Examen final

Clustering, Classification et Règles d'Association

le 12 décembre 2024

Les documents papier sont autorisés, de même qu'une calculatrice classique (comme celle utilisée au collège). Tout autre dispositif électronique est interdit.
L'ordre de résolution des sujets n'est pas imposé.

1. **Clustering**

   (a) **Code**

   Soit `X` un dataframe en R. Que fait le code suivant :

```
n <- dim(X)[1]
indice <- sample(1:n, n)
X2 <- X[indice,]

kcl <- kmeans(X, k = 3)
kcl2 <- kmeans(X2,k = 3)
table(cl$cluster[indice], cl2$cluster)

d <- dist(X)
d2 <- dist(X2)
hcl <- cutree(hclust(d=d, method='single'), k=3)
hcl2 <- cutree(hclust(d=d2, method='single'), k=3)
table(hcl[indice], hcl2)

dcl <- dbscan(data = X, eps = 0.2, MinPts = 3)
dcl2 <- dbscan(data = X2, eps = 0.2, MinPts = 3)
table(dcl$cluster[indice], dcl2$cluster)
```

   A quoi s'attendre comme résultats pour chacune des fonctions `table` ?

   (b) **Calculs**

   Pour les données représentées dans la figure 1b dans l'espace $\mathbb{R}^2$ avec la distance euclidienne, mettez en pratique deux algorithmes de clustering :
   — k-moyennes avec les centres initiaux c et k
   — hiérarchique ascendant avec la stratégie de la distance maximale.
   Travaillez directement sur la dernière feuille que vous détacherez et que vous allez glisser dans la double copie

2. **Itemsets fréquents et règles d'association**

   Soit X_1, X_2, X_3 un 3-itemset fréquent. On essaie de générer toutes les règles d'association à partir de cet itemset.

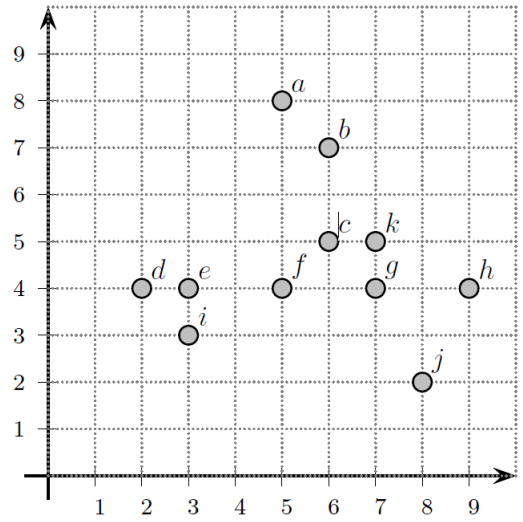

FIGURE 1 – Les points à regrouper en clusters

(a) combien il y en a (sans regarder la confiance) ?

(b) Si on ne cherche que les règles d'association "très fortes", de confiance = 1, avons nous besoin de les générer toutes les règles possibles et calculer expressément la confiance ? Généraliser le raisonnement pour un k-itemset avec un k quelconque.

3. **Calcul d'itemsets fréquent et de règles d'association**

Soit la base de transactions suivante :

TID	items
T1	C, H, A, T
T2	C, A, T
T3	C, H, T
T4	A, H
T5	C, A, T, H
T6	T, H, E
T7	H, T, E, A
T8	H, A, E

(a) Détaillez le déroulement d'un algorithme de calcul des itemsets fréquents pour

$$min\_support\_count = 3$$

Listez explicitement les itemsets fréquents obtenus.

(b) Mettez en évidence les itemsets fermés.

(c) Générez les règles d'association très fortes avec $min\_confiance = 1$, calculez aussi le $LIFT$.

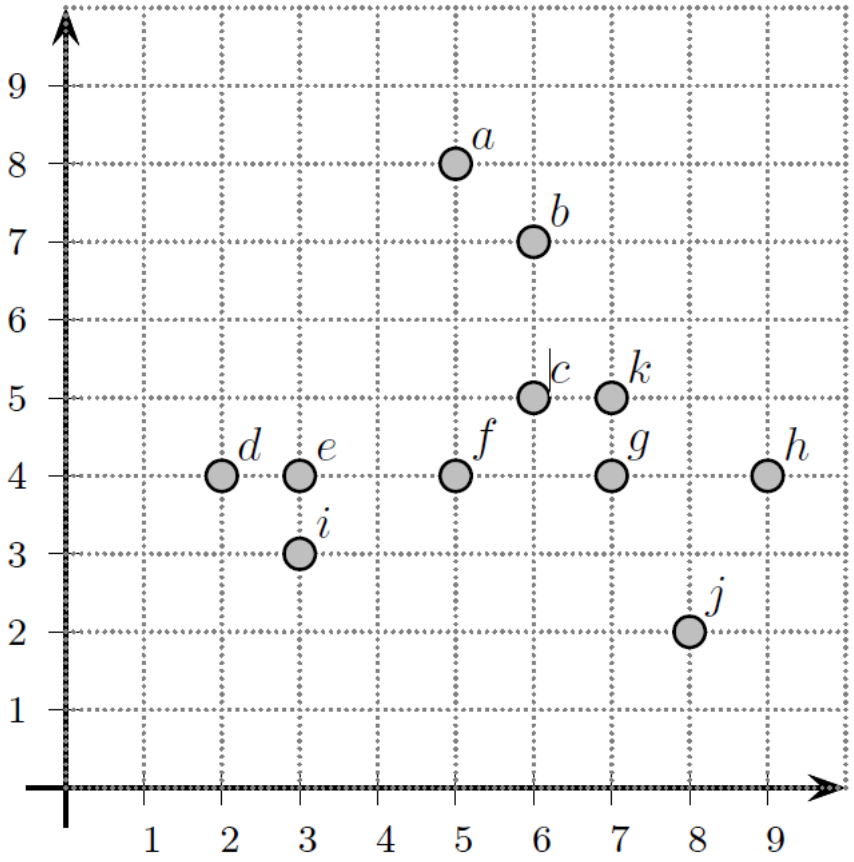

FIGURE 2 – k-means : indiquez successivement les centres mobiles et la séparation en clusters

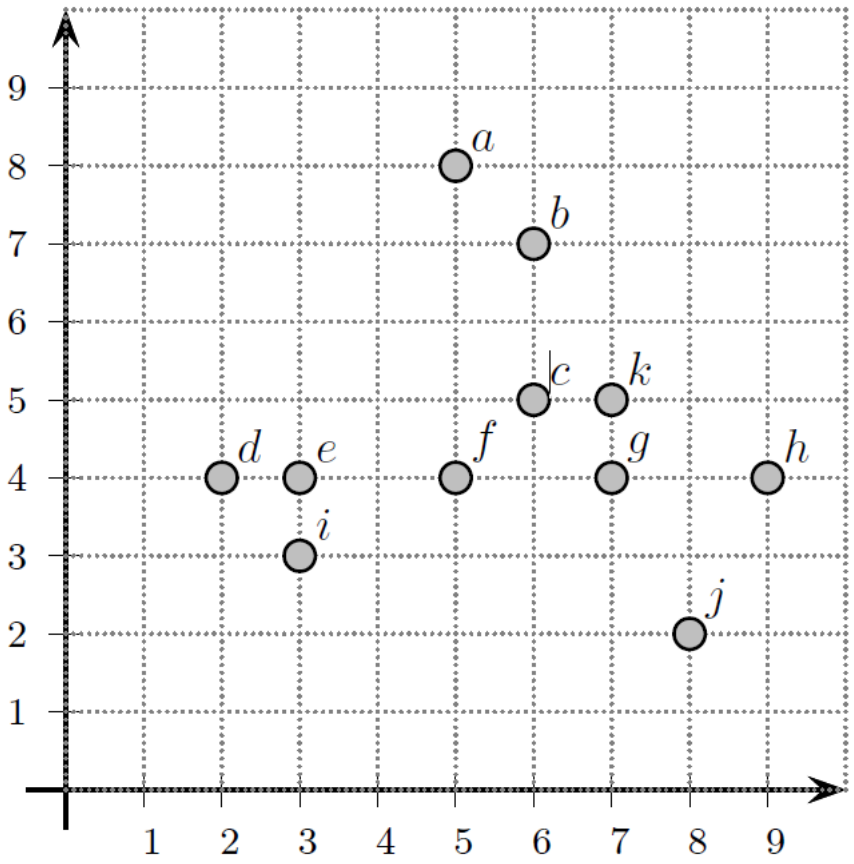

FIGURE 3 – Clustering hiérarchique ascendant avec la distance maximale ; entourez les clusters au fur et à mesure qu'ils se forment, puis dessinez le dendrogramme.