Parallel k-means Clustering

SF2568 – Parallel Computations for Large-Scale Problems

Lukas Bjarre **Gabriel Carrizo**
lbjarre@kth.se **carrizo@kth.se**

---

1 k-means clustering

k-means clustering is a data clustering method which clusters input data from the data set $\mathcal{X}$ into k different classes. The classes are represented by the class means μ_i and points are considered to be in a class S_i if the squared distance to the class mean is the minimum compared to the squared distance to the other class means. Formally:

$$S_i = \{x \in \mathcal{X} : ||x - \mu_i||^2 \leq ||x - \mu_j||^2, \forall 1 \leq j \leq k\}$$

A clustering method aims to find a selection of these classes $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$ which divides the data points in some favorable way. k-means finds the placement of the class means by minimization of the summed squared distance of all class points to the class mean for all k classes:

$$\mathcal{S}_{k\text{-means}} = \arg\min_{\mathcal{S}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

A common algorithm to find this is Lloyd's algorithm, which iteratively classifies points according to current class means and updates them with the average of all classified points until convergence.

while $\forall \mu_k \neq \mu_k^{(new)}$ **do**
  for $\forall x \in \mathcal{X}$ **do**
    class $\leftarrow \min_k ||x - \mu_k||^2$;
    count[class]++;
    $\mu_k^{(new)} \leftarrow \mu_k^{(new)} + x$;
  end
  for $i = 1, \ldots, k$ **do**
    $\mu_k^{(new)} \leftarrow \frac{\mu_k^{(new)}}{\text{count}[i]}$;
  end
end

Algorithm 1: Lloyd's algorithm for finding the k-means clustering class means.