

Chapitre 5 – Classification

I Mesures de ressemblance

1 Ressemblance entre individus

On considère n individus et p variables quantitatives et on cherche à évaluer la ressemblance entre les individus en fonction des valeurs des variables pour ces individus.

Distances. Rappelons qu'une distance est une application $d : E \times E \rightarrow \mathbb{R}_+$ qui vérifie les propriétés suivantes pour tous $(x, y, z) \in E^3$:

1. $d(x, y) = d(y, x)$ (symétrie)
2. $d(x, y) = 0 \iff x = y$ (séparation)
3. $d(x, y) \leq d(x, z) + d(z, y)$ (inégalité triangulaire)

On utilise le plus souvent la distance euclidienne :

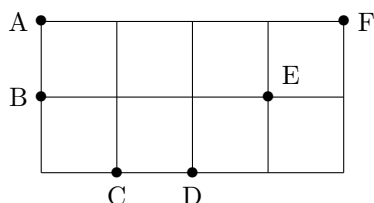
$$d^2(x_i, x_j) = \sum_{k=1}^p (x_{ik} - x_{jk})^2$$

Un exemple de distance non-euclidienne souvent utilisée est la distance dite de Manhattan :

$$d(x_i, x_j) = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

Exemple : A(1, 1, 3), B(1, 1, 1) et C(2, 2, 2).

Exemple : A, B, C, D, E et F sur le dessin ci-dessous.



Similarités et dissimilarités. Une similarité est une application $s : E \times E \rightarrow \mathbb{R}_+$ qui vérifie les propriétés suivantes pour tous $(x, y, z) \in E^3$:

1. $s(x, y) = s(y, x)$ (symétrie)
2. $s(x, x) \geq s(x, y)$

Une dissimilarité est une application $d : E \times E \rightarrow \mathbb{R}_+$ vérifiant les propriétés suivantes pour tous $(x, y, z) \in E^3$:

1. $d(x, y) = d(y, x)$ (symétrie)
2. $d(x, x) = 0$

Exemple. En phytosociologie, on définit un ensemble de stations censées représenter la diversité des milieux d'une zone géographique. On rassemble les données sur les I espèces et les J stations étudiées dans un tableau : x_{ij} vaut 1 si l'espèce i est présente dans la station j et 0 sinon.

La plus ancienne mesure de ressemblance sur ce type de données date de Paul Jaccard (1901) :

$$s(i, j) = \frac{n_{++}}{n_{++} + n_{+-}}$$

où n_{++} est le nombre de stations où les deux espèces sont présentes, et n_{+-} le nombre de stations où seule l'une des deux espèces est présente.

2 Ressemblance entre groupes d'individus

Une fois la notion de ressemblance entre individus définie (souvent par une distance), il faut définir celle entre groupes d'individus.

Saut minimum. La méthode du saut minimum consiste à mesurer la ressemblance entre deux groupes d'individus A et B par la plus petite distance entre un élément de A et un élément de B.

Diamètre. La méthode du diamètre consiste à mesurer la ressemblance entre deux groupes d'individus A et B par la plus grande distance entre un élément de A et un élément de B.

Moyenne. La méthode de la moyenne consiste à mesurer la ressemblance entre deux groupes d'individus A et B par la moyenne des distances entre un élément de A et un élément de B.

Autres méthodes. Dans le cas de distances euclidiennes, on peut également mesurer la ressemblance par la distance entre les centres de gravité des groupes, ou encore par l'inertie inter-classe des deux groupes (d'après le théorème de Huyguens, cf. méthode de Ward ci-dessous).

II Classification hiérarchique ascendante

1 Méthode classique

Algorithme.

1. On construit la matrice de dissimilarités entre individus (qui est symétrique).
2. On regroupe les deux individus i et j les plus proches. La valeur $d(i, j)$ est l'indice d'agrégation entre i et j .
3. On met à jour la matrice en supprimant les lignes et les colonnes correspondant aux individus i et j et en créant une nouvelle ligne et une nouvelle colonne pour le groupe (i, j) , que l'on remplit avec les dissimilarités entre ce groupe et chacun des individus restants.
4. Et on recommence jusqu'à ce qu'il ne reste plus qu'un seul groupe.

Exemple : classification des 6 points A, B, C, D, E et F du dessin ci-dessus.

Partition. On peut obtenir une partition de l'ensemble initial à partir de l'arbre de classification en traçant une ligne horizontale, de la partition la plus grossière (une seule classe contenant tous les individus) à la partition la plus fine (une classe par individu).

On considère généralement que la partition offrant le meilleur compromis est celle obtenue par la ligne tracée à l'endroit du saut le plus grand entre deux nœuds de l'arbre.

Exemple : dans l'exemple ci-dessus, on obtient la partition $\{A, B\}$, $\{C, D\}$, $\{E, F\}$.

2 Méthode de Ward

Principe de la méthode : Elle s'applique à des individus dont la ressemblance est mesurée par une distance euclidienne.

Une bonne partition est telle que

- l'inertie intra-classe de chaque classe (*ie* la somme des écarts entre chaque point et le centre de gravité de la classe) est faible (les individus sont homogènes à l'intérieur d'une classe) ;
- l'inertie inter-classe (*ie* la somme des écarts entre les centres de gravité de chaque classe et le centre de gravité global) est élevée (les individus sont très différents d'une classe à l'autre).

D'après le théorème de Huyguens, l'inertie totale est la somme de ces deux inerties. Quand on regroupe deux classes A et B, on veut minimiser le gain d'inertie intra-classe engendré, donc maximiser la perte d'inertie inter-classe.

Algorithme.

1. On construit la matrice des distances euclidiennes entre individus (qui est symétrique).
2. On regroupe les deux individus i et j les plus proches. La valeur $d(i, j)$ est l'indice d'agrégation entre i et j .
3. On met à jour la matrice en supprimant les lignes et les colonnes correspondant aux individus i et j et en créant une nouvelle ligne et une nouvelle colonne pour le groupe (i, j) , que l'on remplit avec les gains d'inertie entre ce groupe et chacun des individus restants, calculés par la formule :

$$\Delta(A, B) = \frac{N_A N_B}{N_A + N_B} d^2(g_A, g_B)$$

où N_A est le nombre d'éléments dans le groupe A et g_A le centre de gravité du groupe A.

4. Et on recommence jusqu'à ce qu'il ne reste plus qu'un seul groupe.

Remarques. La méthode de Ward a tendance à privilégier le rapprochement de classes dont les centres de gravité sont proches et dont les effectifs sont faibles. Les arbres obtenus sont ainsi harmonieux (les partitions sont composées de classes d'effectifs relativement proches) et ne présentent pas d'inversion.

En outre, la somme de tous les indices d'agrégation est égale à l'inertie totale de l'ensemble des individus, ce qui permet de calculer la perte d'inertie à chaque étape, et donc de choisir la meilleure partition du point de vue de l'inertie perdue.

Exemple : classification des 6 points A, B, C, D, E et F du dessin ci-dessus.

III Partitionnement : méthode des centres mobiles

Contrairement aux méthodes hiérarchiques, le nombre de classes est fixé a priori.

Algorithme.

1. On se donne une partition initiale P_0 .
2. On calcule le centre de gravité de chaque classe de la partition.
3. On réaffecte chaque individu à la classe dont il est le plus proche (en termes de distance au centre de gravité).
4. On obtient une nouvelle partition et on recommence tant que cette partition est différente de la précédente (dans les petits exemples), ou tant que la nouvelle partition améliore la précédente (du point de vue de l'inertie).

Exemple : partitionnement des 6 points A, B, C, D, E et F du dessin ci-dessus en deux classes.