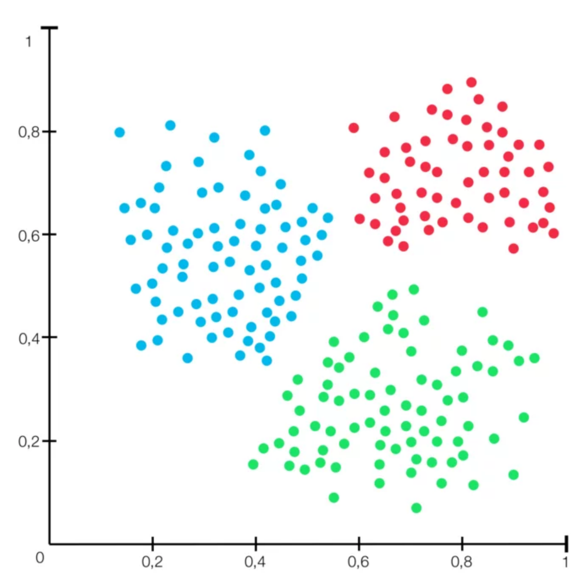
NEARLY-TIGHT AND OBLIVIOUS ALGORITHMS FOR EXPLAINABLE CLUSTERING

Buddhima Gamlath, Xinrui Jia, Adam Polak, Ola Svensson

Classic clustering



Given a set of points $X \subseteq \mathbb{R}^d$, find a set of k centers C that

- k-medians:

$$\text{minimizes} \quad \sum_{x \in X} \min_{c \in C} \ell_1(x, c);$$

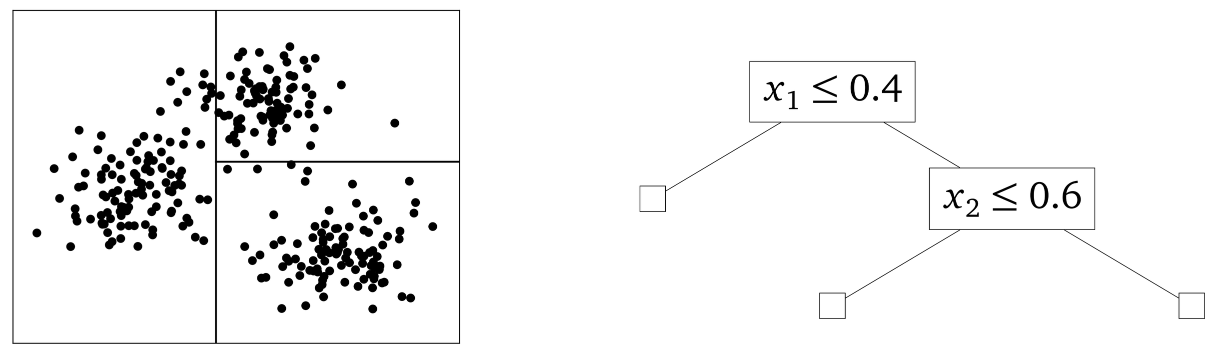
- k-means:

$$\text{minimizes} \quad \sum_{x \in X} \min_{c \in C} \ell_2^2(x, c).$$

There exist constant factor approximation algorithms.
But how can we explain why a point belongs to a particular cluster?

Explainable clustering [Dasgupta, Frost, Moshkovitz, Rashtchian, ICML'20]

Clustering explained by **axis-aligned threshold cuts**.



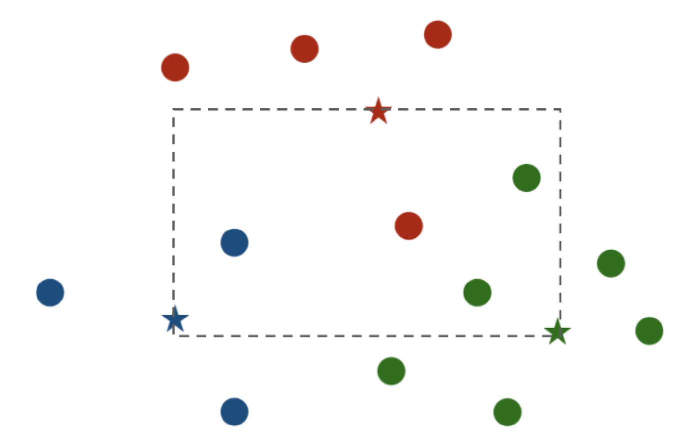
Paths from root to leaf in the **threshold tree** explain why points belong to a cluster.

Price of explainability: How much more expensive is an explainable clustering?

Previous and concurrent work

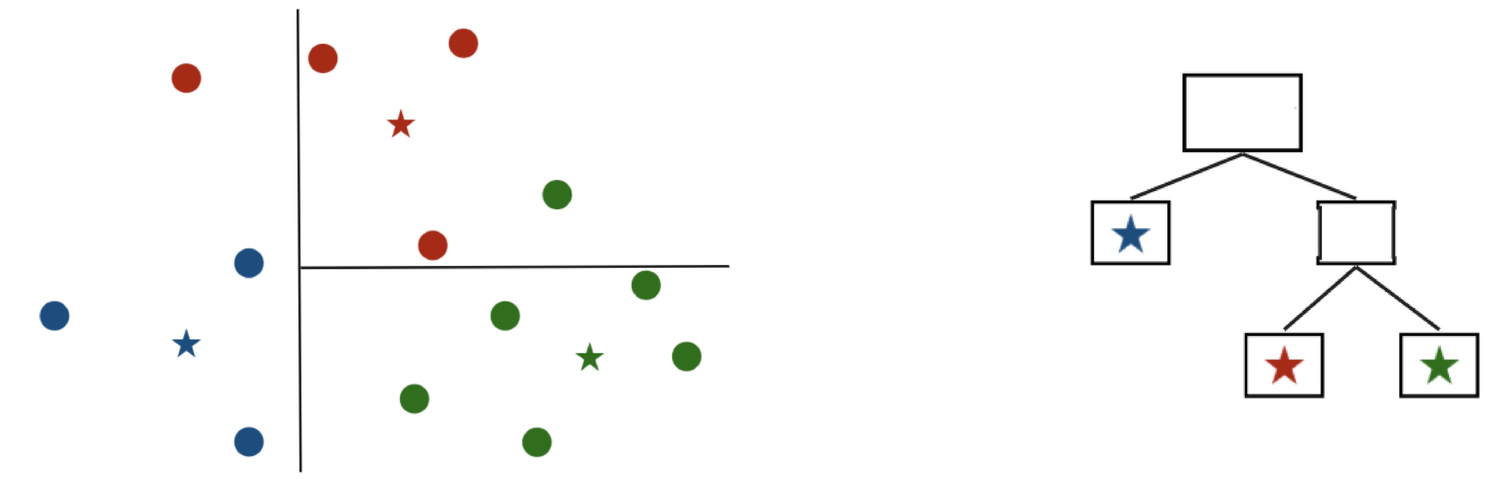
	k-medians	k-means	ℓ_p-norm	
Algorithms	$O(k)$	$O(k^2)$		Dasgupta et al.
	$O(d \log k)$	$O(kd \log k)$		Laber and Murtinho
	$O(\log^2 k)$	$O(k \log^2 k)$	$O(k^{p-1} \log^2 k)$	This paper
	$O(\log k \log\log k)$	$O(k \log k \log\log k)$		Makarychev and Shan
	$O(\log k \log\log k)$	$O(k \log k)$		Esfandiari et al.
	$O(d \log^2 d)$			Esfandiari et al.
		$O(k^{1-2/d} \text{polylog} k)$		Charikar and Hu
Lower bounds	$\Omega(\log k)$	$\Omega(\log k)$		Dasgupta et al.
		$\Omega(k)$	$\Omega(k^{p-1})$	This paper
		$\Omega(k/\log k)$		Makarychev and Shan
	$\Omega(\min(d, \log k))$	$\Omega(k)$		Esfandiari et al.
		$\Omega(k^{1-2/d}/\text{polylog} k)$		Charikar and Hu

Algorithm for explainable k-medians



1. Start with a non-explainable clustering.
2. Compute bounding box of centers. $L = \sum_{i=1}^{d} L_i$, the sum of all side lengths.
3. Keep sampling random threshold cuts (i, θ)
   - i with probability L_i / L,
   - $\theta \in L_i$ uniformly at random.

In the stream of random cuts, take a cut if it separates some centers, until a threshold tree is formed, i.e. each center has its own leaf.

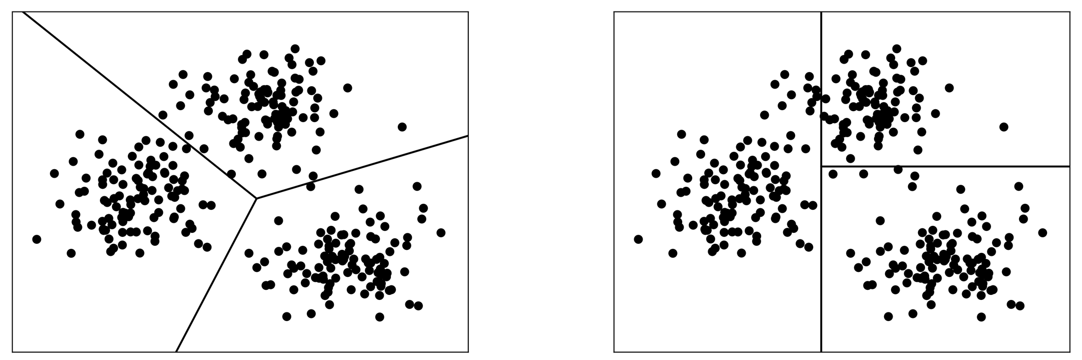


The algorithm is oblivious to data points, and runs in time $\tilde{O}(kd)$.

Summary and open problems

We have a nearly tight understanding of the price of explainability:

$$\Omega(k^{p-1}) \cdot OPT \leqslant \textbf{cost of explainable clustering} \leqslant O(k^{p-1} \log^2 k) \cdot OPT.$$



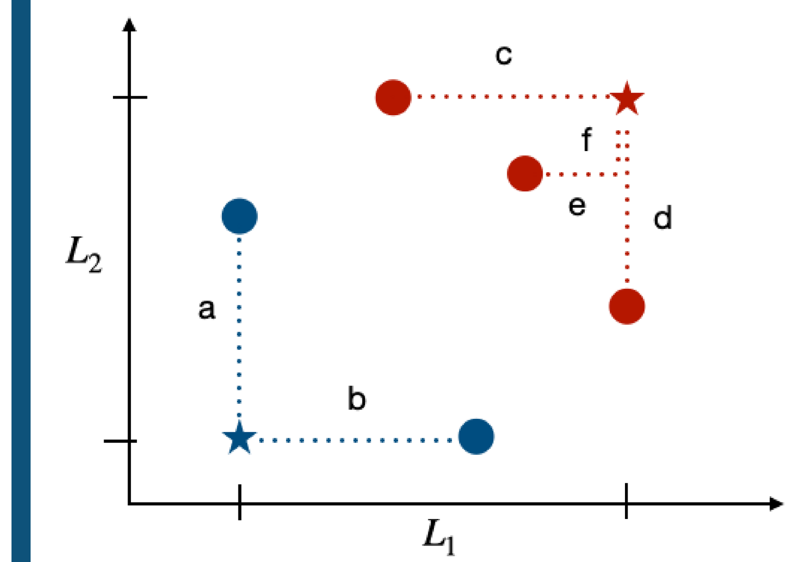
Conjecture. The expected cost of our algorithm for k-medians is at most

$$(1 + H_{k-1}) \cdot OPT \leqslant O(\log k) \cdot OPT.$$

What's next?

- Generalize the notion of explainability, e.g., allow in each node hyperplanes in a small number of dimensions.
- Define natural clusterability assumptions under which the price of explainability is lower.

Analysis of the algorithm



Warm up: two centers

$\Pr[\text{random cut separates } x \text{ from its center } c(x)] \leqslant \ell_1(x, c(x))/L$

$\mathbb{E}[\#\text{ of points that get separated}] \leqslant \sum_x \ell_1(x, c(x))/L = OPT/L$

Cost increase for each separated point $\leqslant L$

$\Rightarrow$ Cost of explainable clustering $\leqslant 2OPT$

A naive bound for k centers

We may need $k-1$ cuts to separate k centers $\Rightarrow$ $OPT + (k-1) \cdot \frac{OPT}{L} \cdot L = k \cdot OPT$

A refined bound for k centers

How many *random* cuts to separate all centers?

- Let $C_{\max}$ be the largest distance between two centers, $C_{\min}$ the smallest
- For a fixed pair of centers at least $C_{\max}/2$ apart,

$$\Pr[\text{random cut does not separate them}] \leqslant 1 - C_{\max}/2L$$

- Take $100 \cdot (2L/C_{\max}) \cdot \log k$ successive random cuts

$$\left(1 - \frac{C_{\max}}{2L}\right)^{100 \cdot (2L/C_{\max}) \cdot \log k} \leqslant \frac{1}{k^{10}}$$

$\Rightarrow$ With high probability, all such pairs of centers are separated

How much these cuts cost?

- Cost increase of $O(L/C_{\max} \cdot \log k)$ cuts?

$$O\left(\frac{L}{C_{\max}} \cdot \log k\right) \cdot \frac{OPT}{L} \cdot C_{\max} = O(\log k) \cdot OPT$$

- Going from $C_{\max} \to C_{\max}/2$ increases cost by $O(\log k) \cdot OPT$
- Repeating $O(\log(C_{\max}/C_{\min}))$ times gives

$$O(\log(C_{\max}/C_{\min}) \cdot \log k) \cdot OPT$$

How to get $O(\log^2 k) \cdot OPT$?

- Automatic if $C_{\max}$ and $C_{\min}$ are polynomially related
- Otherwise, forbid cuts that separate centers that are too close
- While reducing $C_{\max} \to C_{\max}/2$, forbid separating center pairs closer than $C_{\max}/k^4$