COL870: Clustering Algorithms

Ragesh Jaiswal, CSE, IIT Delhi

- Let OPT denote the optimal value for the k-median problem and OPT_{LP} denote the optimal value for the relaxed LP.
- Claim 1: $OPT_{LP} \leq OPT$.
- Let $r_j = \sum_i x_{ij} \cdot D(i,j)$. This may be interpreted as the contribution of the j^{th} point to the cost function.
- Claim 2: $\sum_{j \in S} r_j = OPT_{LP}$
- Let $B(j, r)$ denote the subset of all points that have distance at most r from point j.
- Let $V_j = \{j' \in S | B(j, 2r_j) \cap B(j', 2r'_j) \neq \emptyset\}$

Algorithm

- $T \leftarrow \{\}$
- while $S \neq \{\}$
    - pick the $j \in S$ with smallest r_j
    - $T \leftarrow T \cup \{j\}$
    - $S \leftarrow S \setminus V_j$

- Claim 3: $\Phi(T) \leq 4 \cdot OPT_{LP} \leq 4 \cdot OPT$.
- Claim 4: $|T| \leq 2k$.

Pseudo/bi-criteria-approximation algorithm

What is the use of a pseudo-approximation algorithm for k-median/means?

- Let (X, D) denote any metric space.
- For any $S \subset X$, let $\Psi_k(S, S)$ denote the cost of the optimal k-median solution when the centers are allowed to be chosen from S.
- Claim 1: For any $S, Q \subset X$, $\Psi_k(S, S) \leq 2 \cdot \Psi_k(S, Q)$.
- Let $S \subset X$ and $S_1, ..., S_m$ denote an arbitrary partition of S into m subsets.
- Claim 2: $\sum_i \Psi_k(S_i, S_i) \leq 2 \cdot \Psi_k(S, S)$.

- Let (X, D) denote any metric space.
- For any $S \subset X$, let $\Psi_k(S, S)$ denote the cost of the optimal k-median solution when the centers are allowed to be chosen from S.
- Claim 1: For any $S, Q \subset X$, $\Psi_k(S, S) \leq 2 \cdot \Psi_k(S, Q)$.
- Let $S \subset X$ and $S_1, ..., S_m$ denote an arbitrary partition of S into m subsets.
- Claim 2: $\sum_i \Psi_k(S_i, S_i) \leq 2 \cdot \Psi_k(S, S)$.
- Let $C_i = \{c_{i,1}, c_{i,2}, ..., c_{i,k'}\} \subset S_i$ and let $w_{i,j}$ denote the number of points in S_i for which the closest centre in the set C_i is $c_{i,j}$.
- Let $P = \sum_{i=1}^{m} \Phi(S_i, C_i) = \sum_i \sum_{x \in S_i} \min_{c \in C_i} D(x, c)$.
- Let $c_1^*, c_2^*, ..., c_k^*$ be the optimal centers w.r.t. the discrete k-median problem over S. Let $P^* = \Psi_k(S, S)$.
- Let S' denote a problem instance consisting of the "location" $\cup_i C_i$ and each location $c_{i,j}$ has $w_{i,j}$ points.
- Claim 3: $\Psi_k(S', S') \leq 2 \cdot (P + P^*)$.

- Definition: An (a, b) pseudo-approximation algorithm for the k-median problem outputs at most $a \cdot k$ centers such that the cost of this solution is at most b times the cost of the optimal k-median solution.
- Suppose we have a (a, b) pseudo-approximation algorithm $\mathcal{A}$ and a c-approximation algorithm $\mathcal{B}$. Consider the following approximation algorithm:

An algorithm for k-median

- Input: (S, k)
- Partition S into m equal size sets $S_1, ..., S_m$
- For each $i \in [m]$: Run $\mathcal{A}(S_i, k)$ to obtain centers C_i
- Compute the "weights" $w_{i,j}$ for the centre locations $c_{i,j}$ and consider the instance S'
- Run $\mathcal{B}(S', k)$ and let C be the centers obtained
- Output C

- Definition: An (a, b) pseudo-approximation algorithm for the k-median problem outputs at most $a \cdot k$ centers such that the cost of this solution is at most b times the cost of the optimal k-median solution.
- Suppose we have a (a, b) pseudo-approximation algorithm $\mathcal{A}$ and a c-approximation algorithm $\mathcal{B}$. Consider the following approximation algorithm:

An algorithm for k-median

- Input: (S, k)
- Partition S into m equal size sets $S_1, ..., S_m$
- For each $i \in [m]$: Run $\mathcal{A}(S_i, k)$ to obtain centers C_i
- Compute the "weights" $w_{i,j}$ for the centre locations $c_{i,j}$ and consider the instance S'
- Run $\mathcal{B}(S', k)$ and let C be the centers obtained
- Output C

Theorem

The above algorithm gives an approximation factor of $2c(1 + 2b) + 2b$.

- How do we solve k-median (in metric space) approximately?
  - First Idea: Try writing a Linear Program (LP) relaxation for the discrete version of the problem and round.
  - Second Idea: Try a local search heuristic for the discrete version of the problem.
  - Third Idea: Try simple sampling based approaches.
    - We will analyse an algorithm for the k-means problem in the Euclidean setting which may be very easily generalised for many different settings.

k-means++ *seeding* algorithm

- Pick the first centre c_1 uniformly at random from S
- For $i = 2$ to k
  - Pick a point $x \in S$ to be the centre c_i with probability $\frac{\min_{j \in \{1,...,i-1\}} ||x - c_j||^2}{\sum_{x \in S} \min_{j \in \{1,...,i-1\}} ||x - c_j||^2}$
- Output $T = \{c_1, ..., c_k\}$

- How do we solve k-median (in metric space) approximately?
  - First Idea: Try writing a Linear Program (LP) relaxation for the discrete version of the problem and round.
  - Second Idea: Try a local search heuristic for the discrete version of the problem.
  - Third Idea: Try simple sampling based approaches.
    - We will analyse an algorithm for the k-means problem in the Euclidean setting which may be very easily generalised for many different settings.

k-means++ *seeding* algorithm

- Pick the first centre c_1 uniformly at random from S
- For $i = 2$ to k
  - Pick a point $x \in S$ to be the centre c_i with probability $\frac{\min_{j \in \{1,...,i-1\}} ||x - c_j||^2}{\sum_{x \in S} \min_{j \in \{1,...,i-1\}} ||x - c_j||^2}$
- Output $T = \{c_1, ..., c_k\}$

Theorem (Arthur and Vassilvitskii 2007)

Let $\phi = \Phi(S, T)$ be the random variable denoting the cost of the solution produced by k-means++ and let ϕ_{OPT} denote the cost of the optimal solution. Then $E[\phi] \leq 8 \cdot (\ln k + 2) \cdot \phi_{OPT}$.

k-means++ *seeding* algorithm

- Pick the first centre c_1 uniformly at random from S
- For $i = 2$ to k
    - Pick a point $x \in S$ to be the centre c_i with probability $\frac{\min_{j \in \{1,...,i-1\}} ||x - c_j||^2}{\sum_{x \in S} \min_{j \in \{1,...,i-1\}} ||x - c_j||^2}$
- Output $T = \{c_1, ..., c_k\}$

Theorem (Arthur and Vassilvitskii 2007)

Let $\phi = \Phi(S, T)$ be the random variable denoting the cost of the solution produced by k-means++ and let ϕ_{OPT} denote the cost of the optimal solution. Then $E[\phi] \leq 8 \cdot (\ln k + 2) \cdot \phi_{OPT}$.

- For any set $T \subset S$ of centers and any point $x \in S$, let $D(x, T) = \min_{t \in T} ||x - t||$. We will just use $D(x)$ when T is clear from the context.
- Let C_{OPT} denote the optimal clustering.

Theorem (Arthur and Vassilvitskii 2007)

Let $\phi = \Phi(S, T)$ be the random variable denoting the cost of the solution produced by k-means++ and let ϕ_{OPT} denote the cost of the optimal solution. Then $E[\phi] \leq 8 \cdot (\ln k + 2) \cdot \phi_{OPT}$.

- For any set $T \subset S$ of centers and any point $x \in S$, let $D(x, T) = \min_{t \in T} ||x - t||$. We will just use $D(x)$ when T is clear from the context.
- Let C_{OPT} denote the optimal k-means clustering of S.
- Claim 1: Let A be an arbitrary cluster in C_{OPT}. Let c be a randomly chosen point from A. Then $E[\Phi(A, \{c\})] = 2 \cdot \Phi(A, centroid(A))$.

Theorem (Arthur and Vassilvitskii 2007)

Let $\phi = \Phi(S, T)$ be the random variable denoting the cost of the solution produced by k-means++ and let ϕ_{OPT} denote the cost of the optimal solution. Then $E[\phi] \leq 8 \cdot (\ln k + 2) \cdot \phi_{OPT}$.

- For any set $T \subset S$ of centers and any point $x \in S$, let $D(x, T) = \min_{t \in T} ||x - t||$. We will just use $D(x)$ when T is clear from the context.
- Let C_{OPT} denote the optimal k-means clustering of S.
- Claim 1: Let A be an arbitrary cluster in C_{OPT}. Let c be a randomly chosen point from A. Then $E[\Phi(A, \{c\})] = 2 \cdot \Phi(A, centroid(A))$.
- Claim 2: Let A be an arbitrary cluster in C_{OPT} and let T be an arbitrary set of centers. Let t denote a point chosen from A using D^2 sampling. That is, for any $a \in A$, $\mathbf{Pr}[t = a] = \frac{D(a,T)}{\sum_{x \in A} D(x,T)}$. Then $\mathbf{E}[\Phi(A, T \cup \{t\})] \leq 8 \cdot \Phi(A, centroid(A))$.

End