UMB M'sila Master1 IA

Département de l'Informatique Module ML&DM

TP N°2

Clustering avec R (K-means et CAH)

Objectif du TP :

- Savoir comment utiliser les méthodes de clustering (K-means et Classiffication Ascendante Hiérarchique) pour réaliser le clustering d'un ensemble de données.

Etapes du TP

1- **Ensemble de données (le fichier fromage.txt)**

L'ensemble de données contient les informations sur les différents fromages.

Fromages	calories	sodium	calcium	lipides	retinol	folates	proteines	cholesterol	magnesium
CarredelEst	314	353.5	72.6	26.3	51.6	30.3	21	70	20
Babybel	314	238	209.8	25.1	63.7	6.4	22.6	70	27
Beaufort	401	112	259.4	33.3	54.9	1.2	26.6	120	41
Bleu	342	336	211.1	28.9	37.1	27.5	20.2	90	27
Camembert	264	314	215.9	19.5	103	36.4	23.4	60	20
Cantal	367	256	264	28.8	48.8	5.7	23	90	30
Chabichou	344	192	87.2	27.9	90.1	36.3	19.5	80	36
Chaource	292	276	132.9	25.4	116.4	32.5	17.8	70	25
Cheddar	406	172	182.3	32.5	76.4	4.9	26	110	28
Comte	399	92	220.5	32.4	55.9	1.3	29.2	120	51
Coulomniers	308	222	79.2	25.6	63.6	21.1	20.5	80	13
Edam	327	148	272.2	24.7	65.7	5.5	24.7	80	44
Emmental	378	60	308.2	29.4	56.3	2.4	29.4	110	45
Fr.chevrepatemolle	206	160	72.8	18.5	150.5	31	11.1	50	16
Fr.fondu.45	292	390	168.5	24	77.4	5.5	16.8	70	20
Fr.frais20nat.	80	41	146.3	3.5	50	20	8.3	10	11
Fr.frais40nat.	115	25	94.8	7.8	64.3	22.6	7	30	10
Maroilles	338	311	236.7	29.1	46.7	3.6	20.4	90	40
Morbier	347	285	219	29.5	57.6	5.8	23.6	80	30
Parmesan	381	240	334.6	27.5	90	5.2	35.7	80	46
Petitsuisse40	142	22	78.2	10.4	63.4	20.4	9.4	20	10
PontlEveque	300	223	156.7	23.4	53	4	21.1	70	22
Pyrenees	355	232	178.9	28	51.5	6.8	22.4	90	25
Reblochon	309	272	202.3	24.6	73.1	8.1	19.7	80	30
Rocquefort	370	432	162	31.2	83.5	13.3	18.7	100	25
SaintPaulin	298	205	261	23.3	60.4	6.7	23.3	70	26
Tome	321	252	125.5	27.3	62.3	6.2	21.8	80	20
Vacherin	321	140	218	29.3	49.2	3.7	17.6	80	30
Yaourtlaitent.nat.	70	91	215.7	3.4	42.9	2.9	4.1	13	14

A- Chargement des données et description

- modifier le répertoire par défaut
  setwd("Chemin du répertoire ou le fichier fromage.txt existe")
- charger les données - attention aux options

fromage <- read.table(file="fromage.txt",header=T,row.names=1,sep="\t",dec=".")

- afficher les premières lignes
  print(head(fromage))
- statistiques descriptive (sommaire des données)
  print(summary(fromage))
- graphique - croisement deux à deux
  pairs(fromage)

B- Clustering par K-means

- center = 4 - nombre de groupes demandés
- nstart = 5 - nombre d'essais avec différents individus de départ
  groupes.kmeans <- kmeans(fromage,centers=4,nstart=5)
- affichage des résultats
  print(groupes.kmeans)

C- Clustering avec CAH

- distance entre individus
  d.fromage <- dist(fromage)
- CAH – critère: lien maximum (complete link)
  cah_clink <- hclust(d.fromage,method="complete")
- affichage dendrogramme
  plot(cah_clink)
- dendrogramme avec matérialisation des groupes
  rect.hclust(cah_clink,k=4)
- découpage en 4 groupes
  groupes.cah <- cutree(cah_clink,k=4)
- liste des groupes
  print(sort(groupes.cah))