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

Mars 2016

Examen de rattrapage (Corrigé)

Module "Modèles pour le Datamining"

Durée : 01H30

Exercice 1 :

On dispose de 8 points : de A1 jusqu'a A8, dont les coordonnées sont les suivantes :

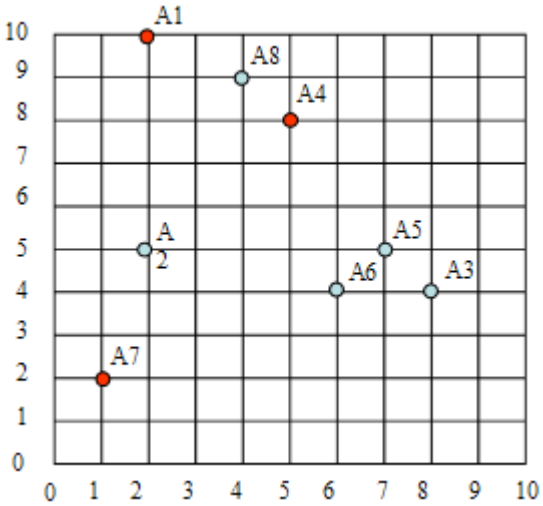
A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

Travail à faire : On veut appliquer l'algorithme Kmeans pour former 3 clusters . Initialement, on choisit comme centres des clusters : A1, A4 et A7.

Montrer toutes les étapes de calcul jusqu'à arriver au résultat final. Dessinez les états intermédiaires et le résultat final.

Réponse :

Soient $\mu 1$, $\mu 2$ et $\mu 3$ les centres de gravité de respectivement : cluster1, cluster2 et cluster3.



1ère itération :

pour A1 :
d(A1, $\mu 1$)=**0**
d(A1, $\mu 2$)= sqrt(13)
d(A1, $\mu 3$)= sqrt(65)
A1 ∈ cluster1

pour A2 :
d(A2, $\mu 1$)=sqrt(25)= 5
d(A2, $\mu 2$)= sqrt(18) = 4.24
d(A2, $\mu 3$)= sqrt(10) = **3.16**
A2 ∈ cluster3

pour A3 :
d(A3, $\mu 1$)= sqrt(36) = 6
d(A3, $\mu 2$)= sqrt(25) = **5**
d(A3, $\mu 3$)= sqrt(53) = 7.28
A3 ∈ cluster2

pour A4 :
d(A4, μ1)= sqrt(13)
d(A4, μ2)=**0**
d(A4, μ3)= sqrt(52)
A4 ∈ cluster2

pour A5 :
d(A5, μ1)=sqrt(50)=7.07
d(A5, μ2)=sqrt(13)=**3.60**
d(A5, μ3)=sqrt(45)=6.70
A5 ∈ cluster2

pour A6 :
d(A6, μ1)=sqrt(52)=7.21
d(A6, μ2)=sqrt(17)=**4.12**
d(A6, μ3)=sqrt(29)=5.38
A6 ∈ cluster2

pour A7 :
d(A7, μ1)=sqrt(65)=7.21
d(A7, μ2)=sqrt(52)=4.12
d(A7, μ3)=**0**
A7 ∈ cluster3

pour A8 :
d(A8, μ1)=sqrt(5)
d(A8, μ2)=**sqrt(2)**
d(A8, μ3)=sqrt(58)
A8 ∈ cluster2

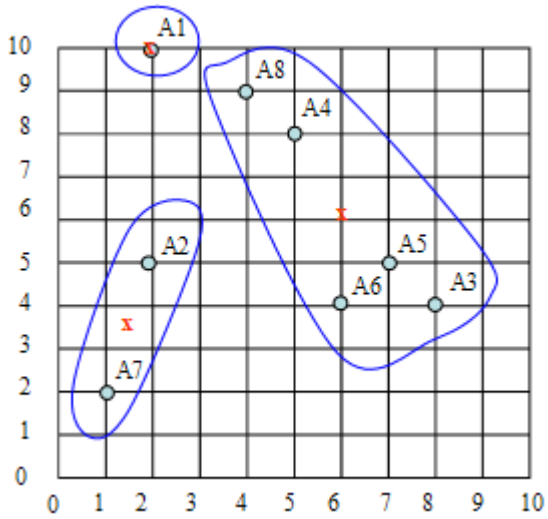
Nouveaux clusters : Cluster1: {A1}, Cluster2: {A3, A4, A5, A6, A8}, Cluster3: {A2, A7}

Actualisation des centres de gravité :

μ1= (2, 10),

μ2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6),

μ3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

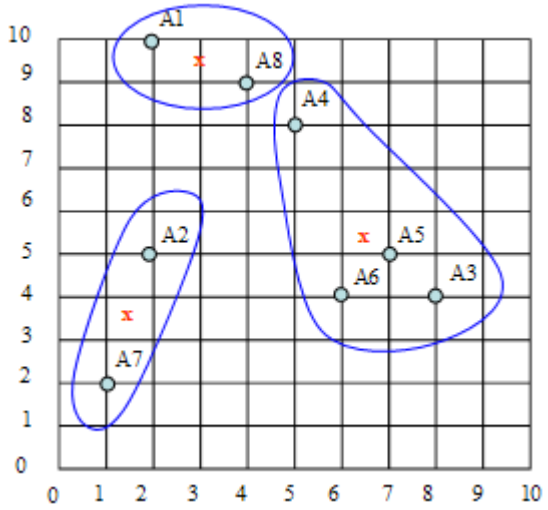


Après la seconde itération : (les candidats sont obligés d'exhiber tous les calculs)

Nouveaux clusters : cluster1: {A1, A8}, cluster2: {A3, A4, A5, A6}, cluster3: {A2, A7}
µ1=(3, 9.5),

µ2=(6.5, 5.25)

µ3=(1.5, 3.5).

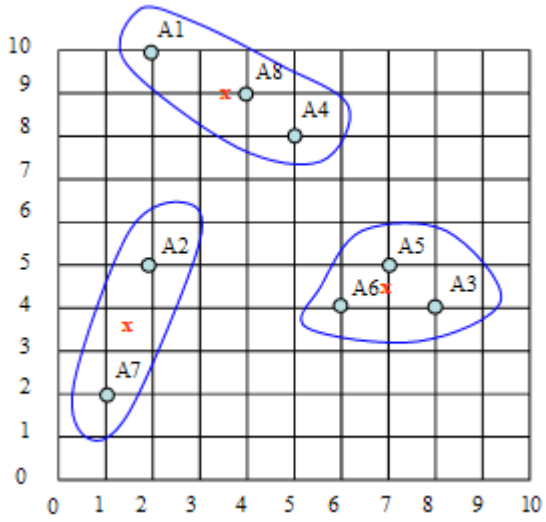


Après la 3ème itération : (les candidats sont obligés d'exhiber tous les calculs)

Nouveaux clusters : cluster1: {A1, A4, A8}, cluster2: {A3, A5, A6}, cluster3: {A2, A7}
µ1=(3.66, 9),

µ2=(7, 4.33)

µ3=(1.5, 3.5).



(12 points)

Exercice 2 :

Qu'est ce que le clustering ?.

Réponse :

> **Le Clustering** , appelé aussi classification automatique non supervisée, est technique d'analyse exploratoire des données visant le structuration des données en classes homogènes : On cherche à regrouper les points en clusters ou classes tels que les données d'un cluster soient les plus similaires possibles.

(2 points)

Donnez trois applications du clustering et les expliquer.

Réponse :

Le clustering est utilisé dans plusieurs domaines, dont :

La classification automatique de mails : l'objectif est de classer les mails (par exemple d'une entreprise commerciale) en fonction de leur objet (par exemple : commande, livraison, service après vente, etc).

La catégorisation de textes : l'objectif est de classer un ensemble de textes en fonction de leur catégorie (politique, économie, sport, etc).

Les études de maketing : l'objectif est de segmenter une population de clients d'une entreprise commerciale en fonction de leur comportement dépensier.

(3 points)

Quels problèmes doit-on confronter si on veut implémenter une méthode de clustering ?

Réponse :

- Nature des observations : Données binaires, textuelles, numériques, etc ?
- Notion de similarité (ou de dissimilarité) entre observations
- Définition d'un cluster
- Evaluation de la validité d'un cluster
- Nombre de clusters pouvant être identifiés dans les données
- Quels algorithmes de clustering ?
- Comparaison de différents résultats de clustering

(3 points)