Projet : l'algorithme PAM

October 20, 2021

Date de remise : Mercredi 8 Décembre.

Constitution des groupes : groupes de 3 étudiants faisant partie obligatoirement du même groupe de TD. Si le nombre d'étudiants n'est pas un multiple de 3 (et uniquement dans ce cas), on peut admettre un ou deux groupes de 4.

L'algorithme PAM est une alternative à l'algorithme des K-means. En entrée on a K un nombre de clusters qui seront retournés par l'algorithme, et D un data set de N objets. En sortie, on obtient K clusters. L'idée de l'algorithme est de construire les clusters autour d'objets représentatifs. Chaque cluster est associé à un objet qui le représente. Noter immédiatement qu'un objet représentatif est un objet de l'ensemble D, alors que le barycentre dans les K-means n'est pas nécessairement dans la population. L'algorithme PAM procède comme suit

1. Choisir arbitrairement (aléatoirement) K objets dans D comme représentation initiale (ou graine) des clusters.
2. répéter
   (a) Affecter chaque objet non représentatif dans le cluster associé à l'objet représentatif qui est le plus similaire (le plus proche au sens de la distance ou la similarité retenue).
   (b) Pour tout objet représentatif m (et donc le cluster associé) et pour tout objet o dans D qui ne soit pas un objet représentatif.
      i. soit E le coût actuel du partitionnement, calculer le coût S de la partition dans laquelle o est un objet représentatif à la place de m
      ii. Si on améliore le coût (is $S < E$), garder la combinaison m et o ainsi que le gain en coût pour le choix du meilleur échange
   (c) Si le meilleur échange (au sens de la diminution du coût) permet de diminuer le coût, effectuer le meilleur échange (m, o)
3. jusqu'à ce qu'il n'y ait plus de changement dans les affectations des objets

On peut utiliser une similarité ou une distance pour cet algorithme. On vous propose d'utiliser la distance L_1 pour ce dataset. Les coûts S et E se calculent comme suit:

- On calcule une matrice de distance (de Manhattan, ou L_1) appelé $d(i, j)$ pour la distance entre i et j.
- Pour une partition $C_1, ..., C_i, ..C_K)$, on cherche à minimiser le coût $\sum_{i=1}^{K} \sum_{j \in C_i} d(i, j)$. Si d est une distance, alors $d(i, i) = 0$. On peut penser à séparer les points représentatifs des autres points pour améliorer la complexité du calcul de E.

1. Programmez en C l'algorithme.

2. Etablissez la complexité de chaque itération de l'algorithme PAM.

3. Comparer les résultats de l'algorithme K-means et de l'algorithme PAM sur le dataset XXX. Ce dataset sera bientôt disponible sur le Moodle du cours. Pour les K-means utiliser R (voir le cours et le TD).