COL870: Clustering Algorithms

Ragesh Jaiswal, CSE, IIT Delhi

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 $\mathbf{E}[\Phi(A, T \cup \{t\})] \leq 8 \cdot \Phi(A, centroid(A))$.

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 $\mathbf{E}[\Phi(A, T \cup \{t\})] \leq 8 \cdot \Phi(A, centroid(A))$.
- For any set of centers $T \subset S$, a cluster $A \in C_{OPT}$ is said to be "uncovered" if $A \cap T = \emptyset$. A cluster that is not uncovered is called covered.
- For any subset $X = \{S_{i_1}, S_{i_2}, ..., S_{i_l}\}$ of optimal clusters $\Phi_{OPT}(X)$ denotes the cost of these clusters in the optimal clustering. That is $\Phi_{OPT}(X) = \sum_{j \in [l]} \Phi(S_{i_j}, centroid(S_{i_j}))$.

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 $\mathbf{E}[\Phi(A, T \cup \{t\})] \leq 8 \cdot \Phi(A, centroid(A))$.
- For any set of centers $T \subset S$, a cluster $A \in C_{OPT}$ is said to be "uncovered" if $A \cap T = \emptyset$. A cluster that is not uncovered is called covered.
- For any subset $X = \{S_{i_1}, S_{i_2}, ..., S_{i_l}\}$ of optimal clusters $\Phi_{OPT}(X)$ denotes the cost of these clusters in the optimal clustering. That is $\Phi_{OPT}(X) = \sum_{j \in [l]} \Phi(S_{i_j}, centroid(S_{i_j}))$.
- Claim 3: Let T be any arbitrary set of centers. Let u be the number of clusters in C_{OPT} that are uncovered (w.r.t. T). Let S_u denote the set of points in these uncovered clusters and $S_c = S \setminus S_u$. Suppose we now add $t \leq u$ centers to T chosen with D^2 sampling. Let ϕ denote the cost w.r.t. T and ϕ' denote the cost w.r.t. T plus the newly added centers. Then $\mathbf{E}[\phi'] \leq (\phi(S_c) + 8 \cdot \phi_{OPT}(S_u)) \cdot (1 + H_t) + \frac{u-t}{u} \cdot \phi(S_u)$. Here H_t denotes the Harmonic sum $H_t = 1 + 1/2 + 1/3 + ... + 1/t$.

End