COL870: Clustering Algorithms

Ragesh Jaiswal, CSE, IIT Delhi

- Claim 1: For any $t \in T$ and $o \in O$, $\Phi(T - \{t\} + \{o\}) - \Phi(T) \geq 0$.
- We will need the following definitions:
  - For any centre $t \in T$, let C_t denote the cluster corresponding to t.
  - For any centre $o \in T$, let C_o denote the cluster corresponding to t.
  - For any point $x \in S$, t_x denotes the closest center in T to x and similarly o_x denotes the closest centre to x in O.
- Claim 2: Let (o, t) denote a swap-pair. Then for any $x \in C_t$, either $o_x = o$ or $t_{o_x} \neq t$.
- Claim 3: For any swap pair (o, t), we have

$$0 \leq \Phi(T - \{t\} + \{o\}) - \Phi(T) \;=\; \sum_{x \in C_o} (d(x, o)^2 - d(x, t_x)^2) + \sum_{x \in C_t \setminus C_o} (d(x, t_{o_x})^2 - d(x, t)^2)$$

- Claim 4: Let $R = \sum_{x \in S} d(x, t_{o_x})$. Then $\Phi(O) - 3\Phi(T) + 2R \geq 0$
- Claim 5: $R \leq 2\Phi(O) + \Phi(T) + 2\sqrt{\Phi(O)}\sqrt{\Phi(T)}$.
- Putting together claims 4 and 5 gives us the result.

- 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.
    - A simple rounding idea gives a "pseudo-approximation" algorithm.
  - Second Idea: Try a local search heuristic for the discrete version of the problem.

- 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.
    - A simple rounding idea gives a "pseudo-approximation" algorithm.
    - A pseudo-approximation algorithm for the k-median problem outputs more than k centers and the approximation factor is computed w.r.t. the optimal solution with k centers.
  - Second Idea: Try a local search heuristic for the discrete version of the problem.

- Recall the Linear Programming relaxation for k-median.

$$\boxed{\begin{array}{l} \text{Minimize } \sum_{i,j} D(i,j) \cdot x_{ij}, \\ \text{subject to :} \\ \sum_i x_{ij} = 1 \quad \text{ for each } j \\ x_{ij} \leq y_i \quad \text{ for each } i,j \\ \sum_i y_i \leq k \\ 0 \leq x_{ij} \leq 1 \quad \text{ for each } i,j \\ 0 \leq y_i \leq 1 \quad \text{ for each } i \end{array}}$$

- Let OPT denote the optimal value for the k-median problem and OPT_{LP} denote the optimal value for the relaxed LP.
- <u>Claim 1</u>: $OPT_{LP} \leq OPT$.

- Recall the Linear Programming relaxation for k-median.

$$\begin{array}{|l|}\hline \text{Minimize } \sum_{i,j} D(i,j) \cdot x_{ij}, \\ \text{subject to :} \\ \sum_i x_{ij} = 1 \quad \text{for each } j \\ x_{ij} \leq y_i \quad \text{for each } i,j \\ \sum_i y_i \leq k \\ 0 \leq x_{ij} \leq 1 \quad \text{for each } i,j \\ 0 \leq y_i \leq 1 \quad \text{for each } i \\ \hline \end{array}$$

- 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 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
    - $T \leftarrow T \cup \{j\}$
    - $S \leftarrow S \setminus V_j$

- 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: $\Phi(T) \leq 4 \cdot OPT_{LP} \leq 4 \cdot OPT$.

- 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: $\sum_i \Psi_k(S_i, S_i) \leq 2 \cdot \Psi_k(S, S)$.

End