Next part of the course: Streaming algorithms for metric problems & graph problems.

Today: Clustering

We'll see: $8$-approx alg for $K$-center

$2 + O(e)$-approx for $K$-center

Data stream: sequence of tokens $\sigma_1, \sigma_2, \ldots, \sigma_m$

from $[n]$ or $\mathbb{R}^n$ or edges of $G$, etc.

Each elt is inserted in the stream sequentially.

Alg can store some info about stream seen so far.

Alg cannot access els not saved that appeared earlier (unless it makes more passes over the data).

Goal: Approximately solve an optimization problem

Use sublinear space in $m, n$
Ideally: $O(\log m, \log n)$ space to output a constant number of els of $[n]$.

Sometimes, cannot do better than $\Omega(m)$ (we'll see lower bounds later in the course).

Today: clustering in metric spaces.

Given a data stream of pts in $\mathbb{R}^n$ $\Gamma = \{ \Gamma_1, \Gamma_2, \ldots, \Gamma_m \} \subseteq \mathbb{R}^n$ output a good "summary" of $\Gamma$.

Reasonable summary: centers of $k$ clusters or representatives of each cluster $c_1, c_2, \ldots, c_k$. 
At this lecture work for more general metrics.

Metric space: \((M, d)\) \(d: M \times M \to \mathbb{R}^+\)

- \(d(x, y) = 0 \) \(\forall x = y\)
- \(d(x, y) = d(y, x)\)
- \(d(x, y) \leq d(x, z) + d(z, y)\)

Eq: \(\mathbb{R}^n \times \mathbb{R}^n\)

\[d(x, y) = \|x - y\|_p^{\frac{1}{p}}\] \(p > 0\).

- \(p = 2\) Euclidean distance.
- \(p = \infty\) max \(\{x_i - y_i\}\)

For graphs \(M = V (\cup + x \text{ set})\)

\(d(x_1, y) = \text{length of shortest path}\)
Clustering

Extend \( d(x,y) \) to \( d(x, S) = \min_{y \in S} d(x, y) \)

Standard objectives: minimize

- \( \Delta \sigma (\sigma, R) = \max_{x \in \sigma} d(x, R) \)

(k-center)

We'll focus on k-center in this lecture.
Other cost measures:

\[ \Delta_1 (\mathcal{O}, R) = \sum_{x \in \mathcal{O}} d(x, R) \]

(k-median)

\[ \Delta_2 (\mathcal{O}, R) = \sum_{x \in \mathcal{O}} d(x, R)^2 \]

(k-means)

Goal: Choose \(|R| \leq k\) centers to

\[ \min \text{ cost } \Delta(\mathcal{O}, R) \]
Hochbaum & Shmoys (Doubling Alg.)

Init: \( R \leftarrow \text{set of first} \ k+1 \ \text{tokens (centers) } (c_1, c_2, \ldots, c_{k+1}) \)

\( z = \text{dist bet closest pair } (z, y) \)

\( R \leftarrow R \setminus \{ z \} \)

On new token \( x \)

. if \( \min d(x, r) \geq 2z \)

\( R \leftarrow R \cup \{ x \} \)

while \( |R| > k \)

. \( R \leftarrow \text{maximal set } R' \subset R \) \( s.t. \ \forall r, s \in R' \)

\( d(r, s) \geq 2z \)

\( R \leftarrow R \)

At end of stream output \( R \).
Analysis:

Invariant:

• separation: \( \forall r \neq s \in R \quad d(r,s) \geq 2 \)

• cost: \( \Delta_c(\sigma, R) \leq 2 \)

Verifying the invs:

inv 1. holds after init phase

assume it holds before processing
we'll show it hold after processing.

By design

inv 2: in initialize phase \( \Delta_c(\sigma, R) = 2 \)

Assume inv 2 holds before processing

\[ Z = 2, \quad \Delta(0, R') = 4 \]
what dist. an r that is kicked out of R is from R':
what dist. an elt y in the cluster of r (that was removed) is from R'

\[ r_2 \leftarrow \text{removed} \]
\[ r_2 \leftarrow \text{by assump} \]
\[ \text{by \Delta \text{ineq.}} \]
\[ 23 \leq 22 \]
\[ 3 \leq 21 = 22 \]

Thus: Alg above use \(0(K)\) space \& produces an 8-approximation to opt clustering.
Pf. Let $R^*$ is an opt clustering producing opt cost $\Delta_\infty (\sigma, R^*)$. Let $\hat{R}, \hat{\varepsilon}$ the final output

$$d (x, y) \geq \frac{\hat{\varepsilon}}{2} \text{ before final update to } \hat{\varepsilon}$$

$$\Delta_\infty (\sigma, \hat{R}) \leq 2 \hat{\varepsilon} \text{ (by inv. 2).}$$

How to relate $\Delta_\infty (\sigma, R^*)$ to $\hat{\varepsilon}$?

**Lemma:** Given $x_i, x_{k+1} \in \sigma$

such that $d (x_i, x_i) \geq t$ for $i, j \in [k+1]$.

Then $\forall R \in M$ of $|R| \leq k$.

$$\Delta_\infty (\sigma, R) \geq \frac{t}{2}$$
Two $x_i$'s represented by same $r_i$.

$$2 \max_{k,j} d(r_i, x_j) \geq d(x_{k+1}, r_k) + d(r_k, x_3) > t$$

$$\forall 2 \Delta_\infty (\sigma, R)$$

$$\Rightarrow \Delta_\infty (\sigma, R) \geq t/2$$

any $R$, so opt too
Applying Lemma to alg analysis:

Lem is true for opt $R^*$, so

$$\Delta_\infty (\gamma, R^*) \geq \frac{t}{2}$$

Recall every $r_i, r_j \in R^1$ are s.t

$$\text{dist} (r_i, r_j) \geq \frac{\varepsilon}{2}.$$

=>

$$\Delta_\infty (\gamma, R^*) \geq \frac{\varepsilon}{4}$$

↑

Recall output $\hat{R}$ s.t. $\Delta_\infty (\gamma, \hat{R}) \leq 2\varepsilon$

=>

$$\Delta_\infty (\gamma, \hat{R}) \leq 8 \Delta_\infty (\gamma, R^*)$$

↑

OPT.
Metric property of cost function:

\[ \Delta \left( \sigma, R \right) \text{ satis. metric prop if } \forall \sigma, \bar{w} \]

If summary \( S \) for \( \sigma \), \( \bar{w} \text{ summary } T \) for \( \sigma \bar{w} \)

\[ \Delta \left( \sigma, S \right) \leq \Delta \left( \sigma \bar{w}, T \right) \]

Above \( \Delta \left( \sigma \bar{w}, T \right) \)

\[ \leq \Delta \left( \sigma, S \right) \]

\[ \leq \Delta \left( \sigma, S \right) + \Delta \left( \sigma \bar{w}, T \right) \]

S is a summary for \( \sigma \)

T is a summary for \( \sigma \bar{w} \) or for \( \sigma \bar{w} \)

Ineq above says that if \( S \) is a good summary for \( \sigma \)

\( T \) is a good summary for \( \sigma \bar{w} \) then \( T \) is still a good summary for \( \sigma \bar{w} \); so we won't make much error on forgetting...
Claim: $\Delta(x, R)$ is a metric cost measure.

Pf idea: Prove defining ineq holds for every element of $\Gamma \circ \tilde{u}$.

How does $x$, $\text{rep}(x)$, $\text{rep}(x)$ compare?

$d(x, S)$ vs $d(x, T)$ vs $d(S, T)$

[Homework]
Def: Let $\Delta$ be an error cost func $x, \Delta \geq 1 \in \mathbb{R}$.

An $\alpha$-threshold alg for $\Delta$:

on input threshold $\tau$ and stream $\sigma$

Either (1): Alg produces a summary $S$ with $\Delta(\sigma, S) \leq \alpha \cdot \tau$

or

(2): Alg fails if no solution $T$ exists of error cost $\Delta(\sigma, T) < \tau$

[i.e., so for opt $\tau^*$ we must have $\Delta(\sigma, \tau^*) \geq \tau$]
Claim:
Doubling alg above is a 2-threshold alg.

Pf: Given thresh 2

either, outputs sol of cost \( \leq 2z \)

or, fails (no sol of cost < 2z exists)

Thus: If \( \Delta \) is a metric cost

& has an \( \Delta \)-approxim. thresh alg

Then \( \exists \) an \( \alpha + O(3) \)-approx alg.

Corollary: For \( \Delta \), we have a \( 2 + O(3)/\text{approx} \)

for k-center

in space is indep. of \( |0|, \text{val cost} \)
Proof of Theorem

Gupta's cascading alg:

• Read 1st B items of input
  while sum. cost = 0 \( (S_0 = 0) \)
  till instance of cost \( > 0 \). Let \( C \) be this cost.

• Run an \( \Delta \)-threshold alg in parallel
  for thresholds
  \[ c, c(1+\varepsilon), \ldots, c(1+\varepsilon)^J \]
  \( \frac{1}{3} \log \frac{\Delta}{\varepsilon} \) many times

  \[ \{ \text{each machine operates on small space} \} \]

• If some thread fails (no set of
  cost \( \Delta_{\infty} \leq t = c(1+\varepsilon)^i \) exists ) \( \leq \) by def
  of \( \Delta \)-thresh alg.
Then update all thresholds $t$ to $t = (1+3)^J c . (1+3)^J$.

$\frac{1}{3}$

- Start processing remaining stream from current summarization.
- Repeat until end of input.
- Output summary of an instance that is still running, with smallest threshold.
Thus: Cascading alg is an $\alpha + O(3^{-\epsilon})$-approx if run on an $\alpha$-thresholding alg, for metric cost $\Delta_3$ in space $O\left(\frac{1}{\epsilon} \log \frac{1}{\epsilon} \cdot \text{space of } \alpha \text{-thresh alg}\right)$.

Corollary: For $\Delta_\omega$ and $k$-center gives $2 + O(\sqrt{\epsilon})$-approx in space $O\left(\frac{k}{\epsilon^2} \log \frac{1}{\epsilon}\right)$.

Proof of Theorem:

Let $U_t$ be the threshold used by the instance with final output. Instance has lowest current threshold.

How many times this threshold changed?

Say $j$ times. The thresholds on that thread are:

$\frac{t_0}{t_1} = \frac{1}{1+\epsilon}$ for some $i$. So $t_i = \frac{t}{1+\epsilon}^{i-1}$. 

$\uparrow$

$\frac{t_0}{t_1} = \frac{1}{1+\epsilon}$ for some $i$. So $t_i = \frac{t}{1+\epsilon}^{i-1}$.
Consider stream $\sigma_i$ to be the stream occurring between the $i-1$ and $i$th change of threshold.

$$\sigma = \sigma_1 \circ \sigma_2 \circ \ldots \circ \sigma_{j+1}$$

Let $S_1$, $S_2$, $S_{j+1}$ be the summaries before each jump.

Note: $S_i$ is a summary of $S_{i-1} \circ \sigma_i$ (not of $\sigma_i$).

$S_0 = \emptyset$.

Claim: For

$$\Delta(\sigma_1 \circ \sigma_2 \circ \ldots \circ \sigma_i, S_i) \leq \Delta(S_{i-1} \circ \sigma_i, S_i)$$

$$+ \Delta(\sigma_1 \circ \sigma_2 \circ \ldots \circ \sigma_{i-1}, S_{i-1})$$
Proof: \[ \Delta(\sigma_1, S_1) \leq \Delta(\sigma_1, S_1) + \Delta(\sigma_0, S_0) \]

\[ \Delta(\sigma_1 \circ \sigma_2, S_2) \leq \Delta(\sigma_1 \circ \sigma_2, S_2) + \Delta(\sigma_1, S_1) \] (by metric property)

\[ \Delta(\sigma_1 \circ \sigma_2 \circ \sigma_3, S_3) \leq \Delta(\sigma_1 \circ \sigma_2, S_2) + \Delta(\sigma_2 \circ \sigma_3, S_3) \]

\[ \Delta(\sigma_1 \circ \sigma_2 \circ \ldots \circ \sigma_0, S_{j+1}) \leq \Delta(\sigma_1 \circ \sigma_2 \circ \ldots \circ \sigma_0, S_{j+1}) + \Delta(\sigma_1 \circ \sigma_2 \circ \ldots \circ \sigma_0, S_{j+1}) \]

\[ \Delta(\sigma_1 \circ \sigma_2 \circ \ldots \circ \sigma_{j+1}, S_{j+1}) \leq \sum_{i=0}^{j+1} \Delta(\sigma_{i+1} \circ \sigma_i, S_i) \]

we'll bound each term.
Recall that since summary $S_i$ is output by an $\alpha$-thresholding alg for threshold $t_{i-1} = \frac{1}{(1+3)^j - j-i}$, we must have output a sol. of cost

$$\Delta(S_{i-1}, \sigma_i, S_i) \leq \alpha \cdot t_{i-1}$$

$$= \alpha \cdot \frac{t}{(1+3)^j} \cdot (j-i+1)$$

So, $\exists \leq \sum_{i=1}^{j+1} \alpha \cdot \frac{t}{(1+3)^j} \cdot (j-i+1)$

$$= \alpha \cdot t \sum_{i=1}^{j+1} \left( \frac{1}{(1+3)^j} \right)^{j-i+1}$$

$$\leq (\alpha + O(3)) t$$  \(\alpha \neq \)
How does \( t \) compare to \( \text{opt} \)?

Recall \( t \) is last threshold that didn't cause any thread to fail.

So some thread must have failed on

\[
\frac{t}{(1+\varepsilon)}
\]

So

\[
\Delta(\sigma_1, \ldots, \sigma_j, \sigma_{j+1}^{\text{opt}}) \geq \frac{t}{(1+\varepsilon)}
\]

Together with

\[
\Delta(\sigma_1, \ldots, \sigma_j, \sigma_{j+1}) \leq (\alpha + O(\varepsilon))(3+1)(1+\varepsilon)
\]

output summary

\( \checkmark \)