Clustering hierarchique ascendant

Caio Corro

1 Méthodes

Les méthodes de clustering ascendant (ou encore agglomératif) commencent par assigner chaque point à son propre cluster. Par exemple, pour un jeu de données de n points $X = [\![\boldsymbol{x}^{(i)}]\!]_{i=1}^{n}$ nous commençons donc par créer n clusters :

$$\pi^{(1)} = \{C^{(1)}, \dots, C^{(n)}\} \qquad \text{avec} \qquad C^{(i)} = \{\boldsymbol{x}^{(i)}\}\,.$$

Ensuite, nous allons itérativement fusionner les clusters deux par deux. Par exemple, à la première étape, nous pouvons choisir de fusionner $C^{(1)}$ et $C^{(2)}$ pour obtenir :

$$\begin{aligned}\pi^{(2)} &= \{C^{(1)} \cup C^{(2)}, C^{(3)}, \dots, C^{(n)}\}\,,\\ &= \{C^{(n+1)}, C^{(3)}, \dots, C^{(n)}\} \qquad \text{avec} \qquad C^{(n+1)} = C^{(1)} \cup C^{(2)}\,.\end{aligned}$$

Après $n-1$ fusions, nous obtenons simplement une partition contenant un seul cluster :

$$\pi^{(n)} = \{C^{(2n-1)}\} \qquad \text{avec} \qquad C^{(2n-1)} = X\,.$$

La Figure 1 illustre cette procédure.

Évidemment, nous devons choisir, à chaque étape, quelles sont les deux clusters de $\pi^{(i)}$ à fusionner pour créer $\pi^{(i+1)}$. Pour cela, nous avons besoin d'introduire une notion de distance entres clusters, et les deux clusters les plus proches selon celle-ci seront fusionné. Nous avons déjà des mesures de distance entre points, mais ici un cluster peut contenir plusieurs points, ce qui rends la tâche plus compliquée. Une première solution consiste à définir la distance entre les clusters C et C' comme la distance minimum entre un point de C et un point de C', distance qui est appelée **single linkage** :

$$D_{\min}(C, C') = \min_{\substack{\boldsymbol{x}\in C,\\ \boldsymbol{x}'\in C'}} d(\boldsymbol{x}, \boldsymbol{x}')\,,$$

où $d(\cdots, \cdot)$ est une mesure de distance entre deux points, par exemples la distance L2 $d(\boldsymbol{x}, \boldsymbol{s}') = \|vx - \boldsymbol{x}'\|_2$. De la même façon, on peut aussi utiliser la distance maximum, distance qui est appelée **complete linkage** :

$$D_{\max}(C, C') = \max_{\substack{\boldsymbol{x}\in C,\\ \boldsymbol{x}'\in C'}} d(\boldsymbol{x}, \boldsymbol{x}')\,.$$

Notons que les deux distances ci-dessus dépendent uniquement d'un point de chaque cluster. Il existe également des variances prenant en compte la totalité des points, par exemple la distance moyenne entre chaque couple de points :

$$D_{\text{avg}}(C, C') = \frac{1}{|C| \times |C'|} \sum_{\boldsymbol{x}\in C} \sum_{\boldsymbol{x}'\in C'} d(\boldsymbol{x}, \boldsymbol{x}')\,,$$

ou encore la distance entre les centroides de chaque cluster :

$$D_{\text{cent}}(C, C') = d\left(\frac{1}{|C|}\sum_{\boldsymbol{x}\in C} \boldsymbol{x}, \frac{1}{|C'|}\sum_{\boldsymbol{x}'\in C'} \boldsymbol{x}'\right)\,.$$

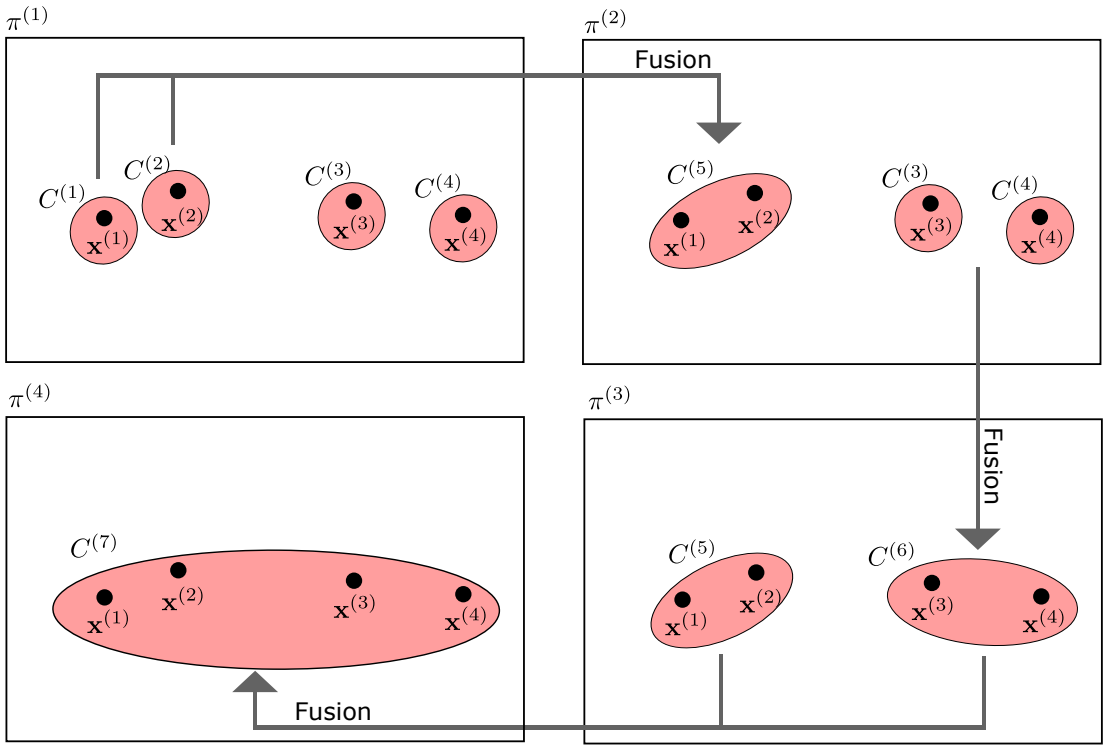


FIGURE 1 – Caption

À noter qu'en fonction du critère utilisé pour la distance entre deux clusters, la suite de solutions $\pi^{(i)}$ pourra être différente (voir exercices). Cependant, si les données sont bien séparées en « bulles compactes », toutes ces distance entre clusters auront tendance à donner le même résultat.

Enfin, notons que le **critère de ward** pour fusionner deux clusters est défini de la façons suivante :

$$D_{\text{ward}}(C, C') = \frac{|C| \times |C'|}{|C| + |C'|} d\left(\frac{1}{C}\sum_{\boldsymbol{x}\in C}\boldsymbol{x}, \frac{1}{|C'|}\sum_{\boldsymbol{x}'\in C'}\boldsymbol{x}'\right),$$

où $d(\cdot,\cdot)$ est la distance euclidienne au carrée, c'est-à-dire qu'on a :

$$= \frac{|C| \times |C'|}{|C| + |C'|} \left\| \frac{1}{C}\sum_{\boldsymbol{x}\in C}\boldsymbol{x} \quad - \quad \frac{1}{|C'|}\sum_{\boldsymbol{x}'\in C'}\boldsymbol{x}' \right\|_2^2 .$$

1.1 Dispersion

Soit $X \subseteq \mathbb{R}^m$ un ensemble de points et $\pi = \{C^{(1)}, ..., C^{(k)}\}$ une partition de X. Le centroïde du des données X, que l'on écrit $\bar{x}$, est défini comme :

$$\bar{x} = \frac{1}{|X|}\sum_{\boldsymbol{x}\in X}\boldsymbol{x}\,.$$

Le centroïde du cluster i, que l'on écrit $\boldsymbol{m}^{(i)}$, est défini comme :

$$\boldsymbol{m}^{(i)} = \frac{1}{|C^{(i)}|}\sum_{\boldsymbol{x}\in C^{(i)}}\boldsymbol{x}\,.$$

Notons que l'on a $\bar{x} \in \mathbb{R}^m$ et $\boldsymbol{m}^{(i)} \in \mathbb{R}^m$. La dispersion intra-clusters de la partition π est :

$$\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C^{(i)}} \|\boldsymbol{x} - \boldsymbol{m}^{(i)}\|_2^2\,.$$

La dispersion inter-clusters de la partition π est :

$$\sum_{i=1}^{k} |C^{(i)}| \times \|\boldsymbol{m}^{(i)} - \bar{\boldsymbol{x}}\|_2^2 \,.$$

En pratique, nous cherchons une partition des données telle que :

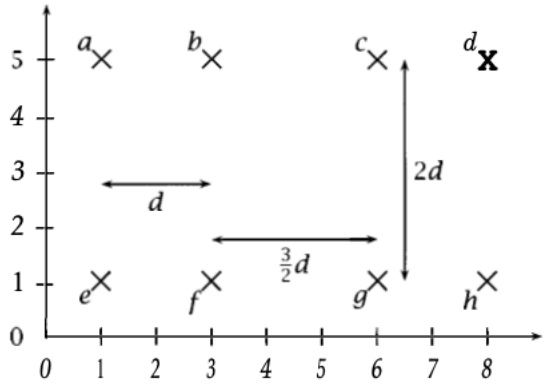
— la dispersion intra-clusters est basse, c'est-à-dire que les clusters doivent contenir des éléments très similaires entre eux ;

— la dispersion inter-clusters élevée, c'est-à-dire que les clusters doivent être différents les un des autres.

2 Exercices

2.1 Comparaison de distances entre clusters (Manning & Schütze)

Supposons le jeu de données suivant :



1. Donnez la séquence de partition et le dendrogramme obtenu en utilisant les deux notion de distance entre clusters suivantes :
   (a) *single linkage*
   (b) *complete linkage*
2. Comparez les partitions qui divisent les données en 2 clusters. Que pouvez-vous dire à propos de la différence entre *single linkage* et *complete linkage* ?

2.2 Décomposition de la dispersion (Robin & Vittaut)

Soit $X \subseteq \mathbb{R}^m$ un ensemble de points et $\pi = \{C^{(1)}, ..., C^{(k)}\}$ une partition de X. Pour simplifier les notations, on notera :

$$n = |X|\,,$$
$$n_i = |C^{(i)}|.$$

Le centroïde du des données X, que l'on écrit $\bar{\boldsymbol{x}}$, est défini comme :

$$\bar{\boldsymbol{x}} = \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} \boldsymbol{x}\,.$$

Le centroïde du cluster i, que l'on écrit $\boldsymbol{m}^{(i)}$, est défini comme :

$$\boldsymbol{m}^{(i)} = \frac{1}{|C^{(i)}|} \sum_{\boldsymbol{x} \in C^{(i)}} \boldsymbol{x}\,.$$

1. Montrer que :
$$\bar{\boldsymbol{x}} = \sum_{i=1}^{k} \frac{n_i}{n} \boldsymbol{m}^{(i)} .$$

2. Montrer que :
$$\sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C^{(i)}} \boldsymbol{x}^\top \boldsymbol{m}^{(i)} = \sum_{i=1}^{k} n_i \|\boldsymbol{m}^{(i)}\|_2^2$$

3. Montrer que :
$$\sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C^{(i)}} \boldsymbol{m}^{(i)\top} \bar{\boldsymbol{x}} = n \|\bar{\boldsymbol{x}}\|_2^2$$

4. Montrer que :
$$\sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C^{(i)}} \boldsymbol{x}^\top \bar{\boldsymbol{x}} = n \|\bar{\boldsymbol{x}}\|_2^2$$

5. Montrer que :
$$\underbrace{\sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C^{(i)}} \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2^2}_{\text{dispersion autour de la moyenne des données}} = \underbrace{\sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C^{(i)}} \|\boldsymbol{x} - \boldsymbol{m}^{(i)}\|_2^2}_{\text{dispersion intra-clusters}} + \underbrace{\sum_{i=1}^{k} n_i \|\boldsymbol{m}^{(i)} - \bar{\boldsymbol{x}}\|_2^2}_{\text{dispersion inter-clusters}} .$$

6. Comment interpréter ce dernier résultat ?