Méthode de classification non supervisée

1 Classification hiérarchique ascendante

La distance entre deux classes contenant plus d'un individu est déterminée suivant l'un des critères d'agrégation explicités ci-dessous :

1. **Critère du saut minimal (Single linkage clustering):**
   - c'est la plus petite distance séparant un individu de la première classe et un individu de la deuxième classe :
   $$D(C_1, C_2) = \underset{x_i \in C_1, x_{if} \in C_2}{Min} d(x_i, x_{if})$$
2. **Critère du saut maximal (Complete linkage clustering):**
   - c'est la plus grande distance séparant un individu de la première classe et un in dividu de la deuxième classe :
   $$D(C_1, C_2) = \underset{x_i \in C_1, x_{if} \in C_2}{Max} d(x_i, x_{if})$$
3. **Critère de la moyenne (Average linkage clustering):**
   - c'est la moyenne des distances entre tous les individus de la première classe et tous les individus de la deuxième classe :
   $$D(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x_i \in C_1} \sum_{x_{if} \in C_2} d(x_i, x_{if})$$
   - Ou $|C|$ désigne l'effectif de la classe C
4. **Critère de Ward (Ward, 1963) :**
   - Ce critère impose l'utilisation du carré de la distance euclidienne.:
   $$D(C_1, C_2) = \frac{|C_1||C_2|}{|C_1| + |C_2|} d_2^2(G_1, G_2)$$
   - Ou $|C|$ désigne l'effectif de la classe C et G_1 le centre de gravité de la première classe et G_2 le centre de gravité de la deuxième classe.
   - On définit la distance euclidienne entre les seux vecteurs d'observation x_i et x_{il}:
   $$d_2(x_i, x_{il}) = \sqrt{\sum_{j=1}^{p} (x_i^j - x_{il}^j)^2}$$
   - On définit le centre de gravité de la classe C_k par:
   $$G_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|}$$

Algorithme: Classification hiérarchique ascendante
Entrée :

- X : **La matrice des individus;**
- K : **Le nombre de classes désiré, sinon K vaut 1;**

Sortie :

- P : **Les partitions emboîtées**

Début

Initialisation :

- $t \leftarrow 0$ $\quad t$: *Iteration courante*

- $nbClasses \leftarrow n$ $\quad$ *Le nombre de classes est egal a celui des individus*

- *Mettre chaque individu dans une classe***:**

- **Pour i de** 1 **à** n **faire**
  $C_i \leftarrow x_i$ $\quad \{Mettre\ chaque\ individu\ dans\ une\ classe\}$**;**
  Fin Pour

- $P[t] = P[0] = \{C_1, C_2, ..., C_n\}$ $\quad$ *Partition initiale*

Agrégations:
Tant que $(nbClasses > K)$ **faire**
$t \leftarrow t + 1$;
Calculer les distances entre les classes :

$$D(C_k, C_l) \quad \forall k \in \{1, ...nbClasses\}\ et\ l \in \{1, ...nbClasses\}$$

Trouver les deux classes les plus proches au sens d'un critère d'agrégation :

$$D(C_q, C_0) \leftarrow Min\, D(C_k, C_l) \quad \forall k \in \{1, ...nbClasses\}\ et\ l \in \{1, ...nbClasses\}$$

Copier les individus de la classe C_0 dans C_q:

$$C_q \leftarrow x_i \quad \forall x_i \in C_0$$

Supprimer la classe C_0
$nbClasses \leftarrow nbClasses - 1$
$P[t] \leftarrow P[t-1] - \{C_0\}$

Fait
Retourner P
Fin

2 Centres-mobiles

Contrairement aux méthodes hiérarchiques qui produisent différentes partitions emboîtées, les algorithmes de classification par partition répartissent les individus en classes dans le but d'obtenir une seule partition optimale. Une partition fixe sur l'ensemble Ω est constituée de K classes tel que chaque individu appartienne à une seule classe. Le nombre des classes doit être fourni à l'algorithme.
Le but de la méthode des centres-mobiles est de parvenir, en un nombre d'itérations limité, à partitionner

les n individus en K classes homogénes en minimisant le critère suivant :

$$G_{cm} = \sum_{k=1}^{K} \sum_{x_i \in C_k} d_2^2(x_i, w_k)$$

Ou w_k est le représentant de la classe C_k qui est aussi son centre de gravité défini par: $w_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|}$

Algorithme: Centres-mobiles
Entrée :

- X : **La matrice des individus;**
- K : **Le nombre de classes désiré;**

Sortie :

- $P_f \leftarrow \{C_1, C_2, ..., C_K\}$ **partition finale**

Début

Initialisation :

- $t \leftarrow 0$ t : *Iteration courante*

- **Choix aléatoire de la partition initiale:**

  - **Pour i de** 1 **à** n **faire**
    $l \leftarrow alea(1, ...K)$ **Choisir une valeur aléatoire entre 1 et K;**
    $C_l \leftarrow x_i$**;**
    Fin Pour

- $P[t] = P[0] = \{C_1, C_2, ..., C_K\}$ *Partition initiale*

Partionnement:
Répéter

- $t \leftarrow t + 1$;

- Etape représentation :

  - Calculer le centre de chaque classe C_k:
  - **Pour** k **de** 1 **à** K **faire**
    $w_k \leftarrow \frac{\sum_{x_i \in C_k} x_i}{|C_k|}$
    Fin Pour

- Etape affectation :

  - générer une nouvelle partition:
  - **Pour** i **de** 1 **à** n **faire**
    $d(w_l, x_i) \leftarrow \underset{k \in \{1,...,K\}}{Min} d(w_k, x_i)$
    $C_l \leftarrow x_i$;
    Fin Pour

jusqu'à ce que $(P_t \neq P_{t-1})$

Retourner $P_f = P_t$

Fin

Algorithme: Centres-mobiles séquentiel
Entrée :

- X : **La matrice des individus;**
- K : **Le nombre de classes désiré;**

Sortie :

- $P_f \leftarrow \{C_1, C_2, ..., C_K\}$ **partition finale**

Début

Initialisation :

- $t \leftarrow 0$ $\quad t$: *Iteration courante*

- **Choix aléatoire représentants des classes:** $\{w_k, k \in \{1, ...K\}\}$

- **Initialiser les effectifs des classes à** 0**:** $|C_1| = |C_2| = \ldots = |C_K| = 0$

Partionnement:
Répéter

- $t \leftarrow t + 1$;

- Etape affectation :

  - acquérir une observation x_i et l'affecter à la classe la plus proche:
  - **Pour** k **de** 1 **à** K **faire**
    $d_2^2(w_l, x_i) \leftarrow \underset{k \in \{1,...,K\}}{Min} d_2^2(w_k, x_i)$
    $C_l \leftarrow x_i$;
    Fin Pour

- Etape représentation :

  - calcul du nouveau centre de gravité de la classe C_l:
  - $n_l \leftarrow n_l + 1$
  - $w_l \leftarrow w_l + \frac{x_i - w_l}{n_l}$

jusqu'à ce que (Plus d'observations disponibles)

Retourner $P_f = P_t$

Fin