Near-optimal Algorithms for Explainable k-Medians and k-Means

Konstantin Makarychev, Liren Shan

Northwestern University

k-medians and k-means

Input:

- A set of n points, $X = \{x_1, x_2, \cdots, x_n\}$
- Number of clusters k

Output:

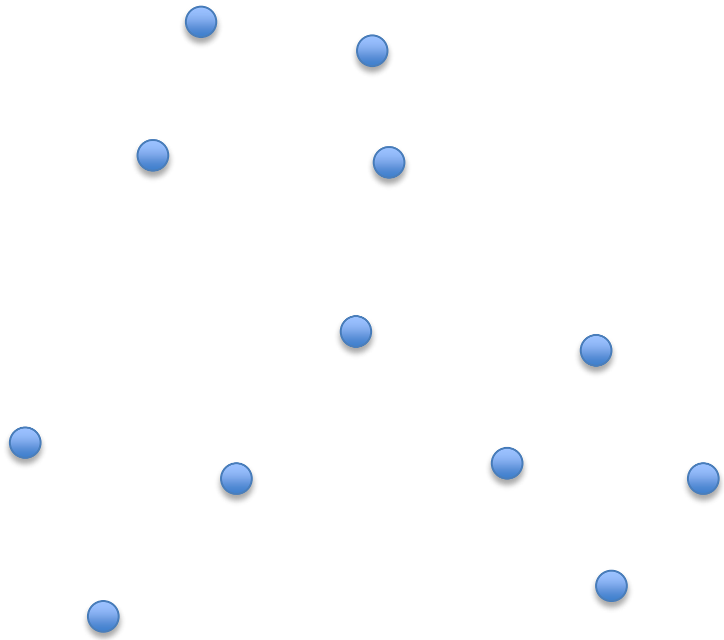
A set of k centers, $C = \{c^1, c^2, \cdots, c^k\}$

Assign x to $c_x = \arg\min_{c \in C} \mathrm{cost}(x, c)$

k-medians in ℓ_1 : $\mathrm{cost}(x, c) = \|x - c\|_1$

k-medians in ℓ_2 : $\mathrm{cost}(x, c) = \|x - c\|_2$

k-means : $\mathrm{cost}(x, c) = \|x - c\|_2^2$



k-means, k=3

k-medians and k-means

Input:

- A set of n points, $X = \{x_1, x_2, \cdots, x_n\}$
- Number of clusters k

Output:

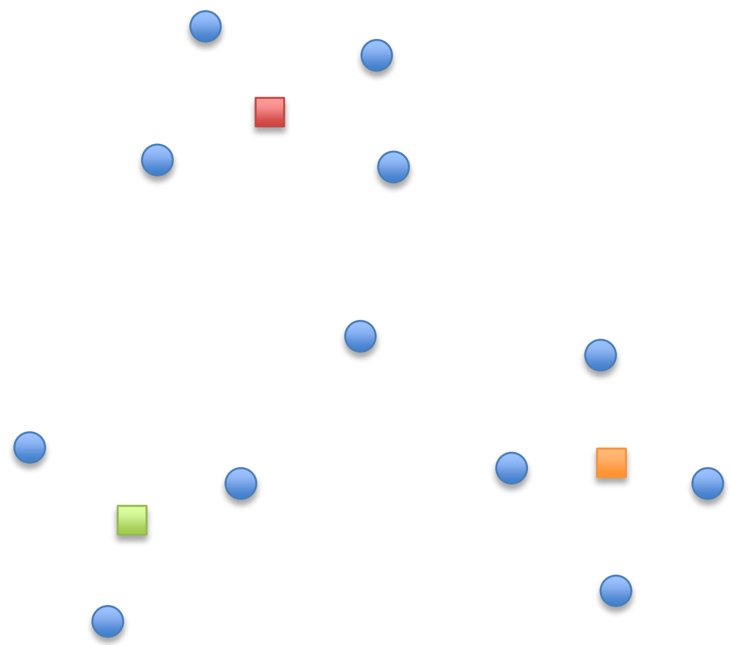
A set of k centers, $C = \{c^1, c^2, \cdots, c^k\}$

Assign x to $c_x = \arg\min_{c \in C} \operatorname{cost}(x, c)$

k-medians in ℓ_1 : $\operatorname{cost}(x, c) = \|x - c\|_1$

k-medians in ℓ_2 : $\operatorname{cost}(x, c) = \|x - c\|_2$

k-means : $\operatorname{cost}(x, c) = \|x - c\|_2^2$



k-means, k=3

k-medians and k-means

Input:

- A set of n points, $X = \{x_1, x_2, \cdots, x_n\}$
- Number of clusters k

Output:

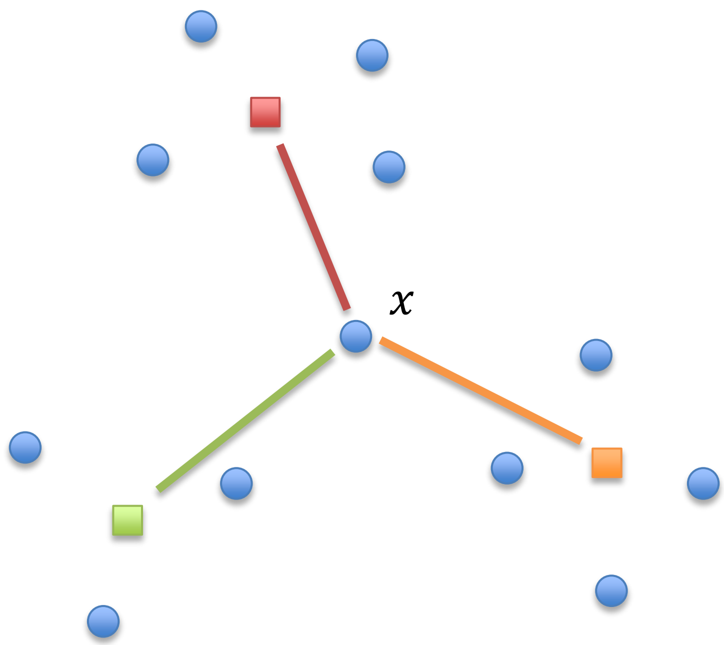
A set of k centers, $C = \{c^1, c^2, \cdots, c^k\}$

Assign x to $c_x = \arg\min_{c \in C} \text{cost}(x, c)$

k-medians in ℓ_1 : $\text{cost}(x, c) = \|x - c\|_1$

k-medians in ℓ_2 : $\text{cost}(x, c) = \|x - c\|_2$

k-means : $\text{cost}(x, c) = \|x - c\|_2^2$



k-means, k=3

k-medians and k-means

Input:

- A set of n points, $X = \{x_1, x_2, \cdots, x_n\}$
- Number of clusters k

Output:

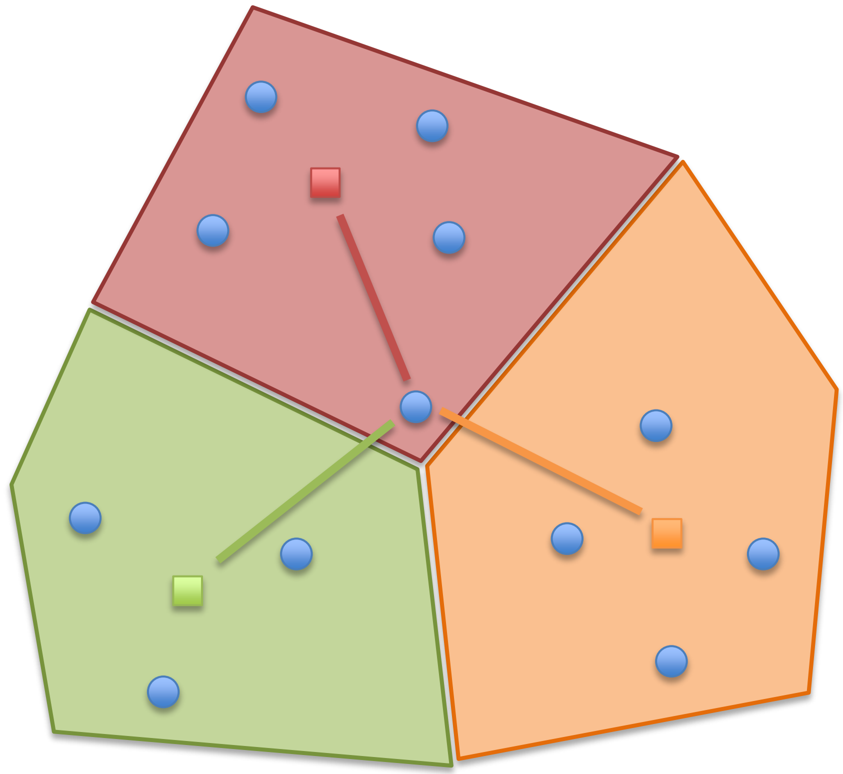
A set of k centers, $C = \{c^1, c^2, \cdots, c^k\}$

Assign x to $c_x = \arg\min_{c \in C} \mathrm{cost}(x, c)$

k-medians in ℓ_1 : $\mathrm{cost}(x, c) = \|x - c\|_1$

k-medians in ℓ_2 : $\mathrm{cost}(x, c) = \|x - c\|_2$

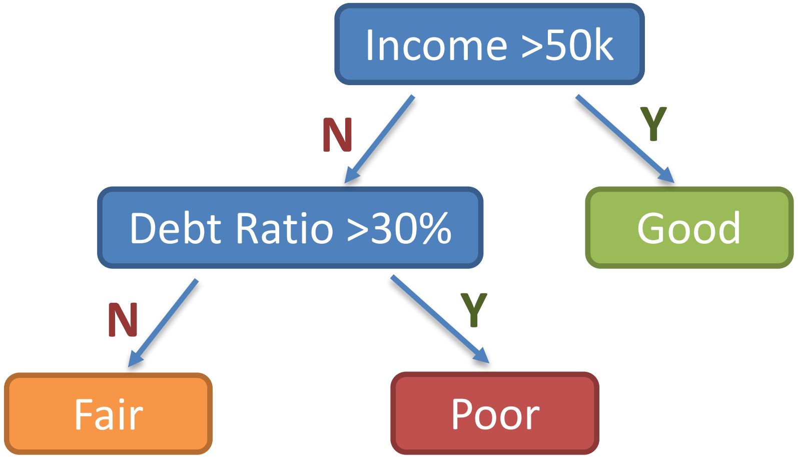
k-means : $\mathrm{cost}(x, c) = \|x - c\|_2^2$



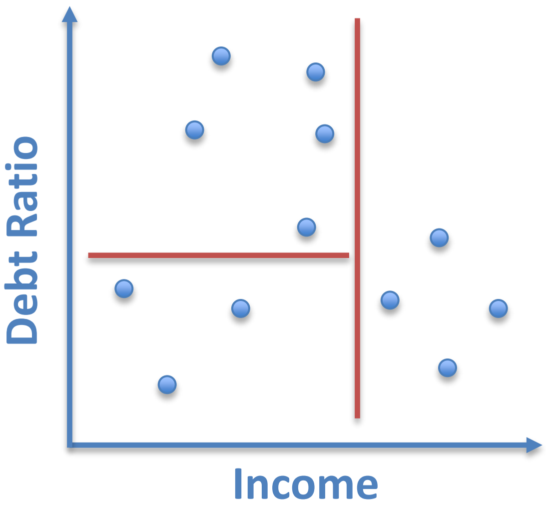
k-means, k=3

Explainable Clustering

[Dasgupta, Frost, Moshkovitz, and Rashtchian, 2020] proposed to use **threshold trees** to describe clusters.

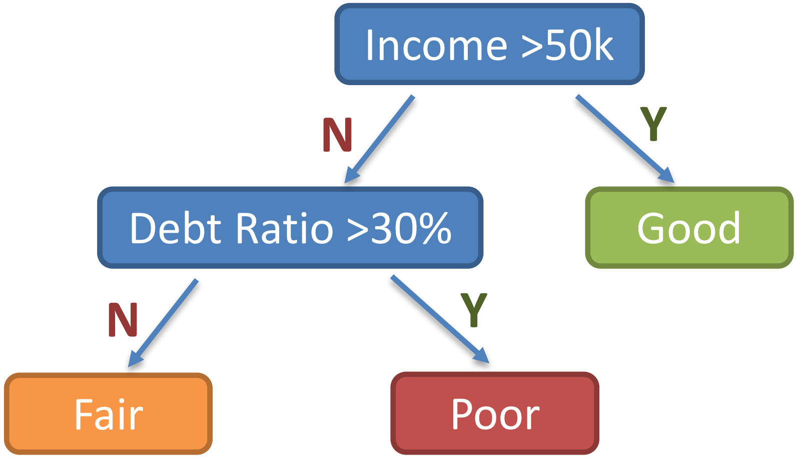


Loan Risk Decision Tree

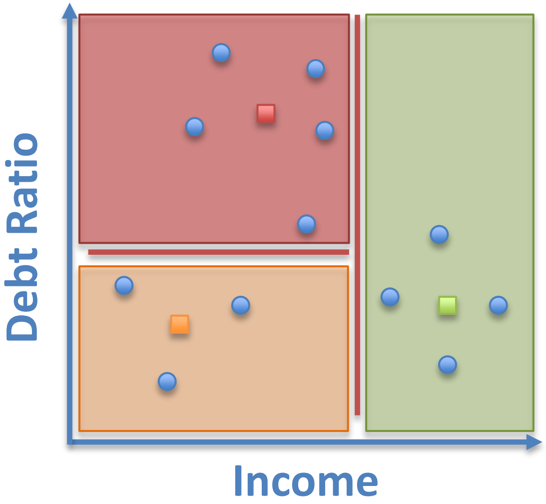


Explainable Clustering

[Dasgupta, Frost, Moshkovitz, and Rashtchian, 2020] proposed to use **threshold trees** to describe clusters.

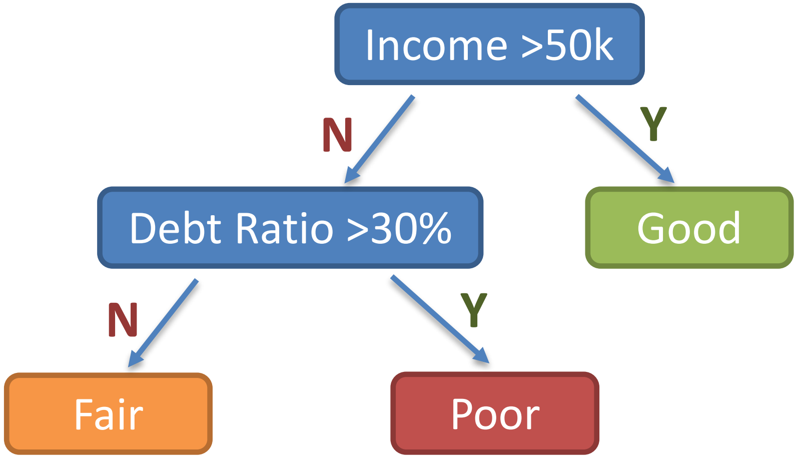


Loan Risk Decision Tree

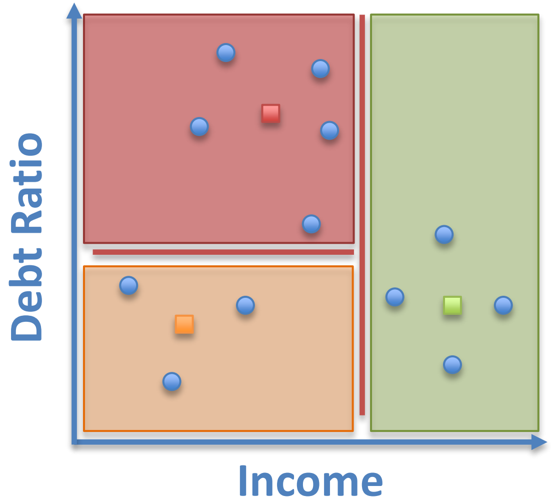


Explainable Clustering

[Dasgupta, Frost, Moshkovitz, and Rashtchian, 2020] proposed to use **threshold trees** to describe clusters.



Loan Risk Decision Tree



Question: Can we find a good explainable clustering?

Explainable Clustering

[Dasgupta et al, 2020] defined the price of explainability as

$$\frac{\mathrm{cost}(X, T)}{\mathrm{OPT}(X)},$$

where $\mathrm{cost}(X, T)$ is the cost of threshold tree T, $\mathrm{OPT}(X)$ is the optimal cost of regular k-medians (k-means) clustering.

Explainable Clustering

[Dasgupta et al, 2020] defined the price of explainability as

$$\frac{\mathrm{cost}(X,T)}{\mathrm{OPT}(X)},$$

where $\mathrm{cost}(X,T)$ is the cost of threshold tree T, $\mathrm{OPT}(X)$ is the optimal cost of regular k-medians (k-means) clustering.

	k-medians in ℓ_1	k-means
Upper Bound	$O(k)$	$O(k^2)$
Lower Bound	$\Omega(\log k)$	$\Omega(\log k)$

Our Results

In this work, we provide almost tight bounds for explainable k-medians in ℓ_1 and k-means clustering.

We also get upper and lower bounds for explainable k-medians in ℓ_2

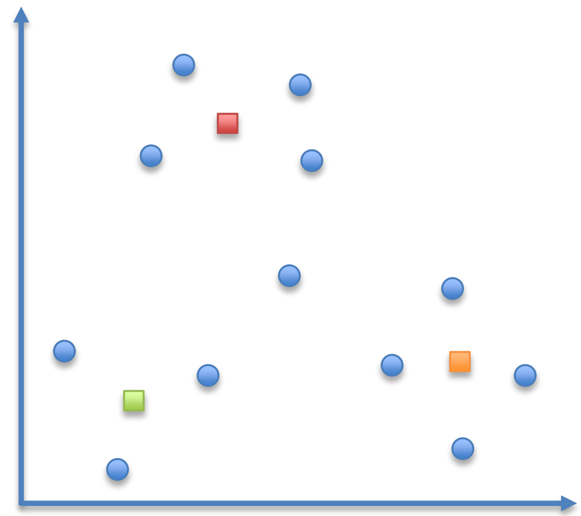
	k-medians in ℓ_1	k-means	k-medians in ℓ_2
Upper Bound	$\tilde{O}(\log k)$	$\tilde{O}(k)$	$O(\log^{3/2} k)$
Lower Bound	$\Omega(\log k)^*$	$\tilde{\Omega}(k)$	$\Omega(\log k)$

*: provided by [Dasgupta, et al, 2020]

Explainable k-medians in ℓ_1

Algorithm:
Input: k centers C
Output: a threshold tree T

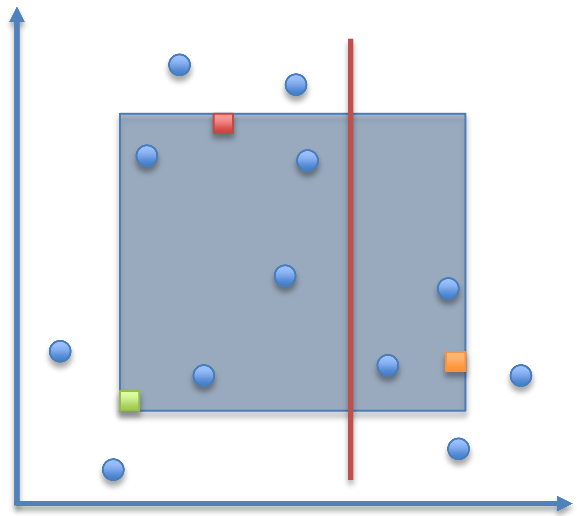
Iteratively split these centers by **uniformly sampling** a threshold cut



Explainable k-medians in ℓ_1

Algorithm:
Input: k centers C
Output: a threshold tree T

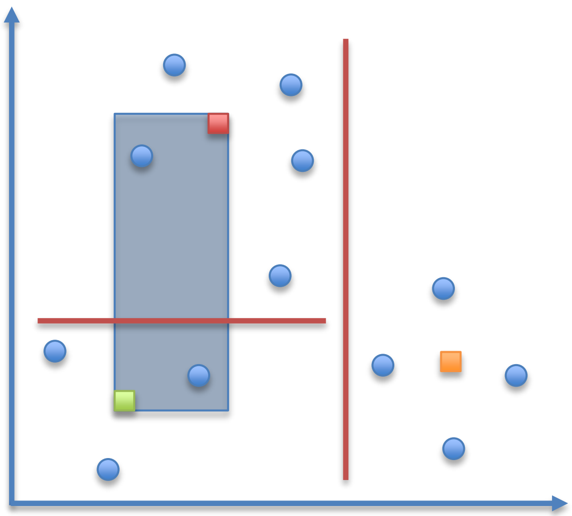
Iteratively split these centers by **uniformly sampling** a threshold cut



Explainable k-medians in ℓ_1

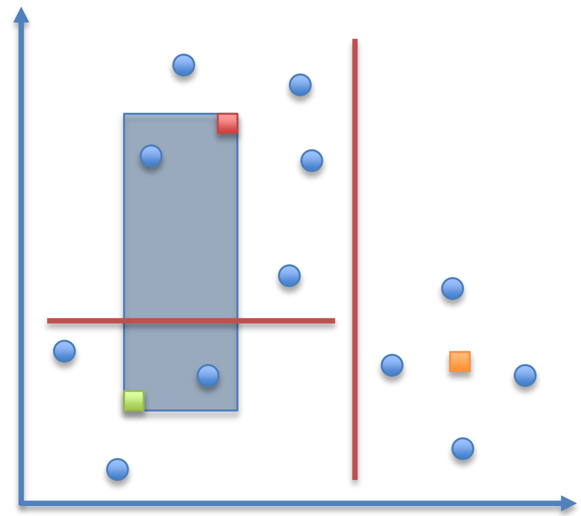
Algorithm:
Input: k centers C
Output: a threshold tree T

Iteratively split these centers by **uniformly sampling** a threshold cut



Explainable k-medians in ℓ_1

Algorithm:
Input: k centers C
Output: a threshold tree T



Iteratively split these centers by **uniformly sampling** a threshold cut

Given a set of points X and a set of centers C, we have $\mathbb{E}_T[\mathrm{cost}(X,T)] \le O(\log k \cdot \log\log k) \cdot \mathrm{cost}(X,C)$.

Explainable k-means

- We use the **Terminal Embedding** $\boldsymbol{\varphi}$ to embed space ℓ_2 into ℓ_1 with distortion $O(k)$, i.e., for every $x \in X, c \in C$

$$\|\varphi(x) - \varphi(c)\|_1 \leq \|x - c\|_2^2 \leq 8k\|\varphi(x) - \varphi(c)\|_1.$$

- Then, we use our algorithm for explainable k-medians in ℓ_1 on the instance after embedding.

Given a set of points X and a set of centers C, we have $\mathbb{E}_T[\mathrm{cost}(X, T)] \leq O(k \log k \cdot \log\log k) \cdot \mathrm{cost}(X, C)$.

Thank you