TD Méthodes de clustering (Réponses)
M1-ATD
Ioannis Partalas, Eric Gaussier

D'après M.-R. Amini & E. Gaussier, *Recherche d'information*, Eyrolles 2013

1. Algorithme des k-moyennes

(a) Montrer qu'entre les itérations t et $t+1$ de l'algorithme 1 on a :

$$\mathcal{L}(G_1^{(t+1)}, \ldots, G_K^{(t+1)}; \mathbf{r}_1^{(t)}, \ldots, \mathbf{r}_K^{(t)}) \leq \mathcal{L}(G_1^{(t)}, \ldots, G_K^{(t)}; \mathbf{r}_1^{(t)}, \ldots, \mathbf{r}_K^{(t)})$$

Réponse : L'inégalité découle de la définition de la réaffectation des exemples :

$$G_k^{(t+1)} \leftarrow \{d : ||\mathbf{d} - \mathbf{r}_k^{(t)}||_2^2 \leq ||\mathbf{d} - \mathbf{r}_l^{(t)}||_2^2, \forall l \neq k, 1 \leq l \leq K\}$$

(b) Montrer aussi que :

$$\mathcal{L}(G_1^{(t+1)}, \ldots, G_K^{(t+1)}; \mathbf{r}_1^{(t+1)}, \ldots, \mathbf{r}_K^{(t+1)}) \leq \mathcal{L}(G_1^{(t+1)}, \ldots, G_K^{(t+1)}; \mathbf{r}_1^{(t)}, \ldots, \mathbf{r}_K^{(t)})$$

Réponse : Considérons la fonction de coût pour une seule classe G :

$$\begin{aligned}
\mathcal{L}(G, z) &= \sum_{d \in G} \parallel d - z \parallel^2 \\
&= \sum_{d \in G} \parallel d - CG + CG - z \parallel^2 \\
&= \sum_{d \in G} \parallel d - CG \parallel^2 + \sum_{d \in G} \parallel CG - z \parallel^2 + 2 \sum_{d \in G} < d - CG, CG - z > \\
&= \sum_{d \in G} \parallel d - CG \parallel^2 + |G| \parallel CG - z \parallel^2 + 2 < \underbrace{\sum_{d \in G} d}_{|G|CG} - |G|CG, CG - z > \\
&= \mathcal{L}(G, CG) + |G| \parallel CG - z \parallel^2
\end{aligned}$$

Nous avons de ce fait :

$$\mathcal{L}(G_1^{(t+1)},\ldots,G_K^{(t+1)};\mathbf{r}_1^{(t+1)},\ldots,\mathbf{r}_K^{(t+1)}) \leq \mathcal{L}(G_1^{(t+1)},\ldots,G_K^{(t+1)};\mathbf{r}_1^{(t)},\ldots,\mathbf{r}_K^{(t)})$$

(c) En déduire que la fonction de coût de l'algorithme k-moyennes décroît à chaque itération.

Réponse : D'après les questions (a) et (b) et en exploitant la positivité de la fonction considérée nous avons $\forall t$:

$$0 \leq \mathcal{L}(G_1^{(t+1)},\ldots,G_K^{(t+1)};\mathbf{r}_1^{(t+1)},\ldots,\mathbf{r}_K^{(t+1)}) \leq \mathcal{L}(G_1^{(t)},\ldots,G_K^{(t)};\mathbf{r}_1^{(t)},\ldots,\mathbf{r}_K^{(t)})$$

La fonction $\mathcal{L}$ décroit donc à chaque itération de l'algorithme.

2. Classification par méthodes agglomératives hiérarchiques ascendantes

Question 1 Montrer que la méthode du lien unique est stable pour la meilleure fusion. On rappelle que la méthode du lien unique est fondée sur la distance entre classes :

$$\mathrm{sim}_{\mathrm{lu}}(G_k, G_l) \quad = \quad \max_{d\in G_k, d'\in G_l} sim(d,d')$$

Réponse : Nous avons :

$$\begin{aligned} sim(G_k^{(r+1)}, G^{(r)}) &= max(sim(G_k^{(r)}, G_{mf(k)}^{(r)}), sim(G_k^{(r)}, G_l^{(r)})) \\ &= sim(G_k^{(r)}, G_{mf(k)}^{(r)}) \end{aligned}$$

Question 2 Quelle est la complexité de l'algorithme 3 qui utilise un tableau de meilleure fusion en lieu et place des files de priorité ?

Réponse : $O(N^2)$

Question 3 Expliquer pourquoi la stabilité de la meilleure fusion est importante pour le bon déroulement de cet algorithme (donner un exemple simple).

Réponse : La stabilité est importante car nous permet d'utiliser un tableau de meilleure fusion et ainsi mettre à jour la matrice C à chaque itération en $O(N)$. On peut considérer un exemple avec 4 documents et l'algorithme du lien complet et montrer qu'on ne peut pas utiliser un tableau de meilleure fusion car la stabilité de la meilleure fusion ne s'applique pas.

Question 4 Montrer que la méthode par lien unique est monotone.

Réponse : Si $G_1^{(r+1)} \neq G^{(r)}$ et $G_2^{(r+1)} \neq G^{(r)}$, l'inégalité découle de la construction même du dendrogramme (sinon les deux classes $G_1^{(r+1)}$ et $G_2^{(r+1)}$ auraient été fusionnées avant les deux classes $G_1^{(r)}$ et $G_2^{(r)}$).

Supposons donc que $G_1^{(r+1)} = G^{(r)}$. Nous avons :

$$sim(G_1^{(r+1)}, G_2^{(r+1)}) = \max_{d \in G_1^{(r)} \cup G_2^{(r)}, d' \in G_2^{(r+1)}} sim(d, d')$$

Cette dernière quantité est équivalente à :

$$\max(\max_{d \in G_1^{(r)}, d' \in G_2^{(r+1)}} sim(d, d'), \max_{d \in G_2^{(r)}, d' \in G_2^{(r+1)}} sim(d, d'),)$$

et :

$$\max(sim(G_1^{(r)}, G_2^{(r+1)}), sim(G_2^{(r)}, G_2^{(r+1)}))$$

Mais $sim(G_1^{(r)}, G_2^{(r+1)}) \leq sim(G_1^{(r)}, G_2^{(r)})$ car sinon ce sont les classes $G_1^{(r)}$ et $G_2^{(r+1)}$ qui auraient été fusionnées à l'étape r. De même, $sim(G_2^{(r)}, G_2^{(r+1)}) \leq sim(G_1^{(r)}, G_2^{(r)})$, ce qui montre la monotonicité du lien simple.