Université de Chlef
Département Informatique
Filière : 2ème Année Master - ISIA

Janvier 2016

Examen semestriel

Module "Modèles pour le Datamining"

Durée : 01H30

Corrigé

Exercice 1 (03 points) :

a/ Expliquez le principe d'une classification supervisée et la différence avec une classification non supervisée.

Réponse :

L'objectif de la classification est d'identifier les classes auxquelles appartiennent des objets à partir de traits descriptifs (attributs, caractéristiques, features).

Dans les méthodes supervisées (dites aussi prédictives): les classes sont connues et l'on dispose d'exemples de chaque classe (fourni par un expert).
Dans les méthodes non supervisées (dites exploratoires) : les classes (clusters) ne sont pas définis à l'avance. On ne dispose pas d'exemples d'apprentissage. Les clusters sont construits en calculant la similarité entre les objets.

(1.5 points)

b/ Parmi les algorithmes présentés en cours, précisez ceux qui font partie des méthodes supervisées et ceux qui appartiennent aux méthodes non supervisés.

Réponse :

Algorithmes de la classification supervisée	Algorithmes de la classification non supervisée
SVM *KNN* *Arbres de décision*	*KMeans*

(1.5 points)

Exercice 2 (07 points) :

Le tableau suivant contient des données sur des individus d'une population décrits selon deux attributs : attribut 1 et attribut 2. On souhaite utiliser la méthode SVM pour classer ces données en deux classes : C1 et C2.

Tableau des données

N°	Attribut 1	Attribut 2	Classe
1	0	3	C1
2	3	3	C1
3	1	4	C1
4	2	5	C1
5	1	6	C2
6	2	7	C2
7	3	7	C2
8	3	8	C2
9	2	9	C2
10	4	9	C2
11	8	9	C2

Dr Mourad LOUKAM

Question 1 :

- Représentez sur le plan de la figure 1 les données du tableau précédent, en symbolisant la classe C1 par une croix (+) et la classe C2 par un cercle (o).
- Représentez l'hyperplan séparateur optimal en montrant clairement les supports de vecteur que vous utilisez.

Question 2 :

- On souhaite classer un douzième individu ayant Attribut1 égal à 4 et Attribut2 égal à 5.6. Redessinez tous les points sur la figure 2, en montrant la classe (croix ou cercle) de l'élément qui vient d'être ajouté.
- Représentez le nouvel hyperplan optimal.

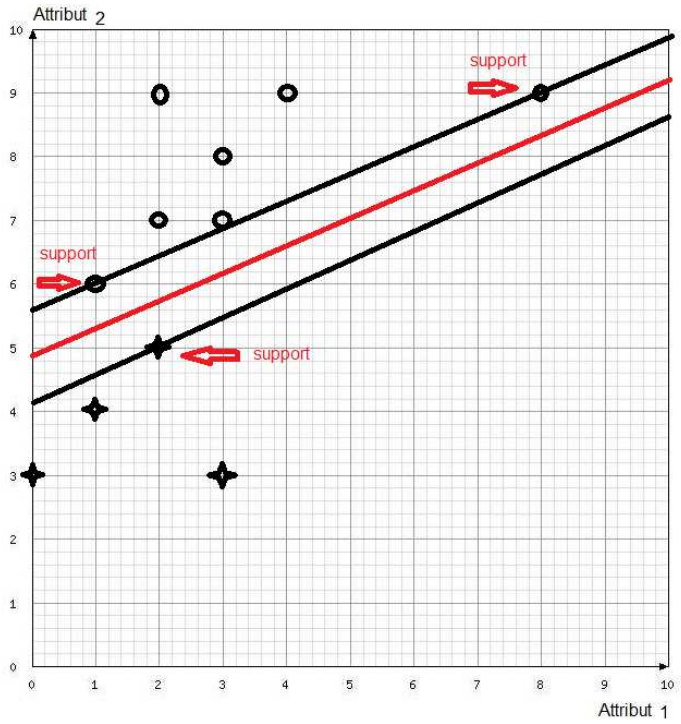

Figure 1

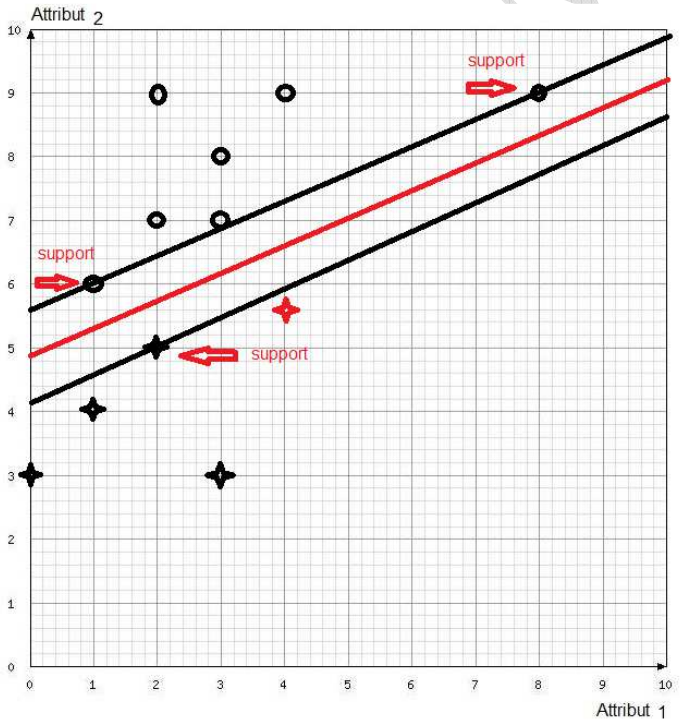

Figure 2

(5 points)

Question 3 : On ajoute un autre point (2, 3) dont on connait la classe, qui est C1.

Les données sont-elles toujours linéairement séparables ? Si oui donnez l'équation du nouvel hyperplan optimal. Sinon, dites ce que prévoit SVM dans ce cas.

Réponse :

Oui, les données sont toujours linéairement séparables. Il y'a toujours un hyperplan optimal H qui sépare les classes 1 et 2. Son équation générale est w. x + b ; w étant le vecteur poids, et b le biais.

(2.5 points)

Exercice 3 (07 points) :

Le tableau suivant contient des données sur des individus d'une population décrits selon deux attributs : attribut 1 et attribut 2. La classe d'un individu peut être : C1, ou C2, ... ou C6.

Tableau des données

N°	Attribut 1	Attribut 2	Classe
1	1	2	C1
2	2	6	C1
3	2	5	C2
4	2	1	C3
5	4	2	C5
6	5	6	C4
7	6	5	C3
8	6	1	C6

Question 1 :

- Représentez sur le plan de la figure 3 suivante les données du tableau précédent.

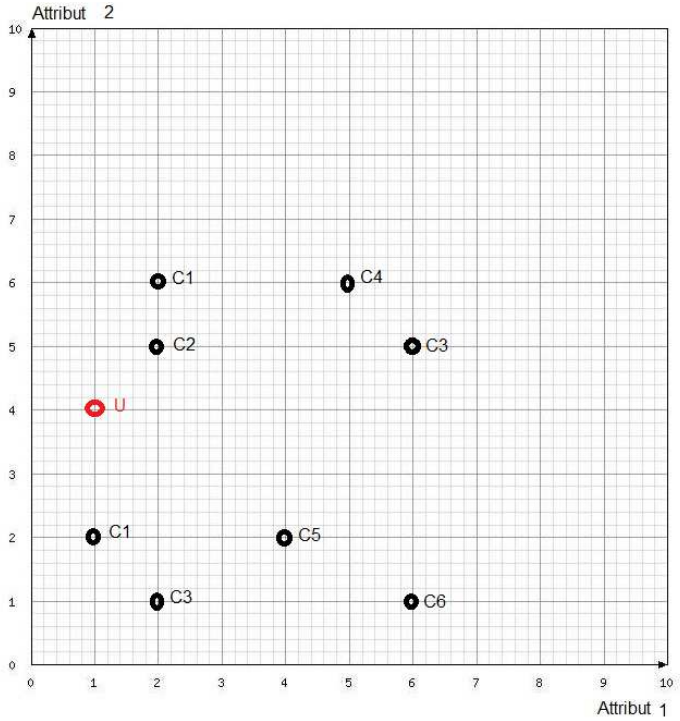

Figure 3

Question 2 :

- On veut classer un nouvel individu U ayant comme attributs (1, 4) en utilisant la méthode KNN. Quelle sera la classe de U si on choisit k=3. Justifiez.

Réponse :

On calcule la distance euclidéenne qui sépare le point U à chacun des points :

Distance	*Expression*	*Valeur*
Distance(U, point 1)	Sqrt($(1-1)^2 + ((4-2)^2$)	02,00
Distance(U, point 2)	Sqrt($(2-1)^2 + ((6-2)^2$)	02,24
Distance(U, point 3)	Sqrt($(2-1)^2 + ((5-2)^2$)	01,41
Distance(U, point 4)	Sqrt($(2-1)^2 + ((1-2)^2$)	03,16
Distance(U, point 5)	Sqrt($(4-1)^2 + ((2-2)^2$)	03,61
Distance(U, point 6)	Sqrt($(5-1)^2 + ((6-2)^2$)	04,47
Distance(U, point 7)	Sqrt($(6-1)^2 + ((5-2)^2$)	05,10
Distance(U, point 8)	Sqrt($(6-1)^2 + ((4-2)^2$)	05,83

On garde les k (k=3) plus proches voisins (ceux qui ont la plus courte distance avec U). Il s'agit des points (1, 2 et 3). Parmi ces 3 points, c'est la classe C1 qui est majoritaire (2 voix pour C1, contre 1 voix pour C2). ***Donc le point U sera affecté à la classe C1****.*

(3.5 points)

Dr Mourad BAM

Question 3 :

- On utilise maintenant la variante de KNN qui utilise la distance $1/d^2$ (inverse de la distance au carré) pour calculer les voisins. Quelle sera la classe de U avec k=3 ?. Justifiez .

Réponse :

Pour les k points voisins du point U trouvés à l'étape précédente, on calcule l'inverse de la distance au carré qui sépare le point U à chacun de ces points :

Distance	*Distance d*	*$1/d^2$*
Distance(U, point 1)	02,00	00,25
Distance(U, point 2)	02,24	00,20
Distance(U, point 3)	01,41	00,50

Pour la classe C1, la somme des poids pondérés est 0,25 + 0,20 = 0,45. Pour la classe C2, la somme est 0,50. ***Le point U sera donc affecté à la classe 2****.*

(3.5 points)

Exercice 4 (03 points) :

Question 1 : Quelles sont les qualités d'un bon clustering ?.

Réponse :

La qualité d'un clustering se mesure par le calcul de l'inertie intra-cluster et l'inertie inter-cluster.

L'inertie intra-cluster mesure la concentration des points autour du centre de gravité du cluster.

L'inertie inter-cluster mesure l'éloignement des centres de gravité des clusters entre eux.

Un "bon" clustering doit œuvrer à minimiser l'inertie intra-cluster et à maximiser l'inertie inter-cluster.

(1.5 point)

Question 2: Quelles sont les recommandations générales pour le choix du paramètre k dans la méthode k plus proches voisins ?

Réponse :

Dans l'algorithme KNN, il est recommandé de choisir K :

pas trop grand, ni trop petit.

impair (pour éviter l'égalité lors du comptage des voix parmi les voisins).

(1.5 points)

Dr Mourad LOUKAM