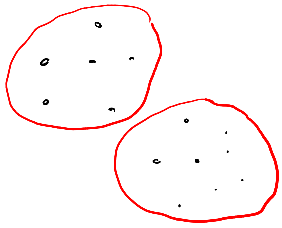
$\mathcal{X} = (x_1, x_2 \ldots x_n)$



$K = 2$

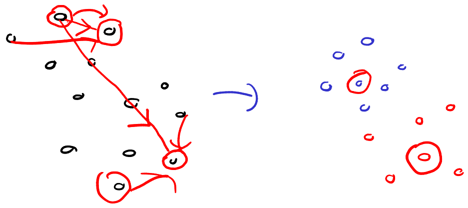
K Centroid

$\{C_1, C_2 \ldots C_k\}$

$$\mathrm{Cost}(\mathcal{X}, C) = \sum_{x \in \mathcal{X}} \min_{c \in C} \|p - c\|_2^2$$

$P = (x_1, y_1) \quad c = (x_2, y_2)$

$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$



- Given an initial set of K-means

$m_1^{(1)}, \ldots, m_k^{(1)} \longrightarrow$ Random

(1) Assignment :- $C_i^{(t)} = \{ x_p : \|x_p - m_i^{(t)}\|_2^2 \le \|x_p - m_j^{t}\|_2^2 \;\forall\; j \in [K] \}$

(2) Update :- $\dfrac{1}{|C_i^{(t)}|} \displaystyle\sum_{x_j \in C_i^{(t)}} x_j$

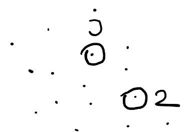
$\mathcal{X} \in \mathbb{R}^d$

Running time $O(nkdi)$

- NP-hard
- 1.07
- <u>Learning Augment K-means Clustering Ergun at al. ICLR 23</u>

$$\mathrm{Cost}(P, C) = \sum_{p \in P} \min_{c \in C} \|P - c\|_2^2$$

$\Pi \rightarrow$

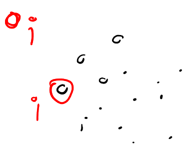


$\{1, \ldots k\}$

K

$\Pi \rightarrow (1 + \alpha)$ approximat Solⁿ

label error rate $\lambda \le \alpha$



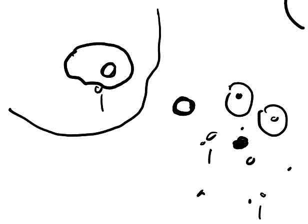
m_i

αm_i

Let $\bar{C}$ be the Centre Output of the algorithm

$$\text{Cost}(P, \bar{C}) \le (1 + O(\alpha))\, \text{Cost}(P, C^{OPT})$$

$$(1 + 20\alpha) \qquad \tilde{O}(\bar{k} n d)$$



Algo

Input :- A Point Set X with labels given by π $\quad d \le \alpha$

Output :- $(1 + O(\alpha))$

for $i = 1$ to K

Let Y_i be the Set of Point with label i

Run CRDEST for each of the d-Coordinates of Y_i

Let C_i'

end for

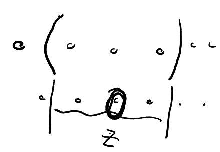
Return $C_1' \dots C_k'$

CRDEST

Input :- $x_1, x_2 \dots x_{2m} \in \mathbb{R}$

Randomly Partition X_1 and X_2 of Size m

Let $I = [a, b]$ Shortest interval s.t. $m(1 - 5\alpha)$ Points of X_1



$$Z = X_2 \cap I$$

$$\dot{z} = \frac{1}{|Z|} \sum_{x \in Z} x$$

Return z

Lemma 1: Let $P, Q \subseteq \mathbb{R}$ be the set of points s.t $|P| \geq (1-\alpha)n$

$|Q| \leq \alpha n$. Let $X = P \cup Q$

C_P be the mean of P, C_X be the mean of X.

Then

$$\mathrm{Cost}(X, C_P) \leq \left(1 + \frac{\alpha}{(1-\alpha)^2}\right) \mathrm{Cost}(X, C_X)$$

Proof:- Suppose w.l.o.g. that $C_X = 0$, $C_P \leq 0$ then $C_Q \geq 0$

Now

$$\mathrm{Cost}(X, C_P) = \underbrace{\mathrm{Cost}(X, C_X) + |X|\,|C_P - C_X|^2}_{\text{Inaba et al.}}$$

$$|X| \cdot |C_P - C_X|^2 \leq \frac{\alpha}{(1-\alpha)^2} \mathrm{Cost}(X, C_X)$$

As we know $C_X = 0$

$$|P| \cdot C_P = -|Q| \cdot C_Q$$

Let $|P| = (1-\delta)n$, $|Q| = \delta n$ $\delta \leq \alpha$

$$C_Q = -\frac{|P| \cdot C_P}{|Q|} = -\frac{(1-\delta)}{\delta} C_P$$

$$\mathrm{Cost}(Q, C_X) \geq |Q|\, C_Q^2$$

$$\geq |Q| \frac{(1-\delta)^2}{\delta^2} C_P^2$$

$$\geq n \frac{(1-\delta)^2}{\delta} C_P^2$$

$$\geq n \frac{(1-\alpha)^2}{\alpha} C_P^2$$

$$|C_P - C_X|^2 = |C_P|^2$$

$$= \frac{\alpha}{n(1-\alpha)^2} \mathrm{Cost}(Q, C_X)$$

$$\leq \frac{\alpha}{\cancel{n}(1-\alpha)^2} \mathrm{Cost}(X, C_X)$$

$I^* \subset \mathbb{R}$

Lemma 2 For a fixed set $X \subseteq \mathbb{R}$, let C be the mean of X & $a^2 = \frac{1}{2|X|} \sum_{x \in X} (x-C)^2$ be the variance.

$I^* = \left[C - \frac{a}{\sqrt{\alpha}},\ C + \frac{a}{\sqrt{\alpha}} \right]$ contains $\geq (1-4\alpha)$ fraction points of X

Proof:-

$$\overset{I^*}{\left[C - \frac{a}{\sqrt{\alpha}} \qquad \underset{C}{\cdot} \qquad C + \frac{a}{\sqrt{\alpha}} \right]}$$

$x \in X \setminus I^*$ satisfy $|x - C|^2 \geq \frac{a^2}{4\alpha}$

Now if more than 4α fraction of point X is that the total variance is larger than a^2.

Lemma 3 Let m be a suff. large constant. $I = [a, b]$ $\geq (1-6\alpha)$ fraction points of X_2 & $b - a \leq \frac{2a}{\sqrt{\alpha}}$ with high Pr.

I^*

Proof:- Suppose we have a set of $2m$ points that X_1 & X_2

Consider an interval J that have atleast $2Cm$ points $C\ (\geq 1-5\alpha)$

Let J_1 & J_2 be the number of points in J that are in X_1 & X_2, resp.

By Chernoff we have

J_1 and J_2 are atleast $mC(1-\alpha)$ with high Pr.

$|J_1 - J_2| \leq \alpha m C$ with high Pr.

As I contains $\geq m(1-5\alpha)$ X_1 and $|J_1 - J_2| \leq \alpha m C$

I must also contain $m(1-6\alpha)$ of X_2 $\square$.

Lemma 4 Let S be a Set of Points $X \subset R$ with $Pr = \frac{1}{2}$.

Let C the optimal Centre of X, and C_S be the Centre of S.

$$Pr\left[Cost(X, C_S) > (1+\alpha)\,Cost(X, C)\right] < \frac{1}{\eta^{k}} \qquad n > 1$$

Thm:- Let $\alpha \in \left(10\frac{\log n}{\sqrt{n}}, 1/7\right)$. Let $P, Q \subset R$ s.t. $|P| \geq (1-\alpha)2m$

$|Q| \leq 2\alpha m$ & $X = P \cup Q$.

Let C be the Centre of P. Then CRDEST

$$Cost(P, C') \leq \underline{(1+18\alpha)(1+\alpha)\left(1 + \frac{\alpha}{(1-\alpha)^2}\right)}\, Cost(P, C)$$

$$\leq (1 + 20\alpha)$$

Proof:- $I \cap X$ Contains atleast $(1-6\alpha)m$ points of $P \cap X_2$, $2\alpha m$ Points of Q in the length $\frac{2a}{\sqrt{\alpha}}$

$$a^2 =$$

From Lemma 1,

$$Cost(P, C_0) \leq \left(1 + \frac{\alpha}{(1-\alpha)^2}\right) Cost(P, C_1)$$

$C_0 \leftarrow I \cap P \cap X_2$ & $C_1 \leftarrow P \cap X_2$

From Lemma 4

$$Cost(P, C_1) \leq (1+\alpha)\, Cost(P, C) \text{ with high } Pr.$$

As C_0 is the Centre of $I \cap P \cap X_2$ & C' $I \cap X_2$
we have

$$|I \cap P \cap X_2|\, C_0 + \sum_{q \in I \cap Q \cap X_2} q = |I \cap X_2|\, C'$$

$$|C' - C_0| \leq 6\sqrt{\alpha}\, a$$

$$
\begin{aligned}
\mathrm{Cost}(P, c') &= \mathrm{Cost}(P, c_0) + |P| \, |c_0 - c'|^2 \\
&\leq \mathrm{Cost}(P, c_0) + |P| \, 36 \, \alpha \, \alpha^2 \\
&\leq \mathrm{Cost}(P, c_0) + \cancel{|P|} \cdot \overset{18}{\cancel{36}} \alpha \, \frac{1}{2 \cancel{|P|}} \mathrm{Cost}(P, C) \\
&\leq \mathrm{Cost}(P, c_0) + 18 \alpha \, \mathrm{Cost}(P, C)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cost}(P, c') &\leq (1 + 18\alpha)(1 + \alpha)\left(1 + \frac{\alpha}{(1-\alpha)^2}\right) \mathrm{Cost}(P, C) \\
&\leq (1 + 20\alpha) \, \mathrm{Cost}(P, C)
\end{aligned}
$$

$\square$