Algorithms Beyond the Worst Case
Lectures 3 & 4: k-Means Method

Bodo Manthey

UNIVERSITY OF TWENTE.

February 23 & March 1, 2016
\textit{k}-Means Clustering

- NP-hard (even for $d = 2$ or $k = 2$)
- $(1 + \varepsilon)$ approximations
- In practice: $k$-means method
$k$-Means Clustering

- NP-hard (even for $d = 2$ or $k = 2$)
- $(1 + \varepsilon)$-approximations
- in practice: $k$-means method
**k-Means Clustering**

**instance**  
$n$ points $X \subseteq \mathbb{R}^d$

**output**  
clustering $C_1, \ldots, C_k$

centers $c_1, \ldots, c_k$

**objective** minimize  
$$\sum_i \sum_{x \in C_i} \|x - c_i\|^2$$

**how to choose centers:**  
$$c_i = \frac{1}{|C_i|} \cdot \sum_{x \in C_i} x$$

- NP-hard (even for $d = 2$ or $k = 2$)
- $(1 + \varepsilon)$ approximations
- in practice: $k$-means method
Clusters vs. Centers

- centers imply clusters: every point is assigned to the center closest to it (there might some be ambiguity)

- clusters imply centers: centers should be the centers of mass according to the following lemma

**Lemma**

\[
X \subseteq \mathbb{R}^d: \text{finite set of points,} \\
c = \text{cm}(X) = \frac{1}{|X|} \cdot \sum_{x \in X} x: \text{center of mass of } X, \\
y \in \mathbb{R}^d \text{ arbitrary;} \\
\text{then } \sum_{x \in X} \|x - y\|^2 = \sum_{x \in X} \|x - c\|^2 + |X| \cdot \|c - y\|^2
\]
**k-Means Method**

**input:** \( X \subseteq \mathbb{R}^d \)

1. choose \( c_1, \ldots, c_k \in \mathbb{R}^d \)
2. repeat
3. partition \( X \)
4. adjust centers
5. until clustering is stable

**output:** \( C_1, \ldots, C_k \)

- worst-case lower bound: \( 2\Omega(n) \) iterations
- worst-case upper bound: \( n^{3kd} \)
- in practice the number of iterations is generally much less than the number of points (Duda et al.)
**k-Means Method**

**input:** \( X \subseteq \mathbb{R}^d \)

1. choose \( c_1, \ldots, c_k \in \mathbb{R}^d \)

2. repeat

3. partition \( X \)

4. adjust centers

5. until clustering is stable

**output:** \( C_1, \ldots, C_k \)

- worst-case lower bound: \( 2 \Omega(n) \) iterations
- worst-case upper bound: \( n^{3kd} \)
- in practice the number of iterations is generally much less than the number of points (Duda et al.)
**k-Means Method**

**input:** $X \subseteq \mathbb{R}^d$

1. choose $c_1, \ldots, c_k \in \mathbb{R}^d$
2. repeat
3. partition $X$
4. adjust centers
5. until clustering is stable

**output:** $C_1, \ldots, C_k$
**k-Means Method**

**input:** $X \subseteq \mathbb{R}^d$

1. Choose $c_1, \ldots, c_k \in \mathbb{R}^d$
2. Repeat
3. Partition $X$
4. Adjust centers
5. Until clustering is stable

**output:** $C_1, \ldots, C_k$

- **worst-case lower bound:** $2^\Omega(n)$ iterations
- **worst-case upper bound:** $n^3$ iterations

In practice, the number of iterations is generally much less than the number of points (Duda et al.)
**k-Means Method**

**Input:** $\mathbf{X} \subseteq \mathbb{R}^d$

1. Choose $c_1, \ldots, c_k \in \mathbb{R}^d$
2. Repeat
3. Partition $\mathbf{X}$
4. Adjust centers
5. Until clustering is stable

**Output:** $C_1, \ldots, C_k$

- Worst-case lower bound: $\Omega(n)$ iterations
- Worst-case upper bound: $n^3 kd$

In practice, the number of iterations is generally much less than the number of points (Duda et al.)
**k-Means Method**

**input:** \( X \subseteq \mathbb{R}^d \)

1. choose \( c_1, \ldots, c_k \in \mathbb{R}^d \)
2. repeat
3. partition \( X \)
4. adjust centers
5. until clustering is stable

**output:** \( C_1, \ldots, C_k \)

- worst-case lower bound: \( 2 \Omega(n) \) iterations
- worst-case upper bound: \( n^{3d} \)
- in practice, the number of iterations is generally much less than the number of points (Duda et al.)
**k-Means Method**

**input:** \( X \subseteq \mathbb{R}^d \)

1. choose \( c_1, \ldots, c_k \in \mathbb{R}^d \)
2. repeat
3. partition \( X \)
4. adjust centers
5. until clustering is stable

**output:** \( C_1, \ldots, C_k \)

- worst-case lower bound: \( 2 \Omega(n) \) iterations
- worst-case upper bound: \( n^3 kd \)
- in practice the number of iterations is generally much less than the number of points (Duda et al.)
**k-Means Method**

**input:** $X \subseteq \mathbb{R}^d$

1. choose $c_1, \ldots, c_k \in \mathbb{R}^d$
2. repeat
3. partition $X$
4. adjust centers
5. until clustering is stable

**output:** $C_1, \ldots, C_k$

- worst-case lower bound: $2\Omega(n)$ iterations
- worst-case upper bound: $n^{3kd}$
- in practice the number of iterations is generally much less than the number of points (Duda et al.)
**k-Means Method**

**input:** \( X \subseteq \mathbb{R}^d \)

1. choose \( c_1, \ldots, c_k \in \mathbb{R}^d \)
2. repeat
3. partition \( X \)
4. adjust centers
5. until clustering is stable

**output:** \( C_1, \ldots, C_k \)

- worst-case lower bound: \( 2\Omega(n) \) iterations
- worst-case upper bound: \( n^{3kd} \)

In practice, the number of iterations is generally much less than the number of points (Duda et al.)
**k-Means Method**

**input:** $X \subseteq \mathbb{R}^d$

1. choose $c_1, \ldots, c_k \in \mathbb{R}^d$
2. repeat
3. partition $X$
4. adjust centers
5. until clustering is stable

**output:** $C_1, \ldots, C_k$

- worst-case lower bound: $2\Omega(n)$ iterations
- worst-case upper bound: $n^3 kd$
- in practice the number of iterations is generally much less than the number of points (Duda et al.)
**k-_means Method**

**Input:** \( X \subseteq \mathbb{R}^d \)

1. choose \( c_1, \ldots, c_k \in \mathbb{R}^d \)
2. repeat
3. partition \( X \)
4. adjust centers
5. until clustering is stable

**Output:** \( C_1, \ldots, C_k \)

- worst-case lower bound: \( 2^{\Omega(n)} \) iterations
- worst-case upper bound: \( n^{3kd} \)
- in practice the number of iterations is generally much less than the number of points

(Duda et al.)
we use the objective function as potential (as for 2-opt and TSP)

Lemma

Let $c_i, c_j$ be two cluster centers, let $x$ be a point closer to $c_j$ than to $c_i$, let $\epsilon$ be the distance of $x$ to the hyperplane $H$ bisecting $c_i$ and $c_j$, and let $\delta = \|c_i - c_j\|$. Then reassigning $x$ from $C_i$ to $C_j$ decreases the potential by $2\delta \epsilon$.

Lemma

Readjusting the cluster center of $C_i$ from some point $z$ to the center of mass of $C_i$ decreases the potential by $|C_i| \cdot \|z - \text{cm}(C_i)\|^2 \geq \|z - \text{cm}(C_i)\|^2$.

Lemma

With a probability of at least $1 - n^{-3kd}$, the potential is bounded by $O(nd^2k \log n)$ after the first iteration of $k$-means.
Potential function argument – ideas

- **initial potential:** $O(nd^2k \log n)$
- **many points change clusters:**
  decrease by reassigning
- **few points change clusters:**
  center moves
- **problem:** $n^{O(kd)}$ possible clusterings after perturbation
- **even worse:** $k^n$ possible clusterings before perturbation

reassigning points: $-2\epsilon \delta$

moving centers: $-|C| \cdot \epsilon^2$
Partitioning of iterations

**dense iterations**
- one cluster loses or gains at least \(2kd\) points

**sparse iterations**
- each cluster exchanges \(\leq 2kd\) points
- after 4 (or: \(2^k\)) iterations: 3rd configuration

---

**Theorem (Arthur, Vassilvitskii)**

smoothed \#iterations of \(k\)-means: \(\text{poly}(n^k, 1/\sigma)\)
Discussion – smoothed analysis of $k$-means

- bound can be improved to (roughly) $O\left(\frac{n^{34}}{\sigma^6}\right)$
  (read: “polynomial in $n$ and $1/\sigma$”)

- improvements seem difficult

- good smoothed approximation ratio is unlikely (see exercises):
  - $k$-means method shows poor approximation ratio in practice
  - $k$-means method only works well because it can be restarted with different initializations very often
  - good smoothed approximation ratio would be a case against smoothed analysis