CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 22 — October 23, 2023

Prof. Samson Zhou *Scribe: Matthew Chang*

1 Clustering

Definition (Clustering)**.** Given input dataset X, partition X so that "similar" points are in the same cluster and "different" points are in different clusters.

1.1 k-clustering

k-clustering provides a parameter to the clustering problem where there can be at most k different clusters.

To measure the "quality" of a clustering partition, we need to assign a "center", c_i, to each of the k clusters and then define a cost function.

The cost function is induced by c_i for all of the points P_i assigned to cluster i. We define $\text{Cost}(P_i, c_i)$ to be a function of $\{dist(x, c_i)\}_{x \in P_i}$ where the distance function is the distance in metric space between a point and its respective cluster center.

1.2 Types of k-clustering

Below we give 4 types of k-clustering and their respective cost functions.

- k-center: $\text{Cost}(X, C) = \max_{x \in X} \text{dist}(x, C)$
- k-median: $\text{Cost}(X, C) = \sum_{x \in X} \text{dist}(x, C)$
- k-means: $\text{Cost}(X, C) = \sum_{x \in X} (\text{dist}(x, C))^2$
- (k, z)-clustering: $\text{Cost}(X, C) = \sum_{x \in X} (\text{dist}(x, C))^z$

For our cases, we apply Euclidean k-clustering which means that our distance function is measured as the Euclidean distance in d dimensions.

$\text{dist}(x, y) = \sqrt{(x_1 - y_1)^2 + ... + (x_d - y_d)^2}$

1.3 Coreset

Definition (Coreset)**.** Subset X' of representative points of X for a specific clustering objective where the $\text{Cost}(X, C) \approx \text{Cost}(X', C)$ for all sets C with $|C| = k$

More formally, given a set X and an accuracy parameter $\varepsilon > 0$, we say a set X' with weight function w is a $(1+\varepsilon)$-multiplicative coreset for a cost function Cost, if for all queries C with $|C| = k$, we have:

$(1-\varepsilon)\,\mathrm{Cost}(X, C) \leq \mathrm{Cost}(X', C, w) \leq (1+\varepsilon)\,\mathrm{Cost}(X, C)$

2 Clustering in the Streaming Model

Clustering in the streaming model can be solved with the merge-and-reduce framework. For now, we assume that we have an algorithm for $(1+\varepsilon)$-coreset construction that uses $f(k, \frac{1}{\varepsilon})$ weighted input points.

2.1 Merge-and-Reduce

The merge-and-reduce framework is outlined below.

1. Partition the stream into blocks containing $f(k, \frac{\log n}{\varepsilon})$
2. Create a $(1 + \frac{\varepsilon}{\log n})$-coreset for each block
3. Create a $(1 + \frac{\varepsilon}{\log n})$-coreset for the set of points formed by the union of two coresets for each block

Step 3 is repeated on multiple levels until we end up with a single coreset. The algorithm is named "merge-and-reduce" because the repeated, basic building block of the algorithm is this: two $(1 + \frac{\varepsilon}{\log n})$-coresets are merged and then reduced into a single $(1 + \frac{\varepsilon}{\log n})$-coreset.

2.2 Analysis

Since there are $O(\log n)$ levels and each coreset is a $(1 + \frac{\varepsilon}{\log n})$-coreset of two coresets, we end up with a single coreset with a total approximation of $(1 + \frac{\varepsilon}{\log n})^{\log n} = (1 + O(\varepsilon))$.

Because every pair of coresets are merged and reduced, the memory requirement for merge-and-reduce is bounded by $O(\log n)$.