Constant factor approximation for k-means

$D(\cdot,\cdot)$ squared Euclidean distance

$D(\cdot, \cdot)$ squared Euclidean distance

Goal

A polynomial time algorithm $\textsc{Alg}$ for which there is a constant $\gamma \geq 1$ such that for every $P \subset \mathbb{R}^d, |P| < \infty$, and every $k \in \mathbb{N}$, algorithm $\textsc{Alg}$ on input (P, k) outputs a set C of size k satisfying

$$D(P, C) \leq \gamma \cdot \text{opt}_k(P).$$

Constant factor approximation for k-means

$P \subset \mathbb{R}^d, D(\cdot,\cdot)$ squared Euclidean distance, $T \subset \mathbb{R}^d, x \in T$

Constant factor approximation for k-means

$P \subset \mathbb{R}^d, D(\cdot,\cdot)$ squared Euclidean distance, $T \subset \mathbb{R}^d, x \in T$

- $N_T(x) := \{q \in P \mid \text{for all } y \in T\colon\ D(q,x) \leq D(q,y)\}$

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Constant factor approximation for k-means

$P \subset \mathbb{R}^d, D(\cdot,\cdot)$ squared Euclidean distance, $T \subset \mathbb{R}^d, x \in T$

- $N_T(x) := \{q \in P \mid \text{for all } y \in T\colon\ D(q,x) \leq D(q,y)\}$
- $q \in P : t_q :=$ closest point in T to q

Constant factor approximation for k-means

$P \subset \mathbb{R}^d, D(\cdot,\cdot)$ squared Euclidean distance, $T \subset \mathbb{R}^d, x \in T$

- $N_T(x) := \{q \in P \mid \text{for all } y \in T\colon\ D(q,x) \leq D(q,y)\}$
- $q \in P : t_q :=$ closest point in T to q
- $K \subseteq \mathbb{R}^d$ called (c,k)-approximate candidate set, if there is a set $S \subset K, |S| = k$, with $D(P,S) \leq c \cdot \text{opt}_k(P)$, i.e. the best k-centroid set S in K is at most c times worse than the optimal set of centroids.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Constant factor approximation for k-means

Lemma 5.1

For all finite sets $P \subset \mathbb{R}^d$, and all $k \in N$, the set P is a $(2, k)$-approximate candidate set for itself.

Lemma 5.1

For all finite sets $P \subset \mathbb{R}^d$, and all $k \in N$, the set P is a $(2, k)$-approximate candidate set for itself.

Observation

If K is a $(2, k)$-approximate candidate set for P and if $D(P, S) \leq c \cdot \min_{T \subset K, |T|=k} D(P, T)$, then $D(P, S) \leq 2c \cdot opt_k(P)$.

Stable sets

$O := \text{argmin}_{S \subset P, |S|=k} D(P, S)$, i.e. optimal set of centroids in P.

$O := \text{argmin}_{S\subset P, |S|=k} D(P,S)$, i.e. optimal set of centroids in P.

Definition 5.2

Let $S \subset P$.

1. *S is called stable, if for all $s \in S, s' \in P \setminus S$*

$$D(P, S - \{s\} \cup \{s'\}) \geq D(P,S).$$

2. *S is called ϵ-stable, if for all $s \in S, s' \in P \setminus S$*

$$D(P, S - \{s\} \cup \{s'\}) \geq (1-\epsilon) D(P,S).$$

Observation

If S is stable, then for all $s \in S, o \in O$

$$D(P, S - \{s\} \cup \{o\}) \geq D(P, S).$$

Local improvement for k-means

k-means-$\text{LI}(P)$

choose a set $S \subset P$ of k initial centroids;

repeat

find $s \in S, s' \in P \setminus S$ with

$D(P, S - \{s\} \cup \{s'\}) < \quad D(P, S)$;

set $S := S - \{s\} \cup \{s'\}$;

until S *is stable*;

Local improvement for k-means

k-means-LI(P)

choose a set $S \subset P$ of k initial centroids;

repeat

find $s \in S, s' \in P \setminus S$ with
$D(P, S - \{s\} \cup \{s'\}) < (1-\epsilon)D(P,S)$;
set $S := S - \{s\} \cup \{s'\}$;

until S *is* ϵ*-stable*;

Local improvement for k-means

Theorem 5.3

- *If S is a stable set, then*

$$D(P,S) \leq 81 \cdot D(P,O).$$

- *If S is a ϵ-stable set, then*

$$D(P,S) \leq \left(\frac{9}{1-\epsilon}\right)^2 \cdot D(P,O).$$

Corollary 5.4

For any $\epsilon > 0$, the k-means problem can be approximated with factor $162 + \epsilon$ in time polynomial in the input size and in $1/\epsilon$.

Capturing points

$O \subset P, |O| = k$ optimal set of centroids in P, S stable set, $|S| = k$, called set of heuristic centroids.

Capturing points

$O \subset P, |O| = k$ optimal set of centroids in P, S stable set, $|S| = k$, called set of heuristic centroids.

- If $s \in S$ is closest point in S to $o \in O$, then s captures o, o is captured by s, and we write $s = s_o$.
- If $s \in S$ captures no element of O, then s is called lonely.

Capturing points

$O \subset P, |O| = k$ optimal set of centroids in P, S stable set, $|S| = k$, called set of heuristic centroids.

- If $s \in S$ is closest point in S to $o \in O$, then s captures o, o is captured by s, and we write $s = s_o$.
- If $s \in S$ captures no element of O, then s is called lonely.

Partitioning centroids

Partition S into $S_1, \ldots, S_m$ and O into $O_1, \ldots, O_m$ such that

- $|S_i| = |O_i|, i = 1, \ldots, m$
- if $s \in S_i$, then either s is lonely or s captures all $o \in O_i$.

Swap pairs

Partitioning centroids

Partition S into $S_1, \ldots, S_m$ and O into $O_1, \ldots, O_m$ such that

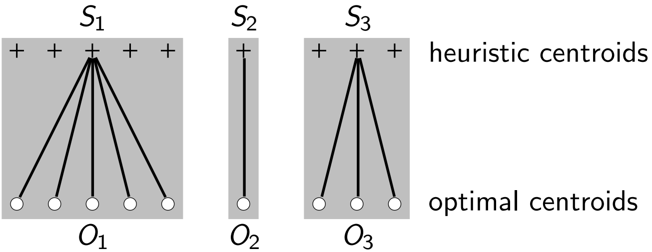
- $|S_i| = |O_i|, i = 1, \ldots, m$
- if $s \in S_i$, then either s is lonely or s captures all $o \in O_i$.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Partitioning centroids

Partition S into $S_1, \ldots, S_m$ and O into $O_1, \ldots, O_m$ such that

- $|S_i| = |O_i|, i = 1, \ldots, m$
- if $s \in S_i$, then either s is lonely or s captures all $o \in O_i$.



Partitioning centroids

Partition S into $S_1, \dots, S_m$ and O into $O_1, \dots, O_m$ such that

- $|S_i| = |O_i|, i = 1, \dots, m$
- if $s \in S_i$, then either s is lonely or s captures all $o \in O_i$.

Swap pairs

$(s_1, o_1), \dots, (s_k, o_k)$ are called swap pairs, if

- $\forall j : (s_j, o_j) \in \bigcup S_i \times O_i$
- each $o \in O$ is contained in exactly one pair,
- each s is contained in at most two pairs,
- for each pair (s_j, o_j) the element s_j captures no $o' \neq o_j$.

Swap pairs

Partitioning centroids

Partition S into $S_1, \ldots, S_m$ and O into $O_1, \ldots, O_m$ such that

- $|S_i| = |O_i|, i = 1, \ldots, m$
- if $s \in S_i$, then either s is lonely or s captures all $o \in O_i$.

Observation

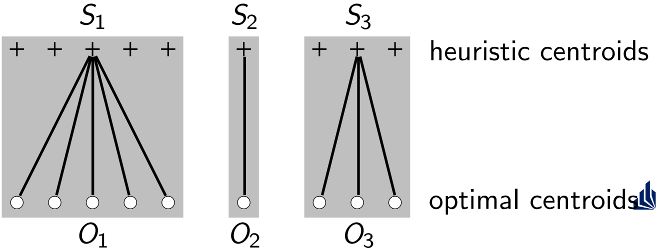
For each partitioning of centroids $S_1, \ldots, S_m$ and $O_1, \ldots, O_m$ there is a set of swap pairs.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Swap pairs

$(s_1, o_1), \ldots, (s_k, o_k)$ are called swap pairs, if

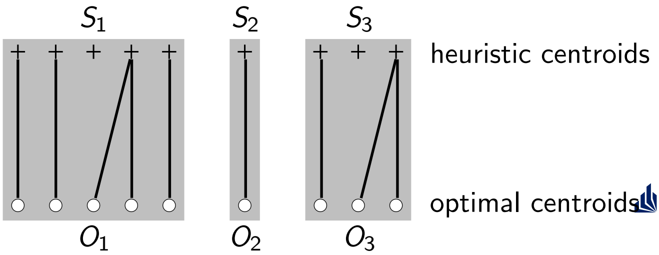
- $\forall j : (s_j, o_j) \in \bigcup S_i \times O_i$
- each $o \in O$ is contained in exactly one pair,
- each s is contained in at most two pairs,
- for each pair (s_j, o_j) the element s_j captures no $o' \neq o_j$.



Swap pairs

$(s_1, o_1), \ldots, (s_k, o_k)$ are called swap pairs, if

- $\forall j : (s_j, o_j) \in \bigcup S_i \times O_i$
- each $o \in O$ is contained in exactly one pair,
- each s is contained in at most two pairs,
- for each pair (s_j, o_j) the element s_j captures no $o' \neq o_j$.



Reassignments

Let (s, o) be a swap pair in set $\{(s_1, o_1), \ldots, (s_k, o_k)\}$ and let $C_1, \ldots, C_k$ be the clusters for set $S = \{s_1, \ldots, s_k\}$.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Reassignments

Let (s, o) be a swap pair in set $\{(s_1, o_1), \ldots, (s_k, o_k)\}$ and let $C_1, \ldots, C_k$ be the clusters for set $S = \{s_1, \ldots, s_k\}$.

Reassigning points

For $S' = S - \{s\} \cup \{o\}$ we define a new clustering of P as follows

- if $q \notin N_S(s) \cup N_O(o)$, then o stays in its old cluster,
- if $q \in N_O(o)$, then q is assigned to o's cluster,
- if $q \in N_S(s) \setminus N_O(o)$ then q is assigned to the cluster belonging to s_{o_q}.

Let (s, o) be a swap pair in set $\{(s_1, o_1), \ldots, (s_k, o_k)\}$ and let $C_1, \ldots, C_k$ be the clusters for set $S = \{s_1, \ldots, s_k\}$.

Reassigning points

For $S' = S - \{s\} \cup \{o\}$ we define a new clustering of P as follows

- if $q \notin N_S(s) \cup N_O(o)$, then o stays in its old cluster,
- if $q \in N_O(o)$, then q is assigned to o's cluster,
- if $q \in N_S(s) \setminus N_O(o)$ then q is assigned to the cluster belonging to s_{o_q}.

Observation

$$0 \leq \sum_{q \in N_O(o)} D(q, o) - D(q, s_q) + \sum_{q \in N_S(s) \setminus N_O(o)} D(q, s_{o_q}) - D(q, s).$$

Lemma 5.5

Let S be a stable set. Then

$$0 \leq D(P, O) - 3D(P, S) + 2R,$$

where $R := \sum_{q \in P} D(q, s_{o_q})$.

Lemma 5.5

Let S be a stable set. Then

$$0 \leq D(P,O) - 3D(P,S) + 2R,$$

where $R := \sum_{q \in P} D(q, s_{o_q})$.

Lemma 5.6

$$R \leq 4D(P,O) + (1 + 4/\alpha)D(P,S),$$

where

$$\alpha^2 := \frac{D(P,S)}{D(P,O)}.$$

Local improvement for k-means - technical lemmas

Lemma 5.7

Let $\beta_n, \dots, \beta_n$ *and* $\gamma_1, \dots, \gamma_n$ *be two sequences of real numbers and set*

$$\alpha^2 := \frac{\sum_{i=1}^n \gamma_i^2}{\sum_{i=1}^n \beta_i^2}.$$

Then

$$\sum_{i=1}^n \gamma_i \beta_i \le \frac{1}{\alpha} \sum_{i=1}^n \gamma_i^2.$$