ALGO3 – Algorithmique et Programmation 3

TP4

Clustering par les k-moyennes (K-means)

Le clustering est le processus de partitionnement d'un groupe de points de données en un petit nombre de clusters. Par exemple, les articles dans un supermarché sont regroupés dans des catégories (beurre, fromage et le lait sont regroupés dans les produits laitiers). Bien sûr, cela est une sorte qualitative de partitionnement. Une approche quantitative serait de mesurer certaines caractéristiques des produits, dire pourcentage de lait et d'autres, et des produits à haut pourcentage de lait serait regroupées.

En général, nous avons n points de données x_i, $i = 1 \dots n$ qui doivent être partitionné en k clusters. L'objectif est d'attribuer chaque point à un cluster. K-means est une méthode de classification qui vise à trouver la position μ_i, $i = 1 \dots k$ des clusters qui réduisent au minimum la distance entre les points au sein de chaque cluster. K-means résout de clustering :

$$\sum_{i=1}^{k} \sum_{x \in c_i} d(x, \mu_i) = \sum_{i=1}^{k} \sum_{x \in c} ||x - \mu_i||_2^2$$

où c_i est l'ensemble des points qui appartiennent au cluster i. Le K-means utilise le carré de la distance euclidienne $d(x, \mu_i) = ||x - \mu_i||_2^2$. Ce problème n'est pas trivial dans le sens où K-means ne permet pas toujours de trouver le minimum global car il peut coincé dans une solution différente.

Algorithme K-means

L'algorithme de Lloyd, surtout connu comme l'algorithme k-means, est utilisée pour résoudre le problème des k-moyennes et fonctionne comme suit. Tout d'abord, il faut décider du nombre des k clusters. Ensuite :

Etape 1 : On choisit aléatoirement k individus comme centres initiaux des classes.

Etape 2 : On attribue chaque objet à la classe la plus proche, ce qui définit k classes

Etape 3 : Connaissant les membres de chaque classe on recalcule les centres d'inertie de chaque classe.

Etape 4 : On redistribue les objets dans la classe qui leur est la plus proche en tenant des nouveaux centre de classe calculés à l'étape précédente.

Etape 5 : On retourne à l'étape 3 jusqu'à ce qu'il y ai convergence, c'est-à-dire jusqu'à ce qu'il n'y ai plus aucun individu à changer de classe.

L'algorithme finit toujours par converger vers un point, bien qu'il soit pas nécessairement le minimum de la somme des carrés. Cela est dû au faite que le problème est non-convexe et que l'algorithme constitue juste une heuristique, convergeant vers un minimum local. L'algorithme arrête lorsqu'aucun changement n'a lieu d'une itération à l'autre.

Question : L'objectif de ce TP est d'implémenter l'algorithme des k-moyennes.