Lecture 8

Barna Saha

AT&T-Labs Research

October 3, 2013

Outline

K-Center

- Given a set of distinct points $P = \{p_1, p_2, \ldots, p_n\}$ find a set of k points $Q \subset P$, $|Q| = k$, that minimizes

$$\max_i \min_{q \in Q} d(p_i, q)$$

  where d is any metric.

Suppose the optimal distance is r. If we know r, can find 2-approx in $O(k)$ space.

Thresholded Algorithm When a new point comes, if the minimum distance of this point from already opened centers is more than $2r$, open a center at that point. Else, assign it to the nearest open center.

Can find $(2+\epsilon)$ approximation in $O(\frac{k}{\epsilon} \log b/a)$ space if we know

$$a < r < b$$

Theorem

$(2+\epsilon)$-approximation in $O(\frac{k}{\epsilon} \log \frac{1}{\epsilon})$ space.

K-Center-Algorithm

- Read the first k items in the input. This has error 0. Keep reading the input as long as the error remains 0.
- Suppose, we see the first input which causes non-zero error. This gives a lower bound a for r.
- Initialize and run the thresholded algorithm for $l_0 = a, l_1 = a(1+\epsilon'), l_2 = a(1+\epsilon)^2, ..., l_J = a(1+\epsilon)^J = O(\frac{1}{\epsilon})$.
- If the thresholded algorithm declares "FAIL" (tries to open $k+1$ centers) for some l_i, $i \in [1, J]$, terminate the algorithm for all $l_{i'}, i' \leq i$. Start running a thresholded algorithm for $l_{i'}(1+\epsilon')^{J+1}$ for $i' \in [0, i]$ using summarization of threshold $l_{i'}$ as the initial input.[Stream-Strapping]
- Repeat the above steps until the end of input. At that time report the centers for the lowest estimate for which the thresholded algorithm is still running.

K-center, Sketch Analysis

- Suppose end threshold is R and it is updated i times: $R_0, R_0(1+\epsilon')^{J+1}, R_0(1+\epsilon)^{2(J+1)}, ..., R_0(1+\epsilon)^{i(J+1)}$
- $i = 0$. $Q_1 = P_1 = [p_1, p_2, .., p_j]$

$$Error(Q_1) = Error(P_1) \leq 2R_0$$

$$OPT(Q_1) > \frac{R_0}{(1+\epsilon')}$$

$$Error(Q_1) \leq 2R_0 \leq (2+2\epsilon)OPT(Q_1)$$

- $i = 1$ $Q_2 = [q_1, q_2, ..., q_k, p_{j+1}, p_{j+2}, .., p_{j'}] =$, $P_2 = p_{j+1}, p_{j+2}, .., p_{j'}$. Terminates with $R_1 = R_0(1+\epsilon)^{J+1}$ but not with $\frac{R_1}{(1+\epsilon)}$.

$$Error(Q_2) \leq 2R_1$$

$$OPT(Q_2) > \frac{R_1}{1+\epsilon}$$

$$Error(Q_2) \leq 2R_1 = (2+2\epsilon)OPT(Q_2)$$

K-center, Sketch Analysis

- Relationships between $Error(Q_2)$ and $Error(P_1 \bigodot P_2)$ and in between $OPT(Q_2)$ and $OPT(P_1 \bigodot P_2)$
  1. $Error(P_1 \bigodot P_2) \leq Error(Q_2) + Error(Q_1) \leq 2R_1 + 2R_0 = 2R_1 \left(1 + \frac{1}{(1+\epsilon)^{J+1}}\right)$
  2. $OPT(P_1 \bigodot P_2) \geq OPT(Q_2) - Error(Q_1) \geq \frac{R_1}{(1+\epsilon)} - 2R_0 = \frac{R_1}{(1+\epsilon)} \left(1 - \frac{2}{(1+\epsilon)^J}\right)$

K-Median

- When we know the optimum solution r: Set $f = \frac{r}{k(1+\log n)}$
- When considering point x, let δ be the distance to the nearest open center. Open a center at x with probability $\frac{\delta}{f}$. Else, assign to the nearest open center.

K-Median

Setting the initial estimate Error after reading $k+1$th point.

How many copies to maintain ? $O(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$. But needs $O(\frac{1}{\epsilon}\log n)$ copies of Stream-Strap to boost the confidence.

When to declare an individual estimate is wrong ? If error becomes more than $4(1+\epsilon)L$ or open more than $k' \simeq k\frac{\log n}{\epsilon'}$ centers.

Initial Summary k' centers weighted by the number of points assigned to those centers.

Final Output Run K-median offline algorithm on the selected k' weighted centers.

K-Means++

- Extension of K-means clustering: minimizes within cluster sum of squared error.
- Initial choice of centers is crucial to guarantee quicker convergence and approximation bound.