Cours 4 – LA METHODE DES K PLUS PROCHES VOISINS

Présentation de la méthode

Méthode des K plus proches voisins (KPPV)

En Anglais, K Nearest Neighbours (KNN)

Méthode d'apprentissage supervisée

classification ou régression

Principe

Se baser sur des cas similaires au cas à résoudre pour faire la prédiction

Description de la méthode

Echantillon d'apprentissage

variables explicatives Xj

variable à expliquer Y

Y(ωi) est égal à yi pour chaque individu ωi

	X1	...	Xj	...	XP	Y
ω1	X1(ω1)	...	Xj(ω1)	...	XP(ω1)	y1
ω2	X1(ω2)	...	Xj(ω2)	...	XP(ω2)	y2
...	...	...	...	...	...	...
ωi	X1(ωi)	...	Xj(ωi)	...	XP(ωi)	yi
...	...	...	...	...	...	...
ωN	X1(ωN)	...	Xj(ωN)	...	XP(ωN)	yN

Problème

Pour un nouvel individu ω

prédire (trouver) y , c'est à dire Y(ω).

ω	X1(ω)	...	Xj(ω)	...	XP(ω)	?

Solution

- Trouver les K plus proches voisins de ω

c'est-à-dire les K individus qui lui ressemblent le plus

On peut prendre K égal à la racine carrée de N

=> Calcul des distances entre ω et les individus de l'échantillon

- Ensuite calculer y en se basant sur les valeurs yi de ces voisins.

Distance entre deux individus

Distance euclidienne

$$D(\omega, \omega i) = \sqrt{\sum_{i=1}^{P} [d(Xj(\omega), Xj(\omega i))]^2}$$

d(Xj(ω),Xj(ωi)) distance par rapport à Xj

Si Xj est de type quantitatif

d(Xj(ω),Xj(ωi)) = Xj(ω) - Xj(ωi)

Si Xj est de type qualitatif nominal

d(Xj(ω),Xj(ωi)) = 0 si Xj(ω) est égal à Xj(ωi)

=1 sinon

Si Xj est de type qualitatif ordinal

d(Xj(ω),Xj(ωi)) = 0 si Xj(ω) est égal à Xj(ωi)

= |Xj(ω) -Xj(ωi) | / Card(Dj) sinon

Par exemple si Xj est la variable Mention qui est ordinale

Dj = {Passable, ABien, Bien, TBien} = {0, 1, 2, 3}

Si Xj(ω) = Passable et Xj(ωi) = Passable

=> d(Xj(ω),Xj(ωi)) = 0

Si Xj(ω) = Passable et Xj(ωi) = Bien

=> d(Xj(ω),Xj(ωi)) = |0 - 2 | / 4 = 0.5

Normalisation des variables quantitatives

Objectif : éviter que la distance dépende principalement de la variable qui a des valeurs plus élevées que les autres.

Normalisation min-max

Transformer les Xj(ωi) en des valeurs $\widehat{Xj}(\omega i)$ entre 0 et 1.

$$\widehat{Xj}(\omega i)=\frac{Xj(\omega i)-min(Xj)}{max(Xj)-min(Xj)}$$

min(Xj) et max(Xj) : minimum et maximum des valeurs de Xj

Cas de la classification

Vote majoritaire

la classe y de ω est égal à la classe majoritaire parmi les classes { y1 , y2 , ..., yK } de ses plus proches voisins.

Vote pondéré

vote majoritaire pondéré

le poids pi de la classe yi d'un voisin ωi est égal à 1 / D(ω , ωi)

Cas de la régression

moyenne simple

y est des valeurs { y1 , y2 , ..., yK } de ses voisins.

moyenne pondérée

y est la moyenne pondérée des valeurs { y1 , y2 , ..., y K } de ses voisins.

Exemple d'application 1

Soit l'ensemble d'apprentissage suivant dans lequel age est une variable quantitative et income, credit_rating, student sont des variables qualitatives considérées comme non ordonnées.

	age	income	student	credit_rating	buys_computer
ω1	30	high	no	fair	no
ω2	30	high	no	excellent	no
ω3	40	high	no	fair	yes
ω4	50	medium	no	fair	ves
ω5	50	low	ves	fair	no
ω6	50	low	ves	excellent	no
ω7	40	low	ves	excellent	ves
ω8	30	medium	no	fair	no
ω9	30	low	ves	fair	ves
ω10	50	medium	ves	fair	ves
ω11	30	medium	ves	excellent	ves
ω12	40	medium	no	excellent	ves
ω13	40	high	yes	fair	no
ω14	50	medium	no	excellent	no

En utilisant la méthode KNN, avec K = 5. prédire la valeur de la variable buys_computer pour un nouvel exemple ayant les caractéristiques suivantes :

age = 35, income = medium, student = yes, credit_rating = fair

Solution

Pour le calcul des distances, on utilise la version normalisée de la variable age obtenue avec la normalisation Min-Max.

Il faut donc considérer le nouveau tableau de données suivant :

	age	age*	income	student	credit_rating	buys_computer	D(ω,ωi)	Poids pi
ω1	30	0	high	no	fair	no	1.43	
ω2	30	0	high	no	excellent	no	1.75	
ω3	40	0.5	high	no	fair	yes	1.43	
ω4	50	1	medium	no	fair	yes	1.25	0.8
ω5	50	1	low	yes	fair	no	1.25	0.8
ω6	50	1	low	yes	excellent	no	1.60	
ω7	40	0.5	low	yes	excellent	yes	1.43	
ω8	30	0	medium	no	fair	no	1.75	
ω9	30	0	low	yes	fair	yes	1.43	
ω10	50	1	medium	yes	fair	yes	0.75	1.33
ω11	30	0	medium	yes	excellent	yes	1.03	0.97
ω12	40	0.5	medium	no	excellent	yes	1.43	
ω13	40	0.5	high	yes	fair	no	1.03	0.97
ω14	50	1	medium	no	excellent	no	1.60	

ω	35	0.25	Medium	Yes	fair	?		

L'age normalisé du nouvel exemple à classer est égal à 0.25

$$D(\omega,\omega1) = ((0.25 -0)^2 + (d(medium, high))^2 + (d(yes, no))^2 + (d(fair, fair))^2)^{1/2}$$
$$= (0.25^2 + 1^2 + 1^2 + 0^2)^{1/2}$$
$$= 1.43$$

Même calcul pour les autres distances D(ω,ω2), …, D(ω,ω14)

Vote simple :

yes : 1+1+1 = 3

no : 1 + 1=2

Donc buys_computer = yes pour le nouvel exemple.

Vote pondéré :

yes : 0.8+1.33+0.97 = 3.1

no : 0.8+0.97 = 1.77

Donc buys_computer = yes pour le nouvel exemple

Exemple d'application 1

Soit l'ensemble d'apprentissage suivant dans lequel

age est une variable quantitative,

income est une variable qualitative ordonnée,

credit_rating, student, buys_computer sont des variables qualitatives considérées comme non ordonnées.

	income	student	credit_rating	buys_computer	age
ω1	high	no	fair	no	30
ω2	high	no	excellent	no	30
ω3	high	no	fair	yes	40
ω4	medium	no	fair	ves	50
ω5	low	ves	fair	no	50
ω6	low	ves	excellent	no	50
ω7	low	ves	excellent	ves	40
ω8	medium	no	fair	no	30
ω9	low	ves	fair	ves	30
ω10	medium	ves	fair	ves	50
ω11	medium	ves	excellent	ves	30
ω12	medium	no	excellent	ves	40
ω13	high	yes	fair	no	40
ω14	medium	no	excellent	no	50

En utilisant la méthode KNN, avec K = 5. prédire la valeur de la variable age pour un nouvel exemple ayant les caractéristiques suivantes :

income = medium, student = yes, credit_rating = fair,

buys_computer = yes