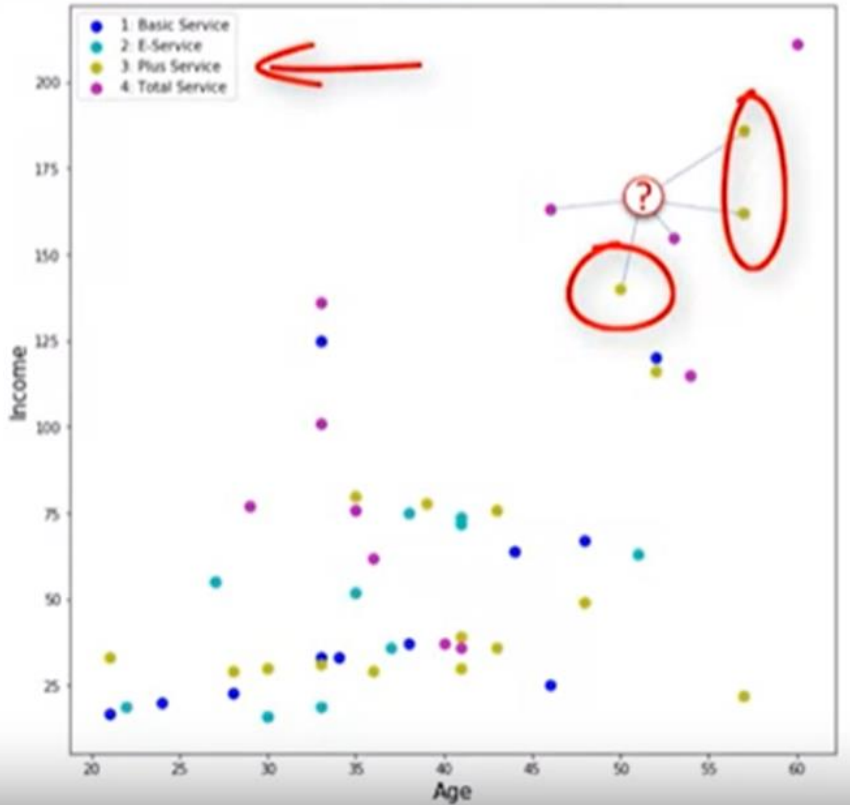
Choisissons K = 5

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



Intuition

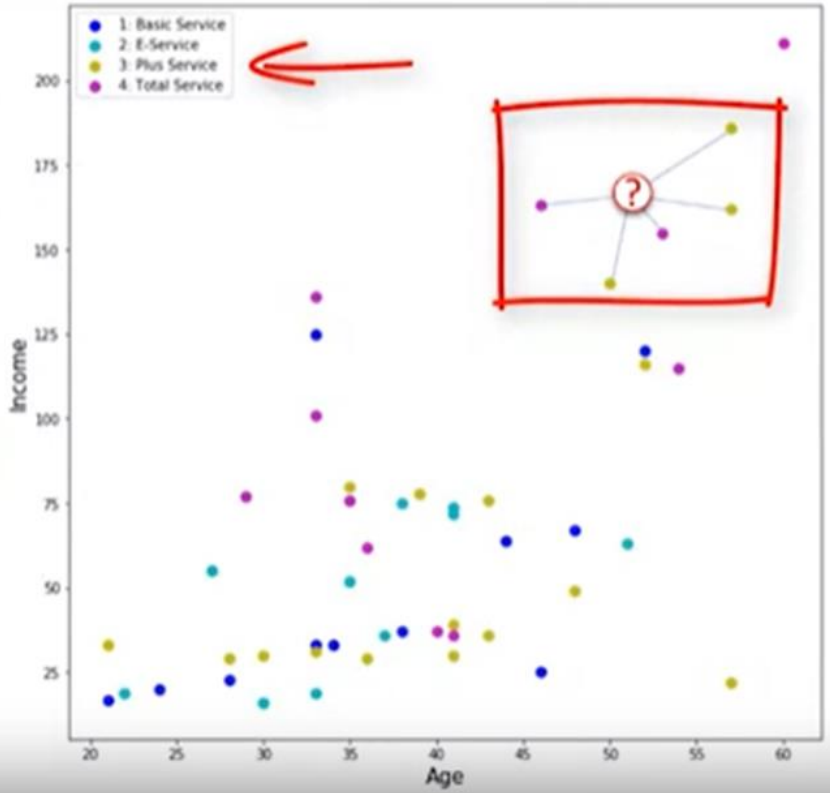
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



Intuition

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

5-NN → 3: Plus Service



KNN, Comment il fonctionne ?

- Prend des points étiquetés et les utilise pour prédire les étiquettes d'autres points.
- Classer un cas en fonction de sa similarité avec d'autres cas
- Dans la méthode KNN, les points proches les uns des autres sont appelés « voisins ».
- KNN est basé sur ce paradigme : les cas similaires ayant la même étiquette sont voisins. En effet, la distance entre deux cas est la mesure de leurs dissimilarité.

KNN Pseudo code:

1- Choisir la valeur de K

2- Calculer la distance de nouveau cas par rapport à tous les cas de jeu de données.

3- Chercher les K observations dans le jeu de données qui sont proches du nouveau cas.

4- Prédire la classe du nouveau cas en utilisant la vote

Exemple de classification

- Supposons que nous ayons un ensemble de données avec deux caractéristiques X1 , X2 et deux classes 0 et 1.

	X1	X2	CLASSE
P1	1	2	0
P2	2	3	0
P3	3	3	1
P4	6	8	1

Nouveau point P5 avec X1= 4 et X2= 5.

- Quelle est sa classe si K = 3?

Problematique ?

Comment calculer la distance ?

Comment choisir la valeur de K?

Comment calculer la distance?

Distance euclidienne

Customer 1		
Age	Income	Education
34	190	3

Customer 2		
Age	Income	Education
30	200	8

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(34-30)^2+(190-200)^2+(3-8)^2} = 11.87$$

Comment choisir K ?

- Si K est trop petit, le modèle sera sensible aux points de bruit.

- Un K plus grand fonctionne bien. Mais un K trop grand peut inclure des points majoritaires des autres classes.

- La règle empirique est K < sqrt(n), n est le nombre d'exemples (échantillons).

Comment choisir K ?

Expérimentez avec différentes valeurs k en utilisant des techniques telles que :

■ Recherche de grille : exécutez une recherche de grille, en testant n_neighbors sur la plage définie.

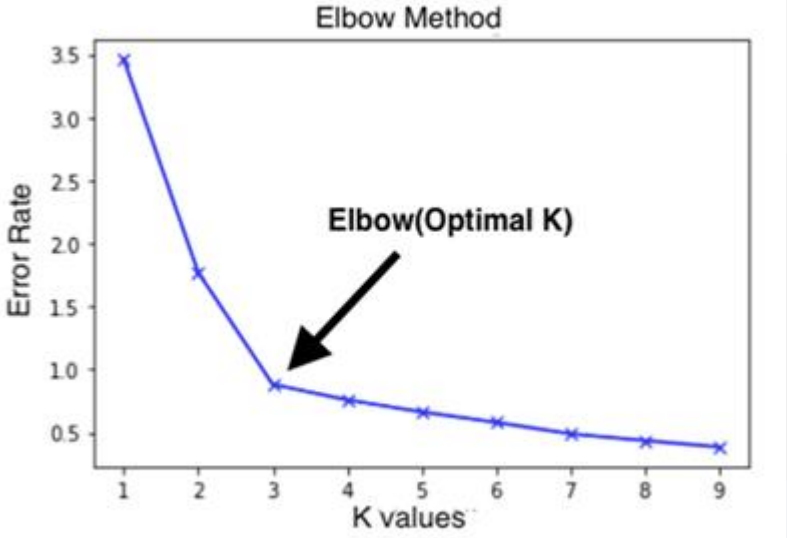
1. Définir la plage de valeurs pour k
2. Créer un modèle KNN
3. La recherche de grille va tester toutes les combinaisons possibles des hyperparamètres spécifiés.
4. Elle utilise la validation croisée pour évaluer chaque combinaison.
5. Exécuter la recherche de grille

Une fois la recherche terminée, vous pouvez identifier la valeur de k qui donne les meilleures performances (par exemple, la plus grande précision ou le plus faible taux d'erreur).

Comment choisir K ?

■ <u>La méthode du coude pour sélectionner le k qui maximise les performances du modèle:</u>

La méthode du coude permet de déterminer la valeur optimale de K en évaluant la performance du modèle (par exemple, l'erreur de classification) pour différentes valeurs de K.



On remarque que l'erreur diminue rapidement jusqu'à K = 3 puis se stabilise. Le "coude" se situe donc autour de K = 3 qui est la valeur optimale.

1. Choisir une plage de valeurs pour K (par exemple, K = 1 à K = 10).
2. Pour chaque valeur de K , entraîner le modèle et calculer l'erreur.
3. Tracer un graphique de l'erreur en fonction de K.
4. Identifier le "coude" dans la courbe, c'est-à-dire le point où l'erreur commence à se stabiliser. Ce point correspond à la valeur optimale de K.

Advantages de KNN

- Simplicité et facilité d'utilisation

- Efficace pour les données non linéaires

- Interprétabilité : possibilité de voir quels points de données ont influencé la prédiction d'un nouveau point de données en examinant les k voisins les plus proches. Cela peut être utile pour comprendre le processus de prise de décision du modèle.

- Rapide à installer et à utiliser (pas de phase de formation).

- Polyvalence : KNN peut être utilisé à la fois pour des tâches de classification et de régression... etc.

Limitations de KNN

- Coût de calcul : trouver les voisins les plus proches pour chaque prédiction peut être coûteux en termes de calcul

- Stockage des données : KNN nécessite le stockage de l'intégralité des données de formation pour la prédiction

- Sensible aux fonctionnalités non pertinentes : les fonctionnalités non pertinentes peuvent fausser les calculs de distance et amener l'algorithme à identifier des voisins les plus proches trompeurs.

- Sensible au bruit

- Le choix de K est crucial.