IT584 Approximation Algorithms January 23, 2024

Lecture 07

Lecturer: Rachit Chhaya *Scribe/s: Vimal Sonagara*

1 The K-Center Problem

In this case study, we will examine a big dataset consisting of data points that vary in degree of resemblance. Possibly in a predetermined number of clusters, the goal is to combine related data. A selection of data points must be chosen as cluster centres in order to accomplish this. As a result, we are able to identify the clusters by matching each data point to the closest cluster centre.

Studying this issue is valuable because it illuminates the underlying structure of our data and the relationships between different data elements. It also has useful applications in a variety of real-world scenarios. Users of Netflix, for example, have a variety of tastes in films. Users with similar movie tastes may be grouped together to recommend related movies.

1.1 Problem Defination

Input : Given an undirected complete graph $G = (V, E)$ with distances $d_{ij} \geq 0$ between each pair $i, j \in V$, where the distances follow the "Metric" rule a positive integer K.

Goal : Find k clusters grouping together the vertices that are most similar to each other, denoted by $S \subseteq V$ with $|S| = k$. Each vertex will assign itself to its closest cluster center. For the K-Center problem, the goal is to minimize the maximum distance of a vertex to its cluster:

1.2 Geometric Interpretation

. The distance function $d(\cdot)$ must satisfy the following properties:

1. Positive Semidefiniteness: $d(x, y) \geq 0$ for all $x, y \in V$, and $d(x, y) = 0$ if and only if $x = y$.
2. Symmetry: $d(x, y) = d(y, x)$.
3. Triangle Inequality: $d(x, y) \leq d(x, z) + d(z, y)$.

Finding a set S of k vertices—also known as cluster centers—that minimises the maximum distance between every vertex and its cluster centre is our goal. The cluster centred on the nearest $s \in S$ is designated as $i \in V$. We refer to the acronym as

$$d(i, S) = \min_{s \in S} d(i, s), \quad \text{where } s \in S$$

Given our cluster centers S, we denote the radius of S as follows:

$$r = \max_{i \in V} d(i, S)$$

Find a set of size K of minimum radius

1.3 A Greedy Algorithm

The algorithm is as follows:Initially, select an arbitrary vertex $i \in V$ and include it in our set S of cluster centers. Subsequently, it is logical for the next cluster center to be positioned as distantly as possible from all other existing cluster centers. While $|S| < k$, iteratively identify a vertex $j \in V$ for which the distance $d(j, S)$ is maximized (or, in simpler terms, determines the diameter of set S), and add it to S. The process halts once $|S| = k$, and the algorithm returns S.

The Greedy Algorithm for K-Centering

Pick an arbitrary $s \in V$ and initialize $S = \{s\}$. Do while $|S| < k$:

1. $s \leftarrow \arg\max_{s \in V} d(sS)$
2. Update $S \leftarrow S \cup \{s\}$

Approximation Analysis

Claim The greedy algorithm is a 2-approximation algorithm for the k-center problem.

Proof: Let $S^* = \{\mathrm{s}_1, s_2, \ldots, s_k\}$ represent the optimal solution, and r^* be its radius. This optimum arrangement divides the nodes V into clusters $\{V_1^*, V_2^*, \ldots, V_k^*\}$, with each $i \in V$ assigned to V^* if it is the nearest to s among all points in S^*. Ties are resolved arbitrarily.

Initially, note that for any pair of points i and j belonging to the same cluster V_l^*, their maximum separation is $2r^*$. In accordance with the triangle inequality, the distance $d(i, j)$ is, at most, the sum of $d(i, s_l)$, representing the distance from i to the center s_l, and $d(j, s_l)$, indicating the distance from the center s_l to j. Both these distances are restricted by r^*, thus:

$$d(i, j) \leq d(i, s_l) + d(j, s_l) \leq r^* + r^* = 2r^*$$

Consider the subset $S \subseteq V$ comprising the points chosen by the greedy algorithm. In the initial iteration, the algorithm selects a point $i \in V_l^*$ for addition to S, despite having previously selected a point $i' \in V_l^*$ in an earlier iteration. Before adding i', each center incorporated into S was chosen from distinct optimal clusters of S^*. For all points j covered by centers added before i', it is imperative that $d(j, S) \leq r^*$ based on the preceding argument. Since the greedy algorithm consistently chooses points farthest from the current set of points in S, the distance for any other point $j \in V$ not yet included in S must be constrained by:

$$d(j, S) \leq d(i, i')$$

In any case, j should have been added to S preceding i'. However, since (i) and (i') are members of the same optimal cluster V_l^*, $d(i, i') \leq 2r^*$ Thus, for all focuses covered later i' is added to the cluster centeres, the distance among it and its closest focus is likewise limited by all things considered $2r^*$. We know that the distance is bounded by $d(i, S) \leq 2r^*$ for all points in V. If r addresses the range of S returned by our Greedy soltuions, we can say that:

$$r^* \leq r \leq 2r^*$$

indicating that the algorithm provides a 2-approximation as we stated that.