

TP 4 : Les algorithmes de classification

Exercice 1 :

1) Appliquer les centres mobiles à la main pour trouver une partition en deux classes des quatre points-individus ind1, ind2, ind3, ind4 en prenant ind1 et ind2 comme centres initiaux.

Id	Var1	Var2
ind1	5	4
ind2	4	5
ind3	1	-2
ind4	0	-3

En pondérant les points-individus par 1, calculer l'inertie intra-classe de cette partition.

2) Recommencez avec le logiciel R et vérifiez que vous retrouvez la même partition (en partant des mêmes deux centres initiaux) et que vous retrouvez l'inertie intra-classe calculée ci-dessus.

3) En vous aidant de R et de la fonction `var()` qui donne la variance empirique (estimateur non biaisé donc avec $1/n-1$), calculer le pourcentage de l'inertie du nuage des 4 points-individus expliqué par cette partition en 2 classes.

Exercice 2 :

1) Appliquer à la main l'algorithme de classification hiérarchique sur le tableau de données de l'exercice 1 en prenant le lien maximum comme mesure d'agrégation. Quelle est alors la hiérarchie H et la hiérarchie indicée (représentée par le dendrogramme). Quelle est la partition en deux classes issue du dendrogramme du lien max ?

2) A la main toujours, faire le dendrogramme obtenu avec la mesure d'agrégation de Ward.

3) Avec les fonctions `hclust()` et `cutree()` de R, construire la hiérarchie indicée du lien maximum, son dendrogramme et la partition en deux classes associée. Vérifier que vous retrouvez les mêmes résultats qu'à la main. Attention : la fonction `hclust()` prend en entrée une matrice de distance calculée avec la fonction `dist()`.

4) Idem pour Ward. Attention, pour Ward, il faut passer la matrice de distance au carré en argument. Les indices calculés par cette fonction sont ceux de Ward au carré.