CMPUT 675: Algorithms for Streaming and Big Data **Fall 2019**

Lecture 18 (Nov 6, 2019): Clustering Problems and Matrix Multiplication

Lecturer: Mohammad R. Salavatipour *Scribe: Candelario Gutierrez*

18.1 Clustering Problems

We will discuss one specific clustering problem and present an algorithm in the streaming model for it.

18.1.1 Definition

A metric space (X, d) where X is a non-empty set and $d : X \times X \to \mathbb{R}$ is a function that assigns values to pairs of points such that for all $x, y \in X$):

1. $d(x, y) = 0 \longleftrightarrow x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Typically, given a set $P \subseteq X$, an integer parameter k, the goal is to partition P into k parts/clusters $C_1, \ldots, C_k$, each having a center c_i to optimize (minimize) some function:

- k-Centre
  - $\underbrace{cost(P, C) = min}_{\triangle_\infty(P,C)} \max_{p \in P} \min \{d(p, c_i)\}$
  - Has 2-approximation in offline setting and $(2 - \varepsilon)$-hard to approximate.
- k-Median
  - $\underbrace{cost(P, C) = min}_{\triangle_1(P,C)} \sum_{p \in P} min_{c_i} \{d(p, c_i)\}$
- k-Means
  - $\triangle_2(P, C) min \sum_{p \in P} min_{c_i}\, d(p, c_i)^2$

Lemma 1 *Suppose $x_1, \ldots, x_{k+1} \in P$ are points, such that $d(x_i, x_j) \geq \varepsilon$ for $\forall x_i, x_j$. Then for all sets $C \subseteq X$, $\Delta_\infty(P, C) \geq \frac{\varepsilon}{2}$.*

Proof. By way of contradiction suppose that there is a set $C \subseteq X$ such that $\Delta_\infty(P, C) < \varepsilon/2$. Then no two points x_i, x_j can be in the same cluster we need at least $k + 1$ clusters to cover all of them. ■

We present an 8-approximation for the k-center problem in the streaming setting. The algorithm, proposed by M. Charikar, C. Chekuri, T. Feder and R. Motwani called the doubling algorithm[CCFM04]. The algorithm always maintains a set C, $|C| = k$ of centres, uses $O(k)$ space and matains a threshold τ as estimate of optimum.

Doubling algorithm

1) Let S be the first $k+1$ points of stream
2) Let $(x,y) \leftarrow arg\ min_{(u,v)\epsilon S}\ d(u,v)$
3) $\tau \leftarrow d(x,y)$, $C \leftarrow S \backslash \{x\}$
4) For each new point p do
5) if $min_{c\epsilon C}\ d(p,c) > 2\tau$ then
6) $C \leftarrow C \cup \{p\}$
7) while $|C| > k$ do
8) $C \leftarrow$ *maximal* $C' \subseteq C$ such that $\forall u,v \in C' : d(u,v) \geq 2\tau$
9) $\tau \leftarrow 2\tau$

Lemma 2 *The following invariants hold for the doubling algorithm each time Step 4 is executed:*

1. $\forall u,v \in C : d(u,v) \geq \tau$.
2. $\Delta_\infty(\sigma, C) \leq 2\tau$.

and the following holds immediately after step 3 and before execution of Step 8: $\forall C^* \subseteq X$ *with* $|C^*| = k$*:* $\Delta_\infty(\sigma, C^*) \geq \tau/2$.

Proof. By induction: The base case is right after initialization. We have that not only $|C| = k$, but also $d(u,v) \geq \tau$ where this holds for the first $k+1$ points and proprty 2) follows immediately follows. Property 3) follows from the previous lemma.

For induction step we have two cases based on $\min_{c\in C} d(p,c)$. In the case that the "if condition" of line (5) is not satisfied, nothing changes and is easy to verify that all 3 invariants hold.

So, lets assume that it is satisfied, i.e. $min_{c\in C} d(p,c) > 2\tau$. Then $\Delta_\infty(\sigma \cup \{p\}, C \cup \{p\}) = \Delta_\infty(\sigma, C) \leq 2\tau$, so Step 6) preserves property 2. Also, using induction hypothesis and since $min_{c\in C} d(p,c) > 2\tau$, property 1 is preserved. But $|C| = k+1$, so previous lemma together with 1) implies that property 3 is preserved. We have to show that properties are maintained after execution of lines 8) and 9). Note that property 1) will be preserved since each time we select C' in line 8) we have $d(u,v) \geq 2\tau$ (so we have property 1 even after doubling τ in line 9). Therefore, we need to show that: $\triangle_\infty(\sigma, C') \leq 2\tau$.

Let x be an arbitrary point from σ. By property 2) $d(x,C) \leq \Delta_\infty(\sigma, C) \leq 2\tau$ before line 8 is executed. Let r be the closest point to x in C. If $r \in C'$ then $d(x,C') \leq d(x,r) \leq 2\tau$. If $r \notin C'$ then since $C' \subseteq C$ is maximal, there is a point $s \in C'$ such that $d(r,s) < 2\tau$. Since $s \in C'$ we have:

$$d(x,C') \leq d(x,s) \leq d(x,r) + d(r,s) < 2\tau + 2\tau = 4\tau.$$

So after line 8), we have $\Delta_\infty(\sigma, C) < 4\tau$ and after line 9) this implies $\Delta_\infty(\sigma, C) \leq 2\tau$, so property 2) still holds. ■

Using property 3) and observing that step 9) doubles the value of τ, and using property 2) we obtain:

Theorem 1 *The doubling algorithm is an* 8*-approx.*

18.2 Matrix Multiplication

Given two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ we would like to compute $C = A \times B$. The standard matrix multiplication takes $O(mnp)$ time. For large values of m, n, p this is way too slow. By using Strassen's like divide and conquer algorithms we can improve the run time.

Our goal is to find an approximate answer fast. More specifically, we would like to find C such that $\| C-AB \|_x < \varepsilon$ with probability $1-\delta$ for some measure of norm.

Frobenius norm: treat matrices like vectors and

$||A||_F = (\sum_{i,j} A_{ij}^2)^{\frac{1}{2}}$. It can be shown that m $||AB||_F \leq ||A||_F||B||_F$.

Spectral norm: $Sup_{||x||_2=1}||Ax||_2$. Aagain, one can show that $||AB||_2 \leq ||A||_2||B||_2$.

For any matrix M, let $M_{i,j}$ denote row i and $M^{(j)}$ denote column j

$$c_{ij} = \langle A_{(i)}, B^{(j)} \rangle = \sum_{k=1}^{n} A_{ik}B_{kj}$$

$$AB = \sum_{j=1}^{n} A^{(j)}B_{(j)}$$

Here we present an approximate multiplication algorithm (proposed by P. Drineas, R. Kannan and M. Mahone [DKM06].)

We first pick a probability distribution over $[n]$, $p_1 + p_2 + ... + p_n = 1$ and then for $k = 1$ to t pick an index $j_k \in [n]$ according to p_{j_k} and let

$$c = \frac{1}{t}\sum_{k=1}^{t} \frac{1}{p_{j_k}} A^{(j_k)}B_{(j_k)}$$

Note that $C = \frac{1}{t}\sum_{k=1}^{t} c_k$ and $E[c_k] = \sum_{\ell=1}^{n} p_\ell \cdot \frac{1}{p_\ell} A^{(\ell)}B_{(\ell)} = AB$, thus $E[c] = AB$.

Suppose that we want $||C-AB||_F \leq \varepsilon||A||_F||B||_F$. We choose $p_j's$ such that each one corresponds to contribution of $A^{(j_k)}B_{(j_k)}$ to $||AB||_F$. By using the fact that

$$\| A^{(j)}B_{(j)} \|_2 = \| A^{(j)} \|_2 \| B_{(j)} \|_2$$

it turns out that using

$$p_j = \frac{\|A^{(j)}\|_2\|B_{(j)}\|_2}{\sum_\ell \|A^{(\ell)}\|_2\|B\|_2}$$

is optimized. In the next lecture we will continue with its expected value and probability.

References

CCFM04 M. Charikar, C. Chekuri, T. Feder, R. Motwani. Incremental Clustering and Dynamic Information Retrieval.*SIAM J. Comput.*, 33. 1417-1440, 2004.

DKM06 P. Drineas, R. Kannan, M. Mahoney. Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication.*SIAM J. Comput.*, 36. 132-157, 2006.