

INF5081

Méthode des k plus proches voisins

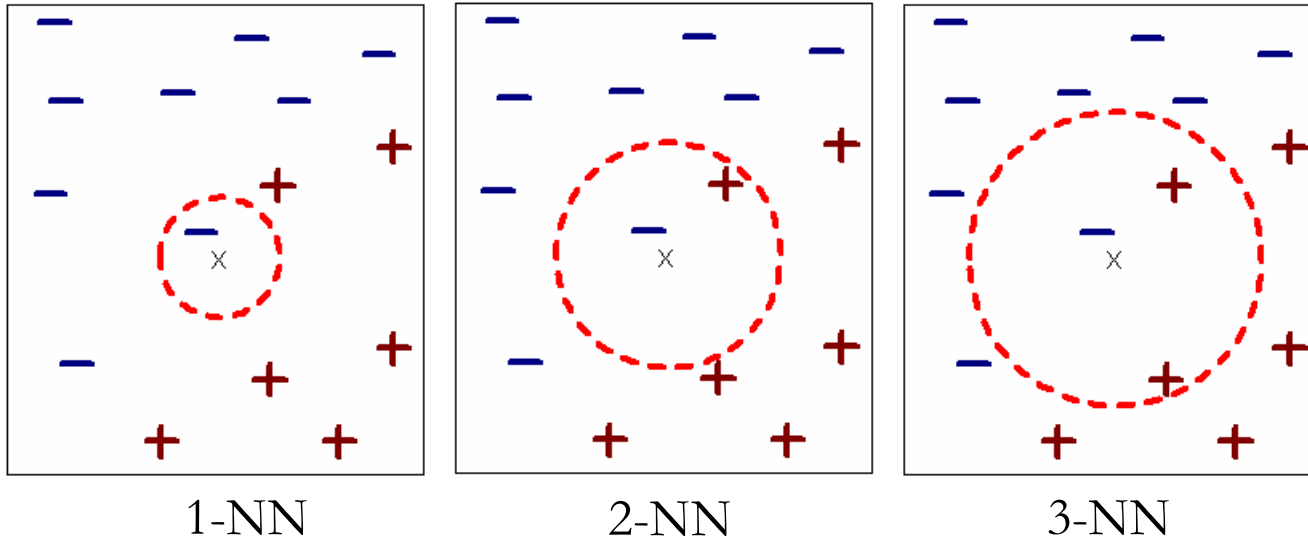
Mohamed Bouguessa



Principe

- Apprendre par analogie
 - Recherche d'un ou des cas similaires déjà résolus
 - « Dis-moi qui sont tes amis, je te dirais qui tu es »
- 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

Notion de k-plus proche voisin (k-NN)









- Les k plus proches voisins du point x sont les points avec la plus petite distance.
- Pour le calcul de la distance, on utilise, généralement, la distance euclidienne

$$D(p, q) = \sqrt{\sum_{k=1}^d (p_k - q_k)^2}$$







Méthode des k plus proches voisins

■ Exemple

Customer	Age	Income	No. credit cards	Loyal
John 	35	35K	3	No
Rachel 	22	50K	2	Yes
Hannah 	63	200K	1	No
Tom 	59	170K	1	No
Nellie 	25	40K	4	Yes
David 	37	50K	2	?

Méthode des k plus proches voisins

- Identification des k plus proche voisin de David
- $k = 3$
- La distance utilisée est la distance euclidienne

Customer	Age	Income	No. credit cards	Loyal	Distance from David
John 	35	35K	3	No	$\text{sqrt} [(35-37)^2+(35-50)^2+(3-2)^2]=15.16$
Rachel 	22	50K	2	Yes	$\text{sqrt} [(22-37)^2+(50-50)^2+(2-2)^2]=15$
Hannah 	63	200K	1	No	$\text{sqrt} [(63-37)^2+(200-50)^2+(1-2)^2]=152.23$
Tom 	59	170K	1	No	$\text{sqrt} [(59-37)^2+(170-50)^2+(1-2)^2]=122$
Nellie 	25	40K	4	Yes	$\text{sqrt} [(25-37)^2+(40-50)^2+(4-2)^2]=15.74$
David 	37	50K	2	Yes	

« Yes » représente la classe dominante des plus proches voisins de David



Méthode des k plus proches voisins

Paramètre :

k – nombre de voisins

Données :

Un échantillon d'apprentissage de N exemples/enregistrements avec leurs classes - La classe d'un exemple x est y)

Entrée :

Un enregistrement z (on doit chercher la classe de z)

Sortie :

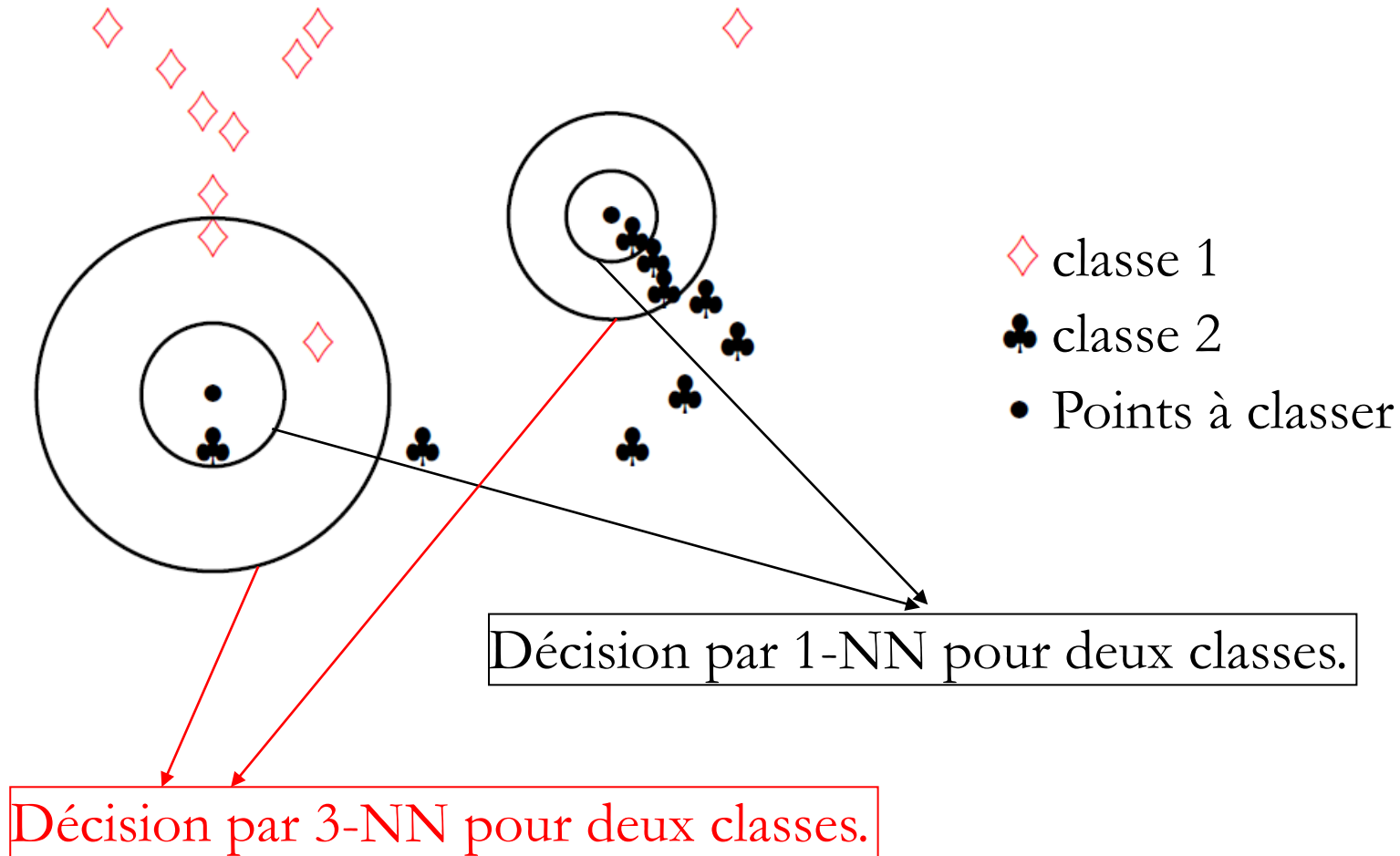
La classe de z est c

Approche :

1. Déterminer les k plus proches exemples de z en calculant les distances.
2. Choisir la classe majoritaire c qui représente les k -plus proches voisins de z .

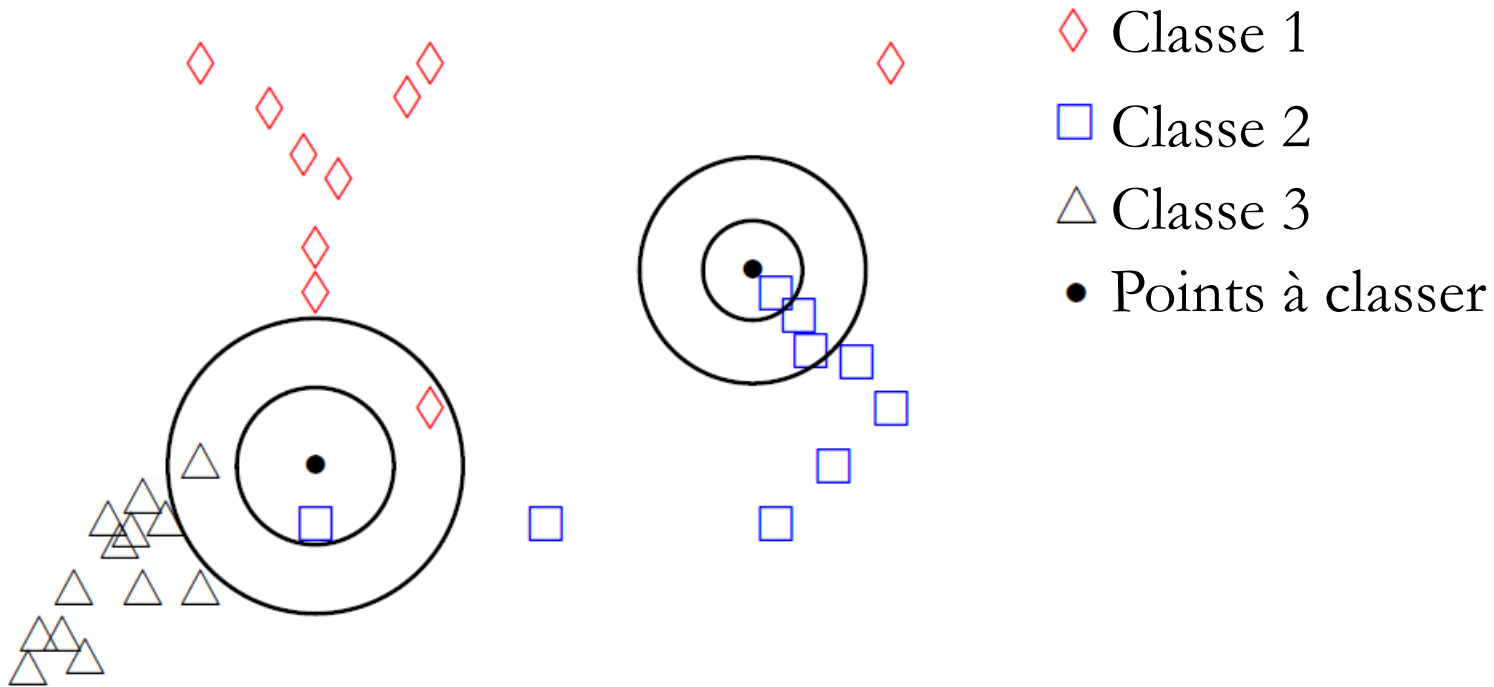
Illustration

Un problème de deux classes



Illustration

Un problème de trois classes



Décision par 1-ppv et 3-ppv pour trois classes.



Choix de la valeur de k

Diverses considérations théoriques et expérimentales mènent à l'heuristique suivante :

$$k \approx \sqrt{n/C}$$

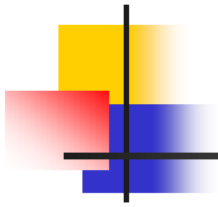
n/C : est le nombre moyen de points d'apprentissage par classe.



Méthode des k plus proches voisins

■ Remarque

- Parfois il est préférable de normaliser les valeurs des attributs en utilisant les méthodes dans les diapositives suivants (méthodes : min-max et Z-score).



Normalisation

- la méthode min-max: normaliser à $[new_min_A, new_max_A]$

➤ Mise à l'échelle pour avoir un petit intervalle spécifié

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

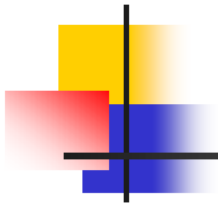
Ancienne valeur (valeur originale)

Nouvelle valeur (valeur normalisée)

Exemple: Intervalle des revenus entre 12 000\$ à 98 000\$ à normaliser entre [0, 1]

Donc la valeur 73 000\$ est transformée

$$\frac{73,000 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$$



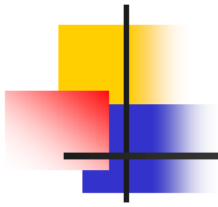
Normalisation

■ **Z-score:** (μ : moyenne, σ : écart type):

➤ Même ordre de grandeurs pour les valeurs des attributs

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Expl. Si $\mu = 54,000$, $\sigma = 16,000$. alors $\frac{73,600 - 54,000}{16,000} = 1.225$



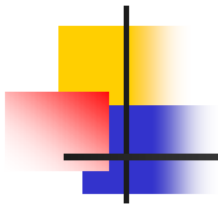
Normalisation

■ Z-score (exemple)

valeur originale

valeur normalisée

v1	v1'			v2	v2'		
0.18	-0.84	Avg	0.68	20	-0.26	Avg	34.3
0.60	-0.14	sdev	0.59	40	0.11	sdev	55.9
0.52	-0.27			5	0.55		
0.25	-0.72			70	4		
0.80	0.20			32	-0.05		
0.55	-0.22			8	-0.48		
0.92	0.40			5	-0.53		
0.21	-0.79			15	-0.35		
0.64	-0.07			250	3.87		
0.20	-0.80			32	-0.05		
0.63	-0.09			18	-0.30		
0.70	0.04			10	-0.44		
0.67	-0.02			-14	-0.87		
0.58	-0.17			22	-0.23		
0.98	0.50			45	0.20		
0.81	0.22			60	0.47		
0.10	-0.97			-5	-0.71		
0.82	0.24			7	-0.49		
0.50	-0.30			2	-0.58		
3.00	3.87			4	-0.55		



k - plus proches voisins - Discussion

- Tous les calculs doivent être effectués lors de la classification : pas de construction de modèle.
- Le modèle est l'échantillon d'apprentissage : cela nécessite l'utilisation
 - d'espace mémoire important pour stocker les données,
 - et des méthodes d'accès rapides pour accélérer les calculs.
- Classifieur sensible au choix de la valeur de k .



k- plus proches voisins - Discussion

- Le k NN est un classifieur qui coute cher en termes de ressource et temps de calcul alors que la classification avec un classifieur à modèle, comme les arbres de décisions, est très rapide une fois le modèle est établi (arbre construit).
- La prédiction avec k NN est basée sur une information locale qui est l'identification des plus proches voisins, alors avec les arbres de décisions un modèle global est construit qui modélise toutes les caractéristiques de l'ensemble de données d'apprentissage.

k-plus proches voisins - Discussion

Un avantage par rapport a un arbre de décision, la bordure de décision (séparation entre les classes) est de forme arbitraire

