Coreset Tree

Upwork project

definitions of coreset

1. Coreset is a small weighted point set that approximates the points from the data stream with respect to the k-means clustering problem.
2. Coreset for a set P is a small weighted set, such that for any set of k cluster centers the (weighted) clustering cost of the coreset is an approximation for the clustering cost of the original set P with small relative error.
3. (k,ε)-coreset is that the clustering cost of the coreset for any arbitrary set of k centers is within (1 ± ε) of the cost of the clustering for the original input

Class Node

The notebook contains 2 data structures. Node class defines a structure to create nodes of a tree
Class node works as follows:
Let's say we have observed N number of points in a data stream so far. Let this set be P. where cardinality of P is N
Each node stores four attributes:

1. A subset of the set P. let's name this subset P_i
2. A representative point of P_i called q_i
3. Cost
4. Weight

Cost is equal to the sum of distance of all points in Pi to qi. Weight is equal to cardinality of Pi.
SubsetFinder function works as follows:

1. Calculate the cost and select the point Xm with maximal distance
2. Create two subsets of Pi named S1 and S2
3. S1 contains points of Pi nearest to qi
4. S2 contains points of Pi nearest to Xm

Class Coreset Tree

A coreset tree T for a point set P is a binary tree that is associated with a hierarchical divisive clustering for P. One starts with a single cluster that contains the whole point set P and successively partitions existing clusters into two subclusters, such that the points in one subcluster are far from the points in the other subcluster. The division step is repeated until the number of clusters corresponds to the desired number of clusters. Associated with this procedure, the coreset tree T has to satisfy the following properties:

1. Each node of T is associated with a cluster in the hierarchical divisive clustering.

2. The root of T is associated with the single cluster that contains the whole point set P.

3. The nodes associated with the two subclusters of a cluster C are the child nodes of the node associated with C.

Visit function

This function recursively visits all the leaf nodes of a tree and selects the leaf which has the highest cost among all the leaf nodes. Let's name this "Lmax"

Add Child function

This function recursively visits all the leaf nodes of a tree and finds the "Lmax" node and creates two child nodes for "Lmax" based ono kmeans++ initialization procedure

Propagate Up function

The cost of parent node is the sum of cost of its child nodes. This function recursively propagates the cost of leaf nodes upwards till it reaches the Root node.