DCD : Découverte de Connaissances dans des Données

TIW09 – Automne 2020–2021

↓ Reporter le numéro d'anonymat de la copie ici ↓

numéro de copie :

↑ Reporter le numéro d'anonymat de la copie ici ↑

Résumé

Durée : 90 min. Documents (papiers) autorisés. L'usage du téléphone est strictement interdit. Les réponses doivent être données sur la feuille.

Exercise 1 : Clustering (8pts)

On considère les 12 points suivants :

Points	(x;y)	Points	(x;y)
A	(3 ; 0.5)	G	(7 ; 5)
B	(4 ; 2)	H	(3 ; 5)
C	(3 ; 2)	I	(1 ; 5.5)
D	(8 ; 2)	J	(2 ; 6)
E	(9.5 ; 2.5)	K	(1 ; 7)
F	(9 ; 3.5)	L	(5 ; 7.5)

1. Effectuer l'algorithme k-means afin d'obtenir une partition en 3 sous-ensembles de ces points. On considérera la distance de Manhattan et les points B, C et D comme centres initiaux. *Bien dérouler les différentes étapes.*

2. *Même question avec les points A, E et F comme centres initiaux*

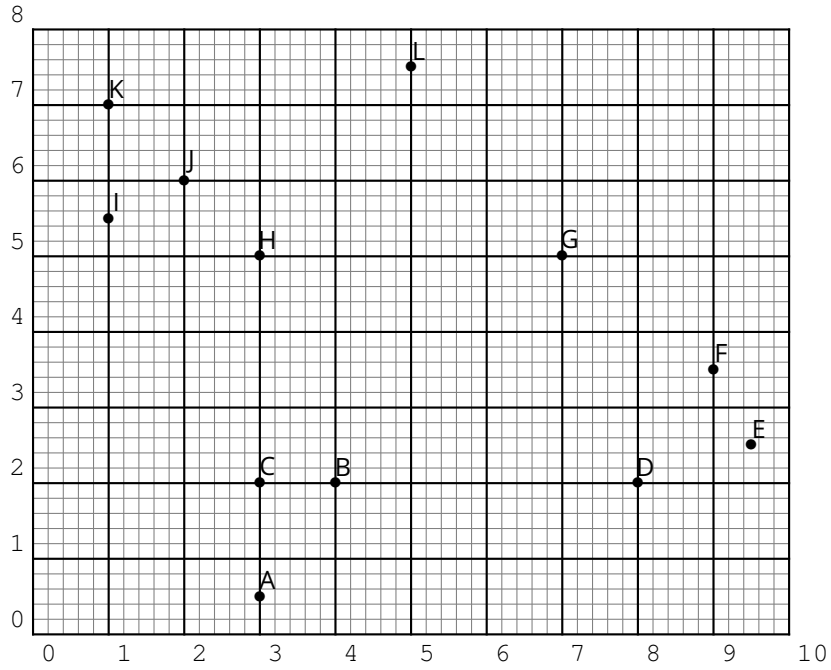
3. *L'algorithme des $k-$ moyennes est un algorithme glouton, qui cherche à minimiser la somme des dissimilarités intra-groupes. Si $(C_i)_{i=1..k}$ est une partition, avec m_k la moyenne du groupe C_k, alors la fonction de score de cette partition est :*

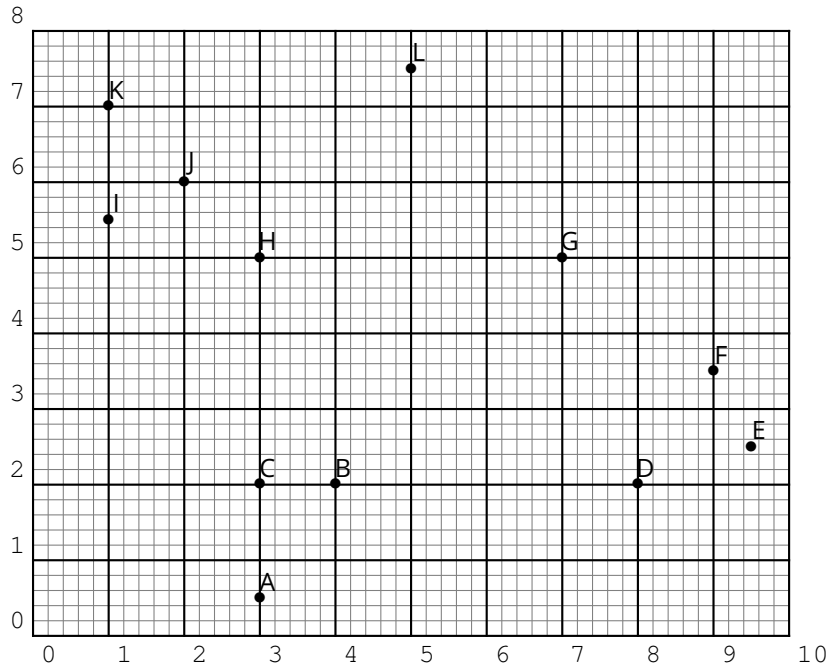
$$\Sigma_{i=1}^{k} \Sigma_{x \in C_k} d(x, m_k)$$

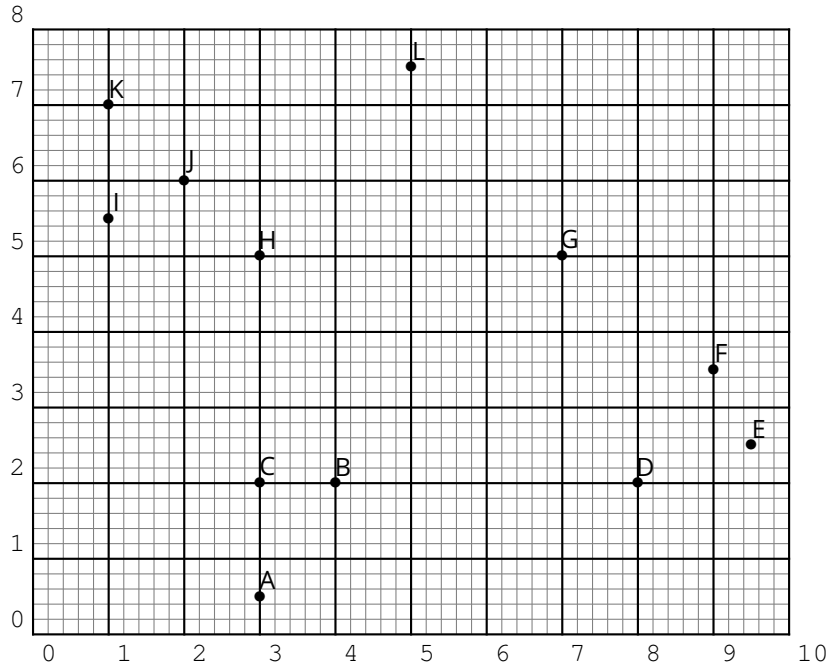
Cette mesure comme fonction de score a comme effet de trouver des groupes assez compacts, de forme circulaire. Evaluez le score des deux partitions trouvées pour en déterminer la meilleure selon ce critère.

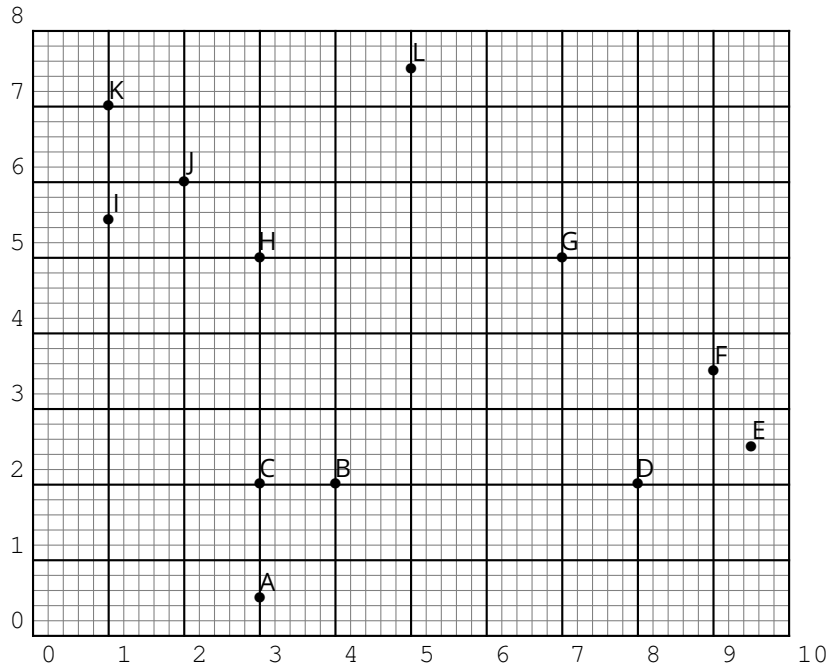
4. Appliquer un clustering hiérarchique. On utilisera toujours la distance de Manhattan et la single link method pour calculer la distance entre deux clusters.

5. *Même question en utilisant la complete link method*


8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
L
K
J
I
H
G
F
E
C
B
D
A


8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
L
K
J
I
H
G
F
E
C
B
D
A


8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
L
K
J
I
H
G
F
E
C
B
D
A


8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10
L
K
J
I
H
G
F
E
C
B
D
A