COL870: Clustering Algorithms

Ragesh Jaiswal, CSE, IIT Delhi

- How do we solve k-median (in metric space) approximately?
  - First Idea: Try writing a Linear Program (LP) for the discrete version of the problem.
  - Second Idea: Try a local search heuristic.
    1. Start with k centers $T \subset S$ chosen arbitrarily.
    2. At every step, replace a centre in T with a point in S given that the cost decreases due to this "swap".
    - We will argue that when the above local search algorithm terminates, we obtain a constant factor approximation.
    - For ease of discussion, we will discuss this heuristic for the k-means problem in Euclidean space and skip the running time analysis.

- How do we solve k-median (in metric space) approximately?
  - First Idea: Try writing a Linear Program (LP) for the discrete version of the problem.
  - Second Idea: Try a local search heuristic.

Local search for k-means

- Initialize centers $T \subset S$ arbitrarily
- While $\exists t \in T$, $t' \in S$, such that $\Phi(T + \{t'\} - \{t\}) < \Phi(T)$
  - $T \leftarrow T + \{t'\} - \{t\}$

Local search for k-means

- Initialize centers $T \subset S$ arbitrarily
- While $\exists t \in T$, $t' \in S$, such that $\Phi(T + \{t'\} - \{t\}) < \Phi(T)$
  - $T \leftarrow T + \{t'\} - \{t\}$

Lemma

Let O be the subset of k data points that minimise $\Phi(O)$ for the discrete k-means problem.Let T be the solution returned by the above local search procedure. Then $\Phi(T) \leq 25 \cdot \Phi(O)$.

- This gives an approximation factor of 50 for the k-means problem.
- Project Topic: There is a much better analysis of this local search than what we see here. This procedure is also efficient.

Local search for k-means

- Initialize centers $T \subset S$ arbitrarily
- While $\exists t \in T$, $t' \in S$, such that $\Phi(T + \{t'\} - \{t\}) < \Phi(T)$
    - $T \leftarrow T + \{t'\} - \{t\}$

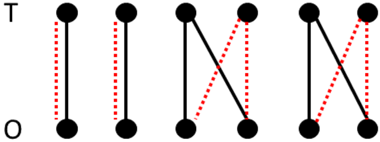
Lemma

Let O be the subset of k data points that minimise $\Phi(O)$ for the discrete k-means problem.Let T be the solution returned by the above local search procedure. Then $\Phi(T) \leq 25 \cdot \Phi(O)$.

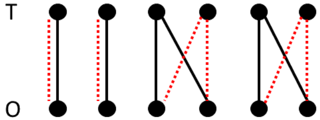
- Claim 1: For any $t \in T$ and $o \in O$, $\Phi(T - \{t\} + \{o\}) - \Phi(T) \geq 0$.

- 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 $o \in O$, let C_o denote the cluster corresponding to o.
    - 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 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 $o \in O$, let C_o denote the cluster corresponding to o.
  - 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.
- Consider a bipartite graph with nodes corresponding to T and O on either side. There is an edge from a node $o \in O$ to a node $t \in T$ iff t is the nearest point in T to o.

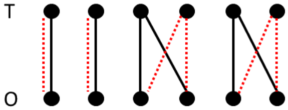


- Claim 1: For any $t \in T$ and $o \in O$, $\Phi(T - \{t\} + \{o\}) - \Phi(T) \geq 0$.
- We will need the following definitions:
  - For any centre $t \in T$, let C_t denote the cluster corresponding to t.
  - For any centre $o \in O$, let C_o denote the cluster corresponding to o.
  - For any point $x \in S$, t_x denotes the closest center in T to x and similarly o_x denotes the closest centre to x in O.
- Consider a bipartite graph with nodes corresponding to T and O on either side. There is an edge from a node $o \in O$ to a node $t \in T$ iff t is the nearest point in T to o.



- We will construct the following k "swap-pairs" from the above bipartite graph:
  - All vertices in T with degree 1 are taken in the swap pair (along with their neighbour in O).
  - All remaining vertices in O are paired with a 0 degree vertex such that each 0-degree vertex is present in at most two of the swap pairs.

- 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 $o \in O$, let C_o denote the cluster corresponding to o.
  - 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.
- Consider a bipartite graph with nodes corresponding to T and O on either side. There is an edge from a node $o \in O$ to a node $t \in T$ iff t is the nearest point in T to o.



- We will construct the following k "swap-pairs" from the above bipartite graph:
  - All vertices in T with degree 1 are taken in the swap pair (along with their neighbour in O).
  - All remaining vertices in O are paired with a 0 degree vertex such that each 0-degree vertex is present in at most two of the swap pairs.
- 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 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 $o \in O$, let C_o denote the cluster corresponding to o.
    - 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

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

- Claim 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 $o \in O$, let C_o denote the cluster corresponding to o.
  - 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

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

- Claim 4: Let $R = \sum_{x \in S} d(x, t_{o_x})$. Then $\Phi(O) - 3\Phi(T) + 2R \geq 0$

- Claim 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 $o \in O$, let C_o denote the cluster corresponding to o.
  - 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

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

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

- 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 $o \in O$, let C_o denote the cluster corresponding to o.
  - 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

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

- 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.

End