Coresets for Probabilistic Clustering

Melanie Schmidt

Joint work with
Christiane Lammersen, Christian Sohler

15.12.2011
Clustering Certain Data: Metric $k$-median
Clustering Certain Data: Metric $k$-median

Given
Clustering Certain Data: Metric $k$-median

Given

- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$
Clustering Certain Data: Metric $k$-median

Given

- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$
- set of center candidates $C \subseteq X$
Clustering Certain Data: Metric $k$-median

Given
- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$
- set of center candidates $C \subseteq X$

Wanted
- A set $C := \{c_1, \ldots, c_k\} \subseteq C$ and
Clustering Certain Data: Metric $k$-median

**Given**
- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$
- set of center candidates $C \subseteq X$

**Wanted**
- A set $C := \{c_1, \ldots, c_k\} \subseteq C$ and
- an assignment $\rho : P \rightarrow C$ minimizing

$$\text{cost}(P, C, \rho) := \sum_{i=1}^{n} D(p_i, \rho(p_i)).$$
Clustering Certain Data: Metric $k$-median

**Given**
- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$
- set of center candidates $C \subseteq X$

**Wanted**
- A set $C := \{c_1, \ldots, c_k\} \subseteq C$ and
- an assignment $\rho : P \rightarrow C$ minimizing

$$\text{cost}(P, C) := \sum_{i=1}^{n} \min_{c \in C} D(p_i, c).$$
Metric Assigned Probabilistic $k$-Median Clustering

Clustering Certain Data
Clustering Certain Data

General Metrics

Usually $\mathcal{C} = P$.

- No $(1 + \varepsilon)$-approximation for $\varepsilon < 0.73$ (Jain et al. 2002)
- First constant factor approximation by Charikar et al. (1999)
- Approximation guarantee consecutively improved, 3-approximation by Arya et al. (2001)
Clustering Certain Data

General Metrics

Usually $\mathcal{C} = P$.
- no $(1 + \varepsilon)$-approximation for $\varepsilon < 0.73$ (Jain et al. 2002)
- first constant factor approximation by Charikar et al. (1999)
- approximation guarantee consecutively improved,
  3-approximation by Arya et al. (2001)

Euclidean Metric

Usually $X = \mathbb{R}^d$, $M = || \cdot ||$ and $\mathcal{C} = \mathbb{R}^d$.
- First $(1 + \varepsilon)$-approximation by Arora et al. (1998)
- Several improvements reducing the running time
- Chen: $(1 + \varepsilon)$-approximation, pol. in the dimension (2006)
Clustering Certain Data

Coresets

Given a set of points $P$, a weighted subset $S \subseteq P$ is a $(k, \varepsilon)$-coreset if for all sets $C \subseteq \mathcal{C}$ of $k$ centers it holds

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \text{cost}(P, C)$$

where $\text{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p) D(p, c)$.
Clustering Certain Data

Coresets

Given a set of points $P$, a weighted subset $S \subset P$ is a $(k, \varepsilon)$-coreset if for all sets $C \subset \mathcal{C}$ of $k$ centers it holds

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \text{cost}(P, C)$$

where $\text{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p)D(p, c)$. 

Coresets for Probabilistic Clustering
Clustering Certain Data

Coresets

Given a set of points $P$, a weighted subset $S \subset P$ is a $(k, \varepsilon)$-coreset if for all sets $C \subset C$ of $k$ centers it holds

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \text{cost}(P, C)$$

where

$$\text{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p)D(p, c).$$
Clustering Certain Data

Coresets

Given a set of points $P$, a weighted subset $S \subset P$ is a $(k, \varepsilon)$-coreset if for all sets $C \subset C$ of $k$ centers it holds

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \text{cost}(P, C)$$

where $\text{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p)D(p, c)$.
Clustering Certain Data

Coresets

Given a set of points $P$, a weighted subset $S \subseteq P$ is a $(k, \varepsilon)$-coreset if for all sets $C \subseteq \mathcal{C}$ of $k$ centers it holds

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \text{cost}(P, C)$$

where $\text{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p)D(p, c)$. 
Clustering Certain Data

Coresets

Given a set of points $P$, a weighted subset $S \subset P$ is a $(k, \varepsilon)$-coreset if for all sets $C \subset C$ of $k$ centers it holds

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \text{cost}(P, C)$$

where $\text{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p)D(p, c)$. 
Clustering Certain Data
Clustering Certain Data

Coreset constructions

’02: Bădoiu, Har-Peled and Indyk:
First coreset construction for clustering problems

’04: Agarwal, Har-Peled and Varadarajan:
Definition of coresets as used nowadays

’04: Har-Peled and Mazumdar, Coreset of size $O(k\varepsilon^{-d} \log n)$, maintainable in data streams

’05: Har-Peled, Kushal: Coreset of size $O(k^2\varepsilon^{-d})$

’05: Frahling and Sohler: Coreset of size $O(k\varepsilon^{-d} \log n)$, insertion-deletion data streams

’06: Chen: Coresets for metric and Euclidean $k$-median and $k$-means, polynomial in $d,n$ and $\varepsilon^{-1}$

’07: Feldman, Monemizadeh, Sohler: weak coresets, poly$(k, \varepsilon^{-1})$
Clustering Uncertain Data
Clustering Uncertain Data

**Uncertain Data**

Input consisting of (discrete) distributions instead of points
Clustering Uncertain Data

Uncertain Data
Input consisting of (discrete) distributions instead of points
Clustering Uncertain Data

Uncertain Data

Input consisting of (discrete) distributions instead of points

Sources of Uncertain Data

- measurements of sensor networks
- linkage across multiple databases
Metric Assigned Probabilistic \( k \)-Median Clustering

Given finite set \( \mathcal{X} = \{x_1, \ldots, x_m\} \) from metric space \( \mathcal{M} = (\mathcal{X}, D) \), set of nodes \( V = \{v_1, \ldots, v_n\} \) probability distribution \( D_i \) for each node \( v_i \), given by realization probabilities \( p_{ij} \) for all \( j \in [m] \), \( \sum_j p_{ij} = 1 \), \( p_{ij} \leq 1 \), set of possible center locations \( \mathcal{C} \subset \mathcal{X} \).
Given
Given

- finite set \( X := \{ x_1, \ldots, x_m \} \) from metric space \( M = (X, D) \),
Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
Given

- finite set \( X := \{x_1, \ldots, x_m\} \) from metric space \( M = (X, D) \),
- set of nodes \( V : \{v_1, \ldots, v_n\} \)
- probability distribution \( D_i \) for each node \( v_i \), given by realization probabilities \( p_{ij} \) for all \( j \in [m] \), \( \sum_{j=1}^{m} p_{ij} \leq 1 \),
Given

- finite set $X := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $D_i$ for each node $v_i$, given by realization probabilities $p_{ij}$ for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $C \subset X$. 
Metric Assigned Probabilistic $k$-Median Clustering

Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $D_i$ for each node $v_i$, given by realization probabilities $p_{ij}$ for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $C \subset X$. 
Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node $v_i$, given by realization probabilities $p_{ij}$ for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $\mathcal{C} \subset X$. 

Coresets for Probabilistic Clustering
Given

- finite set \( \mathcal{X} := \{x_1, \ldots, x_m\} \) from metric space \( M = (X, D) \),
- set of nodes \( V : \{v_1, \ldots, v_n\} \)
- probability distribution \( D_i \) for each node \( v_i \), given by realization probabilities \( p_{ij} \) for all \( j \in [m] \), \( \sum_{j=1}^{m} p_{ij} \leq 1 \),
- set of possible center locations \( C \subset X \).

Wanted

- A set \( C := \{c_1, \ldots, c_k\} \subseteq C \) and
- an assignment \( \rho : V \rightarrow C \) minimizing

\[
E_D [\text{cost}(V, C, \rho)] := \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \cdot D(x_j, \rho(v_i)).
\]
Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $D_i$ for each node $v_i$, given by realization probabilities $p_{ij}$ for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $C \subset X$.

Wanted

- A set $C := \{c_1, \ldots, c_k\} \subseteq C$ and
- an assignment $\rho : V \rightarrow C$ minimizing

$$E_D [\text{cost}(V, C)] := \min_{\rho : V \rightarrow C} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \cdot D(x_j, \rho(v_i)).$$
Related work

Cormode, McGregor (PODS 2008)
- $(1 + \varepsilon)$-approximation for a variant of the above problem
- $(1 + \varepsilon)$-approximation for uncertain $k$-means
- Constant approximation for (assigned) metric $k$-median
- Bicriteria approximations for uncertain metric $k$-center

Guha and Munagala (PODS 2009)
- Constant approximation for uncertain metric $k$-center
Probabilistic Coresets

What should a probabilistic coreset look like?
Probabilistic Coresets

What should a probabilistic coreset look like?

- Consists of probabilistic points (nodes)
Probabilistic Coresets

What should a probabilistic coreset look like?

- Consists of **probabilistic** points (nodes)
- The probability distributions of the nodes should be **sparse**
Probabilistic Coresets

What should a probabilistic coreset look like?
- Consists of probabilistic points (nodes)
- The probability distributions of the nodes should be sparse

Coresets

Given a set of uncertain nodes $V$, a weighted subset $U$ is a $(k, \varepsilon)$-coreset if for all sets $C$ of $k$ centers it holds

$$|E_{D'}[\text{cost}_w(U, C)] - E_D[\text{cost}(V, C)]| \leq \varepsilon E_D[\text{cost}(V, C)]$$

where $E_{D'}[\text{cost}_w(U, C)] := \min_{\rho:U\rightarrow C} \sum_{v_i \in U} \sum_{j=1}^{m} p'_{ij} w(v_i) D(x_j, \rho(v_i))$. 
Metric $k$-median

Idea

Extend cost function to a metric (so far only defined for a tuple of a node and a center)

Point $c \in X \mapsto$ node with all probability at $c$

Generalization of cost function to distance between nodes?
Metric $k$-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
Metric $k$-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at $c$
Metric $k$-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \sim$ node with all probability at $c$
- Generalization of cost function to distance between nodes?
Metric $k$-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at $c$
- Generalization of cost function to distance between nodes?

- Expected distance?
Metric \( k \)-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point \( c \in X \sim \) node with all probability at \( c \)
- Generalization of cost function to distance between nodes?

- Expected distance?
- Expected distance between two copies of the same probabilistic node is not zero
Metric $k$-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at $c$
- Generalization of cost function to distance between nodes?

- Expected distance?
- Expected distance between two copies of the same probabilistic node is not zero
- $\rightsquigarrow$ expected distance is not a metric
Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \sim$ node with all probability at $c$
- Generalization of cost function to distance between nodes?
Metric $k$-median

**Idea**

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \sim$ node with all probability at $c$
- Generalization of cost function to distance between nodes?

```
0.2 0.2

0.2 0.4 0.2

0.5 0.3

0.2 0.2

0.5
```
Earth Mover Distance

Let \( v_{i_1}, v_{i_2} \in V \) and let \( p_{i_1} = p_{i_2} \). A mapping \( \varrho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0} \) morphs \( v_{i_1} \) into \( v_{i_2} \) if for all \( x_{j_1}, x_{j_2} \in \mathcal{X} \):

\[
\sum_{x_{j_1} \in \mathcal{X}} \varrho(x_{j_1}, x_j) = p_{i_1j_1} \quad \text{and} \quad \sum_{x_{j_2} \in \mathcal{X}} \varrho(x_j, x_{j_2}) = p_{i_2j_2}.
\]

The cost of \( \varrho \) is defined as

\[
\sum \sum \varrho(x_{j_1} x_{j_2}) \cdot D(x_{j_1}, x_{j_2}).
\]

The earth mover distance EMD between \( v_{i_1} \) and \( v_{i_2} \) is the minimum cost of a mapping that morphs \( v_{i_1} \) into \( v_{i_2} \).
Morphing Probability Distributions
**EMD is a metric**
Morphing Probability Distributions

- EMD is a metric
- EMD is a generalization of the cost function
EMD is a metric
EMD is a generalization of the cost function
for each \( x \in C \), create an artificial node \( \sim \rightarrow C' \)
- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in C$, create an artificial node $\sim \rightarrow C'$
- A deterministic $(k, \varepsilon)$-coreset for $V$ with center set $C'$ and metric EMD is a probabilistic $(k, \varepsilon)$-coreset
Morphing Probability Distributions

- EMD is a metric
- EMD is a generalization of the cost function
- for each \( x \in C \), create an artificial node \( \sim \to C' \)
- A deterministic \((k, \varepsilon)\)-coreset for \( V \) with center set \( C' \) and metric EMD is a probabilistic \((k, \varepsilon)\)-coreset
  - if we thin out the probability distributions
EMD is a metric

EMD is a generalization of the cost function

for each $x \in C$, create an artificial node $\sim C'$

A deterministic $(k, \varepsilon)$-coreset for $V$ with center set $C'$ and metric EMD is a probabilistic $(k, \varepsilon)$-coreset

– if we thin out the probability distributions

– for uniform realization probabilities.
EMD is a metric
EMD is a generalization of the cost function
for each $x \in C$, create an artificial node $\sim \in C'$
A deterministic $(k, \varepsilon)$-coreset for $V$ with center set $C'$ and metric EMD is a probabilistic $(k, \varepsilon)$-coreset
– if we thin out the probability distributions
– for uniform realization probabilities.

Fixing Issues
**Morphing Probability Distributions**

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in C$, create an artificial node $\sim C'$
- A deterministic $(k, \varepsilon)$-coreset for $V$ with center set $C'$ and metric EMD is a probabilistic $(k, \varepsilon)$-coreset
  - if we thin out the probability distributions
  - for uniform realization probabilities.

**Fixing Issues**

- for general realization probabilities, group nodes and round to $p_{\min}(1 + \varepsilon)^{\ell} \rightarrow$ error is a factor $(1 + \varepsilon)$
Morphing Probability Distributions

- EMD is a metric
- EMD is a generalization of the cost function
- for each \( x \in C \), create an artificial node \( \sim \rightarrow C' \)
- A deterministic \((k, \varepsilon)\)-coreset for \( V \) with center set \( C' \) and metric EMD is a probabilistic \((k, \varepsilon)\)-coreset
  - if we thin out the probability distributions
  - for uniform realization probabilities.

Fixing Issues

- for general realization probabilities, group nodes and round to \( p_{\min}(1 + \varepsilon)^{\ell} \rightarrow \) error is a factor \((1 + \varepsilon)\)
- to compute the EMD efficiently, use det. \((1, \varepsilon)\)-coresets of the nodes \( \rightarrow \) also ensures sparsity of coreset nodes
Theorem

A probabilistic $(k, \varepsilon)$-coreset of size

\[ \mathcal{O}(k\varepsilon^{-3} \cdot \text{polylog}(|C|, n, \delta, 1/p_{\text{min}})) \]

can be computed in time

\[ \mathcal{O}(nm + \varepsilon^{-10}kn \cdot \text{polylog}(|C|, n, m, \delta, 1/p_{\text{min}})) \]

with error probability $\delta$. The probability distributions have size

\[ \mathcal{O}(\varepsilon^{-3} \cdot \text{polylog}(|C|, n, \delta, 1/p_{\text{min}})). \]
Partitioning nodes

Coresets for Probabilistic Clustering
Does the same approach work in the Euclidean case?
Does the same approach work in the Euclidean case?

- In the general metric case, $C$ is usually finite (e.g. $P$)
Does the same approach work in the Euclidean case?

- In the general metric case, $C$ is usually finite (e.g. $P$).
- In the Euclidean case, one usually sets $C = \mathbb{R}^d$.

→ Develop coreset construction
→ Use deterministic coreset construction by Chen
Does the same approach work in the Euclidean case?

- in the general metric case, $C$ is usually finite (e.g. $P$)
- in the Euclidean case, one usually sets $C = \mathbb{R}^d$.

⇝ algorithms for the general case do not work here

coreset construction

⇝ Develop coreset construction
⇝ Use deterministic coreset construction by Chen
Does the same approach work in the Euclidean case?

- in the general metric case, $C$ is usually finite (e.g. $P$)
- in the Euclidean case, one usually sets $C = \mathbb{R}^d$.

Therefore, algorithms for the general case do not work here.

Even though probabilistic Euclidean $k$-median can be seen as deterministic metric $k$-median, we cannot use deterministic algorithms.

Therefore, develop coreset construction.

Therefore, use deterministic coreset construction by Chen.
Introduction

Metric $k$-median

Euclidean $k$-median

Partitioning nodes

Chen (2006) compute bicriteria approximation partition input points into subsets of points which are close to each other compared to the optimal clustering cost. Sample representatives from each subset.
Chen (2006)
Chen (2006)

- compute bicriteria approximation
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset
Chen (2006)

- Compute bicriteria approximation
- Partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- Sample representatives from each subset
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset
Probabilistic Coreset Construction
Probabilistic Coreset Construction

- Compute 1-medians
Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition
Introduction

Metric $k$-median

Euclidean $k$-median

End

Partitioning nodes

Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

Analysis
### Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

### Analysis

- Show that 1. yields a bicriteria approximation for the probabilistic problem
Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

Analysis

- Show that 1. yields a bicriteria approximation for the probabilistic problem
- Find nice formulation of error in 4.
Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

Analysis

- Show that 1. yields a bicriteria approximation for the probabilistic problem
- Find nice formulation of error in 4.
- Bound error by analyzing properties of 3.
Result for Euclidean $k$-median

Theorem

A probabilistic $(k, \varepsilon)$-coreset of size

$$O(k^2 \varepsilon^{-2} d \cdot \text{polylog}(n, \delta, \varepsilon^{-1}, 1/p_{\min}))$$

can be computed in time

$$O(knm \cdot \text{polylog}(n, \delta, \varepsilon^{-1}, 1/p_{\min}))$$

with error probability $\delta$. The probability distributions have size

$$O(\varepsilon^{-2} d \cdot \text{polylog}(n, \delta, \varepsilon^{-1}, 1/p_{\min})).$$
Thank you for your attention!