Chapitre 5

k plus proches voisins (k-ppv)

Zied Elouedi
2018/2019

Plan

- Algorithme standard k-ppv
- Paramètres
  - Ensemble d'apprentissage
  - Choix du k
  - Distance
  - Classification
- Variantes de k-ppv
- Avantages et inconvénients

Introduction

Apprentissage par analogie: Recherche d'un ou de plusieurs cas similaires déjà résolus.

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

Un nouvel objet sera affecté à la classe la plus commune parmi les classes des k objets qui sont les plus proches de lui.

k plus proche voisins (k-ppv)
k nearest neighbors (k-nn)

Technique transductive

- Il y a une seule étape (éventuellement rétirée) et qui permet, au cours de laquelle, chaque individu est classé directement par référence aux autres individus déjà classées.

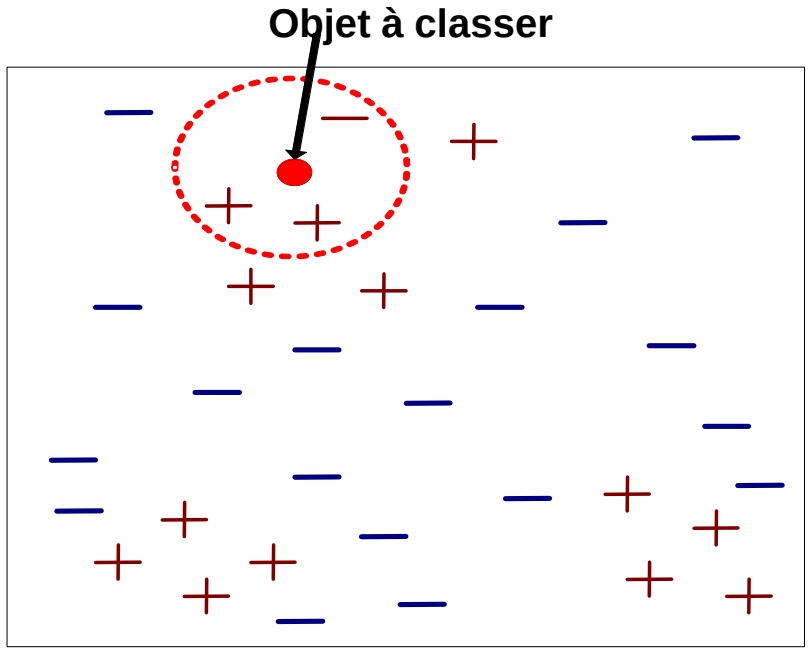
- Pas de construction de modèle à partir des données.

⇒ La méthode k plus proches voisins est une technique transductive.

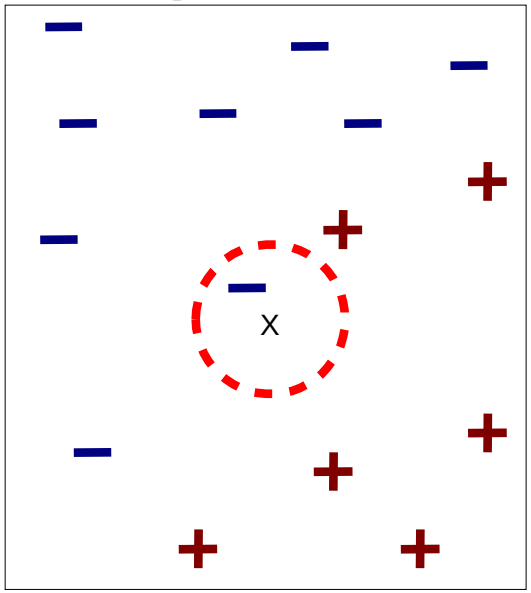
Algorithme k-ppv

Algorithme k-ppv

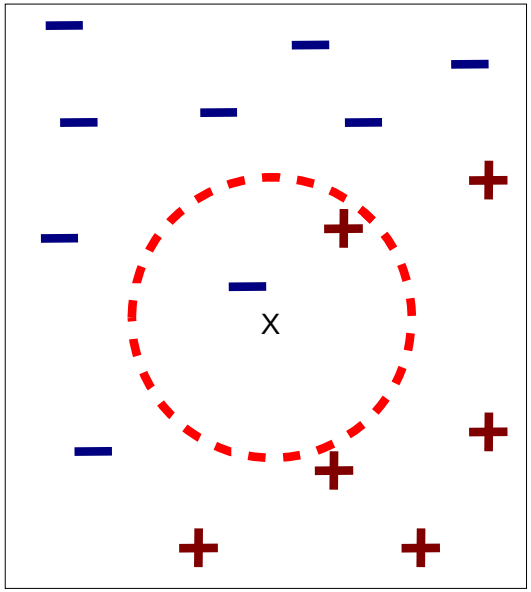
- Pour déterminer la classe d'un nouvel objet O:
  - Calculer la distance entre O et tous les objets de l'ensemble d'apprentissage.
  - Choisir les k objets de l'ensemble d'apprentissage qui sont les plus proches de O.
  - Affecter O à la classe majoritaire parmi les classes des k plus proches voisins.



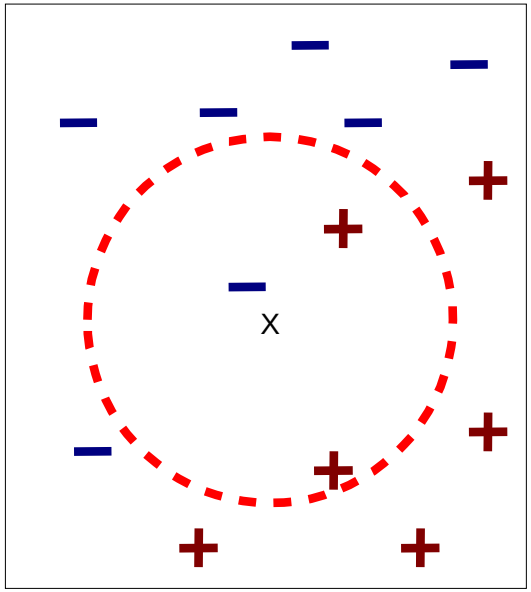
Exemples



(a) 1-plus proche voisin



(b) 2-plus proches voisins



(c) 3-plus proches voisins

Paramètres

Paramètres

- L’ensemble d’apprentissage.
- La métrique de distance pour calculer la distance entre deux objets.
- La valeur de k représentant le nombre des voisins les plus proches.
- Choix de la classe de l’objet à classer.

Ensemble d’apprentissage

- Ensemble d’objets tel que pour chaque objet, on connaît:
  - La valeur de ses attributs.
  - Sa classe.

Distance

- Le choix de la distance est primordial au bon fonctionnement de la méthode.

- Les distances les plus simples permettent d'obtenir des résultats satisfaisants (lorsque c'est possible).

- Propriétés de la distance:
  - Réflexivité: $d(A,B)=0$ SSi $A = B$
  - Non négativité: $d(A, B) \geq 0$
  - Symétrie: $d(A,B)= d(B,A)$
  - Inégalité triangulaire: $d(A,B) \leq d(A,C) + d(B,C)$

Distance Euclidienne

- L'une des distances utilisées quand les attributs sont numériques est la distance Euclidienne.

  - La distance euclidienne entre deux objets $O1=(x_{11}, x_{12}, x_{13},\ldots x_{1p})$ et $O2 =(x_{21},x_{22}, x_{23},\ldots x_{2p})$ est définie comme suit:

$$d(O1,O2) = \sqrt{\sum_{i=1}^{p}(x_{1i} - x_{2i})^2}$$

Distance Euclidienne pondérée

$$d(O1, O2) = \sqrt{\sum_{i=1}^{p} w_i (x_{1i} - x_{2i})^2}$$

Remarque: il y a plusieurs distances à utiliser comme celles utilisées en clustering (Minkowski, Manhattan, etc).

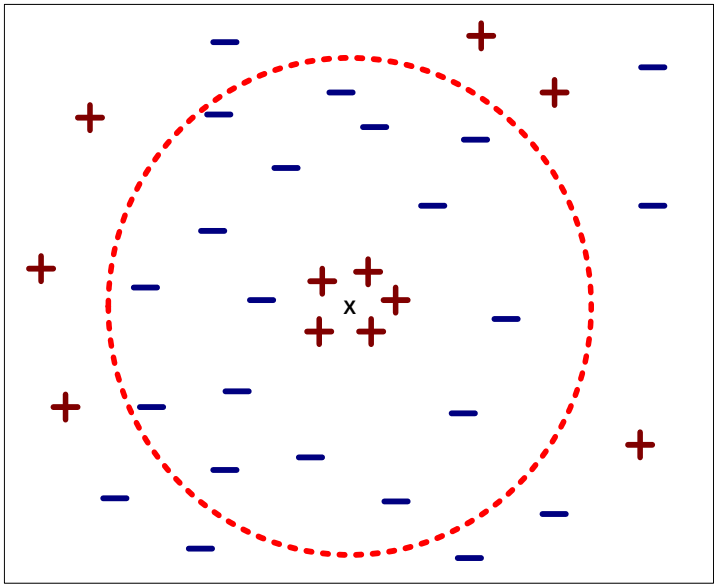
Distance selon les variables

- Variables numériques
- Variables catégoriques
- Variables binaires
- Variables ordinales

Il suffit d’utiliser la distance appropriée à chaque type de variable.

Choix du k

- Si k est trop petit, sensible au bruit.
- Si k est trop grand, le voisinage peut contenir des objets de plusieurs classes.



- Choix du nombre k de voisins peut être déterminé par l'utilisation d'un ensemble test ou par validation croisée.
- Une heuristique fréquemment utilisée est de prendre k égal au nombre d'attributs + 1.

Classification

- Pour déterminer la classe à partir de la liste des k plus proches voisins:
  - Choix de la classe majoritaire.
  - Choix de la classe majoritaire pondérée:
    Chaque classe d'un des k voisins sélectionnés est pondéré.
    Par exemple $w = 1/d^2$.

Cas d'égalité

- **Quelle décision prendre en cas d'égalité ?**
  - Augmenter la valeur de *k* de 1 pour trancher. L'ambiguité peut persister.
  - Choisir au hasard la classe parmi les classes ambigues.
  - Pondération des exemples par leur distance au point *x.*

Complexité

- La complexité de l'algorithme naïf appliquant la règle des *k*-ppv est de O(*kdn*).
  - *d* est la dimensionnalité de l'espace (nombre d'attributs).
  - *n* est le nombre d'échantillons.

Remarques

- Pas de construction de modèle

C'est l'échantillon d'apprentissage, associé à une fonction de distance et d'une fonction de choix de la classe en fonction des classes des voisins les plus proches, qui constitue le modèle.

Exemple

Exemple (1)

Client	Age	Revenu	Nombre cartes de crédit	Classe (Réponse)
Mohamed	35	350	3	Non
Ali	22	500	2	Oui
Samia	63	2000	1	Non
Sami	59	1700	1	Non
Meriem	25	400	4	Oui
Lotfi	37	500	2	?

Exemple (2)

Client	Age	Revenu	Nombre cartes de crédit	Classe (Réponse)	Distance(Client, Lotfi)
Mohamed	35	350	3	Non	Sqrt((35-37)^2+(350-500)^2+(3-2)^2)=**150.01**
Ali	22	500	2	Oui	Sqrt((22-37)^2+(500-500)^2+(2-2)^2)= **15**
Samia	63	2000	1	Non	Sqrt((63-37)^2+(2000-500)^2+(1-2)^2)=**1500.22**
Sami	59	1700	1	Non	Sqrt((59-37)^2+(1700-500)^2+(1-2)^2)=**1200.2**
Meriem	25	400	4	Oui	Sqrt((25-37)^2+(400-500)^2+(4-2)^2)=**100.74**
Lotfi	37	500	2	?	

Exemple (3)

Client	Age	Revenu	Nombre cartes de crédit	Classe (Réponse)	Distance(Client, Lotfi)
Mohamed	35	350	3	Non	Sqrt((35-37)^2+(350-500)^2+(3-2)^2)=**150.01**
Ali	22	500	2	Oui	Sqrt((22-37)^2+(500-500)^2+(2-2)^2)= **15**
Samia	63	2000	1	Non	Sqrt((63-37)^2+(2000-500)^2+(1-2)^2)=**1500.22**
Sami	59	1700	1	Non	Sqrt((59-37)^2+(1700-500)^2+(1-2)^2)=**1200.2**
Meriem	25	400	4	Oui	Sqrt((25-37)^2+(400-500)^2+(4-2)^2)=**100.74**
Lotfi	37	500	2	Oui	?

Il faut normaliser puis calculer les distances

Normalisation des variables

On va supposer que la valeur minimale d'âge, revenu et Nombre de cartes de crédits est 0 et que la valeur maximale d'âge est 63, celle de revenu est 2000 et celle du nombre de cartes de crédits est 4

Client	Age	Revenu	Nombre cartes de crédit	Classe (Réponse)
Mohamed	0.56	0.18	0.75	Non
Ali	0.35	0.25	0.5	Oui
Samia	1	1	0.25	Non
Sami	0.94	0.85	0.25	Non
Meriem	0.4	0.2	1	Oui
Lotfi	0.59	0.25	0.5	?

Exemple (4)

Client	Age	Revenu	Nombre cartes de crédit	Classe (Réponse)	Distance(Client, Lotfi)
Mohamed	0.56	0.18	0.75	Non	$\text{Sqrt}((0.56-0.59)^2+(0.18-0.25)^2+(0.75-0.5)^2)=\mathbf{0.26}$
Ali	0.35	0.25	0.5	Oui	$\text{Sqrt}((0.35-0.59)^2+(0.25-0.25)^2+(0.5-0.5)^2)=\mathbf{0.24}$
Samia	1	1	0.25	Non	$\text{Sqrt}((1-0.59)^2+(1-0.25)^2+(0.25-0.5)^2)=\mathbf{0.89}$
Sami	0.94	0.85	0.25	Non	$\text{Sqrt}((0.94-0.59)^2+(0.85-0.25)^2+(0.25-0.5)^2)=\mathbf{0.74}$
Meriem	0.4	0.2	1	Oui	$\text{Sqrt}((0.4-0.59)^2+(0.2-0.25)^2+(1-0.5)^2)=\mathbf{0.54}$
Lotfi	0.59	0.25	0.5	?	

Exemple (5)
k = 3

Client	Age	Revenu	Nombre cartes de crédit	Classe (Réponse)	Distance(Client, Lotfi)
Mohamed	0.56	0.18	0.75	Non	Sqrt((0.56-0.59)^2+(0.18-0.25)^2+(0.75-0.5)^2)=**0.26**
Ali	0.35	0.25	0.5	Oui	Sqrt((0.35-0.59)^2+(0.25-0.25)^2+(0.5-0.5)^2)= **0.24**
Samia	1	1	0.25	Non	Sqrt((1-0.59)^2+(1-0.25)^2+(0.25-0.5)^2)=**0.89**
Sami	0.94	0.85	0.25	Non	Sqrt((0.94-0.59)^2+(0.85-0.25)^2+(0.25-0.5)^2)=**0.74**
Meriem	0.4	0.2	1	Oui	Sqrt((0.4-0.59)^2+(0.2-0.25)^2+(1-0.5)^2)=**0.54**
Lotfi	0.59	0.25	0.5	Oui	

Avantages et inconvénients

Avantages

☺ Simple et facile à implémenter et à utiliser.

☺ Compréhensible : La classification est facile à expliquer.

☺ Robuste aux données bruitées.

☺ Efficace pour des classes réparties de manière irrégulière.

☺ Des applications intéressantes.

Inconvénients

☹ Nécessité de capacité de stockage et de puissance de calcul.

☹ Pas de modèle construit.

☹ Prend du temps pour classer un nouvel objet:
Comparaison des distances du nouvel objet avec tous les autres de l'ensemble d'apprentissage.

☹ Choix du k.

Travail à faire

Une usine fait une étude sur le statut de ses employés s'ils sont titulaires ou contractuels. Chaque employé de l'ensemble d'apprentissage ci-joint est caractérisé par trois attributs à savoir l'âge, le salaire mensuel et le genre permettant de déterminer son statut.

Employé	Age	Salaire mensuel	Genre	Statut
E1	27	190	F	Contractuel
E2	51	640	M	Titulaire
E3	52	1000	M	Titulaire
E4	33	550	F	Titulaire
E5	45	450	M	Contractuel

1) Transformer l'ensemble d'apprentissage en un ensemble normalisé. Pour l'attribut Genre remplacer F par 1 et M par 0. Cet attribut sera désormais considéré comme un attribut numérique.

2) Soit l'ensemble test composé par les employés E6 et E7:

Employé	Age	Salaire mensuel	Genre	Statut
E6	32	310	M	?
E7	42	700	F	?

Appliquer l'algorithme standard de k plus proche voisins (k = 3) pour classer les employés E6 et E7.

Solution (1)

1) La normalisation donne

Employé	Age	Salaire mensuel	Genre	Statut
E1	0	0	1	Contractuel
E2	0,96	0,56	0	Titulaire
E3	1	1	0	Titulaire
E4	0,24	0,44	1	Titulaire
E5	0,72	0,32	0	Contractuel

Employé	Age	Salaire mensuel	Genre	Statut
E6	0,2	0,15	0	?
E7	0,6	0,63	1	?

Solution (2)

2) **D(E6, E1) = Sqrt((0,2)*2+ (0,15)*2 +(1)*2)= 1,03**
D(E6, E2) = Sqrt((0,76)*2+ (0,41)*2 +(0)*2)= 0,86
D(E6, E3) = Sqrt((0,8)*2+ (0,85)*2 +(0)*2)= 1,17
D(E6, E4) = Sqrt((0,04)*2+ (0,31)*2 +(1)*2)= 1,05
D(E6, E5) = Sqrt((0,52)*2+ (0,17)*2 +(1)*2)= 0,55

Donc E6 contractuel

D(E7, E1) = Sqrt((0,6)*2+ (0,63)*2 +(0)*2)= 0,87
D(E7, E2) = Sqrt((0,36)*2+ (0,07)*2 +(1)*2)= 1,07
D(E7, E3) = Sqrt((0,4)*2+ (0,37)*2 +(1)*2)= 1,14
D(E7, E4) = Sqrt((0,36)*2+ (0,19)*2 +(0)*2)= 0,41
D(E7, E5) = Sqrt((0,12)*2+ (0,31)*2 +(1)*2)= 1,05

Donc E7 contractuel

Conclusion (1)

- k-ppv est une méthode de classification non-paramétrique puisqu'aucune estimation de paramètres n'est nécessaire, pas comme pour la régression linéaire.
- Tous les calculs doivent être effectués lors de la classification (pas de construction de modèle).
- Le modèle est l'échantillon: Espace mémoire important nécessaire pour stocker les données, et méthodes d'accès rapides nécessaires pour accélérer les calculs.
- Les performances de la méthode dépendent du choix de la distance, du nombre de voisins et du mode de combinaison des réponses des voisins.
- La méthode permet de traiter des problèmes avec un grand nombre d'attributs. Cependant, plus le nombre d'attributs est important, plus le nombre d'exemples doit être grand.

Conclusion (2)

- Plusieurs extensions de k plus proches voisins:
  - Système de classification hybrids.
  - Fuzzy k-NN.
  - Belief k-NN.
  - Possibilistic k-NN.

.
.
.

TD

Exercice

Une usine fait une étude sur le statut de ses employés s'ils sont titulaires ou contractuels. Chaque employé de l'ensemble d'apprentissage ci-joint est caractérisé par trois attributs à savoir l'âge, le salaire mensuel et le genre permettant de déterminer son statut.

Employé	Age	Salaire mensuel	Genre	Statut
E1	27	190	F	Contractuel
E2	51	640	M	Titulaire
E3	52	1000	M	Titulaire
E4	33	550	F	Titulaire
E5	45	450	M	Contractuel

1) Transformer l'ensemble d'apprentissage en un ensemble normalisé. Pour l'attribut Genre remplacer F par 1 et M par 0. Cet attribut sera désormais considéré comme un attribut numérique.

2) Soit l'ensemble test composé par les employés E6 et E7:

Employé	Age	Salaire mensuel	Genre	Statut
E6	32	310	M	?
E7	42	700	F	?

Appliquer l'algorithme standard de k plus proche voisins (k = 3) pour classer les employés E6 et E7.

Exercice 2

Tester la méthode k-plus proches voisins (IBK sur Weka) en utilisant Weka sur la base labor en faisant varier:

- k (tester k=1, k=2, k=3, k=5)

- Validation croisée (10 folds, 5 folds)

Comparer les résultats selon le pourcentage de classification correcte.

Exercice 2

Validation croisée (10 folds)
K=1, PCC = 82,45%
K=2, PCC = 84,21%
K=3, PCC = 91,22%
K =5, PCC = 85,96%

5 Validation croisée (5 folds)
K=1, PCC = 89,47%
K=2, PCC = 85,96%
K=3, PCC = 91,22%
K =5, PCC = 85,96%

Meilleure résultat avec k = 3 (pour 10 et 5 folds).