Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

The k-means problem

Outline

- k-means problem
- Lloyd's algorithm and its properties
- k-means++: Approximation guarantees
- Consistency of k-means
- Explainable k-means

Clustering problem

- $\mathcal{X} = \{x_1, \ldots, x_m\}$
  - Finite set to be clustered ... we mostly assume $\mathcal{X} \subset \mathbb{R}^p$
- Problem: Cluster $\mathcal{X}$ into k groups
  - $C_1, \ldots, C_k =$ disjoint partition of $\mathcal{X}$
  - $\mathcal{X} = \bigcup_{i=1}^{k} C_i$ and $C_i \cap C_j = \emptyset$ for $i \neq j$
- Similarity / dissimilarity examples
  - Euclidean distance $d(x, y) = \|x - y\|$ or Gaussian similarity $w(x, y) = e^{-\|x-y\|^2/\gamma^2}$

Fundamental challenge in analysing clustering

- Definition of Cluster analysis:
  - Cambridge dictionary: A way of studying or examining large amounts of data to find groups that are **more like each other** than they are like the data in other group
  - Wikipedia: The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms
- Difference from classification:
  - Objective may not be clear ... many methods are intuitive (not formal)
  - How do we define accuracy / empirical risk?

k-means algorithm (Lloyd's algorithm)

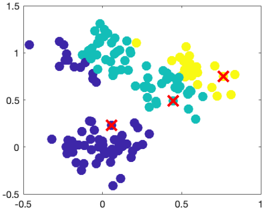
1. Initialise cluster centers $\mu_1, \ldots, \mu_k$

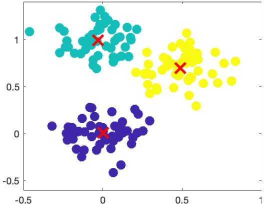
2. Define cluster $C_i = \{x \in \mathcal{X} \;:\; \mu_i \text{ is closest center for } x\}$

3. Re-estimate all centers $\mu_i = \dfrac{1}{|C_i|} \displaystyle\sum_{x \in C_i} x$

4. Repeat from step-2 until convergence

Initial clusters





After convergence

The k-means problem

- $\mathcal{X} = \{x_1, \ldots, x_m\} \subset \mathbb{R}^p$

- k-means cost:

  For set of k-centroids $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_k\} \in \mathbb{R}^p$,

$$\begin{aligned} G(\boldsymbol{\mu}) &= \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2 \qquad \ldots\ C_j = \{x \in \mathcal{X}\ :\ \mu_j \text{ is closest center for } x\} \\ &= \sum_{x \in \mathcal{X}} d(x, \boldsymbol{\mu})^2 \qquad \ldots\ d(x, \boldsymbol{\mu}) = \min_{\mu_j \in \boldsymbol{\mu}} \|x - \mu_j\| \end{aligned}$$

- k-means problem: $\quad \underset{\boldsymbol{\mu}\,:\,|\boldsymbol{\mu}| \leq k}{\text{minimise}}\ G(\boldsymbol{\mu})$

  $\ldots$ NP-Hard problem. Lloyd's algorithm is a greedy solution

Rewriting Lloyd's algorithm

1. Initialise cluster centers $\mu_1^{(0}, \ldots, \mu_k^{(0}$

2. Define cluster $C_j^{(0)} = \Big\{ x \in \mathcal{X} \ : \ \mu_j^{(0)} \text{ is closest center for } x \Big\}$

3. Continue until convergence: $t = 1, 2, \ldots$

   i. Estimate cluster centers

   $$\mu_j^{(t)} = \frac{1}{\left|C_j^{(t-1)}\right|} \sum_{x \in C_j^{(t-1)}} x \qquad \ldots \text{ solves } \quad \underset{\mu}{\text{minimise}} \sum_{x \in C_j^{(t-1)}} \|x - \mu\|^2$$

   ii. Reassign points to clusters

   $$C_j^{(t)} = \Big\{ x \in \mathcal{X} \ : \ \mu_j^{(t)} \text{ is closest center for } x \Big\} \qquad \ldots \text{ solves } \quad \underset{j}{\text{minimise}} \ \|x - \mu_j\|^2$$

Key questions about Lloyd's iterations

- Convergence: Do the iterations converge?
- Approximation: How good is the solution compared to optimal k-means cost?

$$G(\widehat{\boldsymbol{\mu}}) \quad \overset{?}{\leq} \quad G_{opt} \cdot \underbrace{\text{function}(m, k, \text{dimension})}_{\text{ideally we don't want } m}$$

  - $\widehat{\boldsymbol{\mu}} =$ solution of Lloyd's iterations
  - $G_{opt} = \min_{\boldsymbol{\mu}\,:\,|\boldsymbol{\mu}| \leq k} G(\boldsymbol{\mu})$

Convergence of Lloyd's iterations

Lemma kmeans.1 (k-means cost after one iteration of Lloyd's algorithm)

Let centers obtained from two consecutive iterations of Lloyd's algorithm be

$$\boldsymbol{\mu}^{(t)} = \left\{\mu_1^{(t)}, \ldots, \mu_k^{(t)}\right\} \quad \textit{and} \quad \boldsymbol{\mu}^{(t+1)} = \left\{\mu_1^{(t+1)}, \ldots, \mu_k^{(t+1)}\right\}$$

Then

$$G\left(\boldsymbol{\mu}^{(t+1)}\right) \leq G\left(\boldsymbol{\mu}^{(t)}\right)$$

Consequence: Iterations must convergence to a local minimum

- Reason: $G(\cdot)$ cannot decrease infinitely
- No guarantee on #iterations, or goodness of solution

Proof: Notation for iterations

- $C_j^{(t)} \;=\;$ cluster of points with closest center $\mu_j^{(t)}$ ... closest among $\boldsymbol{\mu}^{(t)}$

- k-means cost at iteration-t:

$$G\left(\boldsymbol{\mu}^{(t)}\right) \;=\; \sum_{x\in\mathcal{X}} d\big(x,\boldsymbol{\mu}^{(t)}\big)^2 \;=\; \sum_{j=1}^{k}\sum_{x\in C_j^{(t)}} \left\|x-\mu_j^{(t)}\right\|^2$$

- Re-computing centers: $\mu_j^{(t+1)} = \dfrac{1}{\left|C_j^{(t)}\right|}\displaystyle\sum_{x\in C_j^{(t)}} x$

- Reassignment: $C_j^{(t+1)} \;=\;$ points with closest center $\mu_j^{(t+1)}$ ... closest among $\boldsymbol{\mu}^{(t+1)}$

Proof: Center computation

- center of a cluster minimises the sum of squared distances to all points

$$\mu_j^{(t+1)} = \arg\min_{\nu \in \mathbb{R}^p} \sum_{x \in C_j^{(t)}} \|x - \nu\|^2$$

- From center computation step:

$$\sum_{j=1}^{k} \sum_{x \in C_j^{(t)}} \left\| x - \mu_j^{(t+1)} \right\|^2 \quad \leq \quad \sum_{j=1}^{k} \sum_{x \in C_j^{(t)}} \left\| x - \mu_j^{(t)} \right\|^2 \quad = \quad G\left(\boldsymbol{\mu}^{(t)}\right)$$

Proof: Reassignment

- Sum of squared distances to $\boldsymbol{\mu}^{(t+1)}$ minimised if

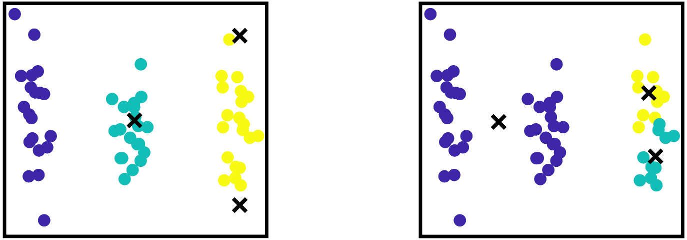
$$\text{point closest to } \mu_j^{(t+1)} \text{ moved from } C_j^{(t)} \text{ to } C_j^{(t+1)}$$

- From center computation step:

$$\sum_{i=1}^{k} \sum_{x \in C_j^{(t)}} \left\| x - \mu_j^{(t+1)} \right\|^2 \geq \sum_{i=1}^{k} \sum_{x \in C_j^{(t+1)}} \left\| x - \mu_j^{(t+1)} \right\|^2 = G\left(\boldsymbol{\mu}^{(t+1)}\right)$$

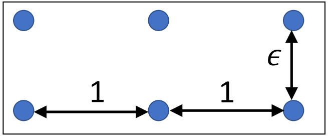
Sub-optimality of Lloyd's algorithm

- Lloyd's iterations always converge to a local optimum
- Can be arbitrarily worse than global optimum

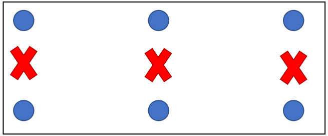


Sub-optimality of Lloyd's algorithm (Verify)

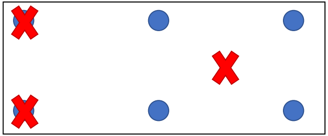
- Consider configuration of 6 points in $\mathbb{R}^2$
- Verify that optimal centers have k-means cost: $G_{opt} = 6\epsilon^2$ (assume $\epsilon \ll 1$)
- No updates if we initialise Lloyd's iterations with configuration on right
    - How does the cost compare to G_{opt}?



Six points



Optimal centers



Possible solution

k-means++

- Most popular practical implementation of k-means
- Idea:
  - Careful choice of centers (seeding)
  - Define clusters given by chosen centers
- Merits:
  - Not iterative; completes in $O(km)$-runtime
  - Comes with an approximation guarantee

k-means++ Algorithm

1. Pick $x \in \mathcal{X}$ uniformly at random and set $\widehat{\mu}_1 = x$
2. For $j = 2, \ldots, k$
   i. Define $w_i = \min\limits_{r \in \{1,\ldots,j-1\}} \|x_i - \widehat{\mu}_r\|^2$ for every $x_i \in \mathcal{X}$
   ii. Normalise weights $w_1, \ldots, w_m$ such that $\sum\limits_{i=1}^{m} w_i = 1$
   iii. Sample $x \in \mathcal{X}$ according to probabilities $w_1, \ldots, w_m$
   iv. Set $\widehat{\mu}_j = x$
3. Define $C_j = \{x \in \mathcal{X} \ : \ \mu_j \text{ is closest center for } x\}$

Approximation guarantee for k-means++

Theorem kmeans.2 (k-means++ approximation guarantee (Arthur & Vassilvitskii 2007))

- *Given $\mathcal{X}$ and k, let $G(\cdot) = k$-means cost, and optimal cost* $G_{opt} = \min\limits_{\boldsymbol{\mu} : |\boldsymbol{\mu}| \leq k} G(\boldsymbol{\mu})$
- $\widehat{\boldsymbol{\mu}} =$ *solution of k-means++*

$$\mathbb{E}\left[G(\widehat{\boldsymbol{\mu}})\right] \quad \leq \quad 8(\ln k + 2) G_{opt}$$

Expectation is with respect to randomness of the k-means++ algorithm

- Proof skipped. Will prove simpler cases $k = 1$ and $k = 2$
- A more complicated polynomial-time algorithm achieves $G(\widehat{\boldsymbol{\mu}}) \leq 6.357 \cdot G_{opt}$
- NP-Hard to find $\widehat{\boldsymbol{\mu}}$ such that $G(\widehat{\boldsymbol{\mu}}) \leq 1.0013 \cdot G_{opt}$ . . . there is a gap

k-means++ for $k = 1$

Lemma kmeans.3 (Squared distance of $\mathcal{X}$ from a random sample)

- *For any $\mathcal{X}$, denote $\mu = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} x$*

- *If $g(\mathcal{X}, z) = \sum_{x\in\mathcal{X}} \|x - z\|^2$, then*

$$g(\mathcal{X}, z) \quad = \quad g(\mathcal{X}, \mu) \; + \; |\mathcal{X}| \cdot \|z - \mu\|^2$$

- *Let $\widehat{\mu}_1 \sim Uniform(\mathcal{X})$* ... *first sample in k-means++*

$$\textit{Then} \qquad \mathbb{E}_{\widehat{\mu}_1}\left[g(\mathcal{X}, \widehat{\mu}_1)\right] \quad = \quad 2g(\mathcal{X}, \mu)$$

Proof: Part 1

- Prove using the relation:

$$\|x - z\|^2 = \|x - \mu\|^2 + \|z - \mu\|^2 - 2\langle x - \mu, z - \mu \rangle$$

- $\sum_{x \in \mathcal{X}} x - \mu = 0 \implies$ third term above is zero after summation

- Summing up

$$g(\mathcal{X}, z) = \sum_{x \in \mathcal{X}} \|x - z\|^2 = \sum_{x \in \mathcal{X}} \|x - \mu\|^2 + \underbrace{\sum_{x \in \mathcal{X}}}_{|\mathcal{X}| \text{ terms}} \|z - \mu\|^2$$

Proof: Part 2

- Note $\widehat{\mu}_1 \sim \text{Uniform}(\mathcal{X})$

$$\begin{aligned}
\mathbb{E}_{\widehat{\mu}_1}\left[g(\mathcal{X}, \widehat{\mu}_1)\right] &= \frac{1}{|\mathcal{X}|} \sum_{z \in \mathcal{X}} g(\mathcal{X}, z) \\
&= \frac{1}{|\mathcal{X}|} \sum_{z \in \mathcal{X}} \left(g(\mathcal{X}, \mu) + |\mathcal{X}| \cdot \|z - \mu\|^2\right) \\
&= g(\mathcal{X}, \mu) \ + \underbrace{\sum_{z \in \mathcal{X}} \|z - \mu\|^2}_{= g(\mathcal{X}, \mu)}
\end{aligned}$$

k-means++ for $k = 2$

Algorithm:

– Sample $\widehat{\mu}_1$ uniformly from $\mathcal{X}$

– Sample $\widehat{\mu}_2$ such that $\mathbb{P}(\widehat{\mu}_2 = z) \propto \|z - \widehat{\mu}_1\|^2$

Theorem kmeans.4 (Approximation guarantee of k-means++ for k=2)

- $\boldsymbol{\mu} = (\mu_1, \mu_2) =$ *optimal centers for* k*-means with clusters* C_1, C_2
- $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \widehat{\mu}_2) =$ *centers obtained by* k*-means++*

$$\textit{Then} \qquad \mathbb{E}_{\widehat{\mu}_1, \widehat{\mu}_2}\left[G(\widehat{\boldsymbol{\mu}})\right] \quad \leq \quad 8 \cdot \underbrace{G(\boldsymbol{\mu})}_{=G_{opt}}$$

Proof

- Recall distance to a set of centers

$$d(x, \widehat{\boldsymbol{\mu}}) = \min\left\{\|x - \widehat{\mu}_1\|\,,\ \|x - \widehat{\mu}_1\|\right\}$$

- Can write $G(\widehat{\boldsymbol{\mu}})$ as

$$\begin{aligned} G(\widehat{\boldsymbol{\mu}}) &= \sum_{x \in \mathcal{X}} d(x, \widehat{\boldsymbol{\mu}})^2 \\ &= \sum_{x \in C_1} d(x, \widehat{\boldsymbol{\mu}})^2 + \sum_{x \in C_2} d(x, \widehat{\boldsymbol{\mu}})^2 \end{aligned}$$

where C_1, C_2 are optimal clusters (name clusters such that $\widehat{\mu}_1 \in C_1$)

Proof: Conditioning on $\widehat{\mu}_1$

$$\mathbb{E}_{\widehat{\mu}_1,\widehat{\mu}_2}\left[G(\widehat{\boldsymbol{\mu}})\right] \quad = \quad \mathbb{E}_{\widehat{\mu}_1}\Big[\underbrace{\mathbb{E}_{\widehat{\mu}_2 \,|\, \widehat{\mu}_1}\left[G(\widehat{\boldsymbol{\mu}})\right]}_{\text{we compute this first}} \Big]$$

- Fix $\widehat{\mu}_1$ and compute conditional expectation

$$\mathbb{E}_{\widehat{\mu}_2 \,|\, \widehat{\mu}_1}\left[G(\widehat{\boldsymbol{\mu}})\right] \quad = \quad \underbrace{\mathbb{E}_{\widehat{\mu}_2 \,|\, \widehat{\mu}_1}\left[\sum_{x\in C_1} d(x,\widehat{\boldsymbol{\mu}})^2\right]}_{\text{term I}} + \underbrace{\mathbb{E}_{\widehat{\mu}_2 \,|\, \widehat{\mu}_1}\left[\sum_{x\in C_2} d(x,\widehat{\boldsymbol{\mu}})^2\right]}_{\text{term II}}$$

Proof: Bounding term I

- Bounding first term using $d(x, \widehat{\boldsymbol{\mu}}) \leq \|x - \widehat{\mu}_1\|$

$$\mathbb{E}_{\widehat{\mu}_2 \mid \widehat{\mu}_1}\left[\sum_{x \in C_1} d(x, \widehat{\boldsymbol{\mu}})^2\right] \leq \sum_{x \in C_1} \|x - \widehat{\mu}_1\|^2 = g(C_1, \widehat{\mu}_1)$$

- We know $\widehat{\mu}_1$ uniformly chosen, and $\widehat{\mu}_1 \in C_1$
  - So $\widehat{\mu}_1 \sim \text{Uniform}(C_1)$, and for $k = 1$, we saw

$$\mathbb{E}_{z \sim \text{Unif}(C)}[g(C, z)] = 2 \cdot g(C, \mu)$$

- But, we cannot say $\widehat{\mu}_2 \in C_2 \implies$ above trick does not work

Proof: Expanding term II

- Sampling of $\widehat{\mu}_2$ (D^2-sampling)

$$\mathbb{P}_{\widehat{\mu}_2 \mid \widehat{\mu}_1}\left(\widehat{\mu}_2 = z\right) = \frac{\|z - \widehat{\mu}_1\|^2}{\sum\limits_{y \in \mathcal{X}} \|y - \widehat{\mu}_1\|^2} = \frac{\|z - \widehat{\mu}_1\|^2}{g(\mathcal{X}, \widehat{\mu}_1)}$$

- For second term, compute conditional expectation

$$\mathbb{E}_{\widehat{\mu}_2 \mid \widehat{\mu}_1}\left[\sum_{x \in C_2} d(x, \widehat{\boldsymbol{\mu}})^2\right] = \sum_{z \in \mathcal{X}} \sum_{x \in C_2} \frac{\|z - \widehat{\mu}_1\|^2}{g(\mathcal{X}, \widehat{\mu}_1)} \cdot d\big(x, \{\widehat{\mu}_1, z\}\big)^2$$

  - Separately considers sums over $z \in C_1$ and $z \in C_2$

Proof: Bounding term II, summands for $z \in C_1$

- Summation over $z \in C_1$:

$$\sum_{z \in C_1} \sum_{x \in C_2} \frac{\|z - \widehat{\mu}_1\|^2}{g(\mathcal{X}, \widehat{\mu}_1)} \cdot d\big(x, \{\widehat{\mu}_1, z\}\big)^2$$

  - In the case, $\widehat{\mu}_2$ sampled from C_1 $\quad\Longrightarrow\quad$ Hard to show $\|x - \widehat{\mu}_2\|$ small for $x \in C_2$
  - We try to cancel $d(x, \widehat{\boldsymbol{\mu}})^2$ with the denominator

$$\sum_{x \in C_2} d(x, \widehat{\boldsymbol{\mu}})^2 \;\leq\; \sum_{x \in C_2} \|x - \widehat{\mu}_1\|^2 \;\leq\; g(C_2, \widehat{\mu}_1) \;\leq\; g(\mathcal{X}, \mu_1)$$

- Summation at top $\;\leq\; g(C_1, \widehat{\mu}_1)$

Proof: Bounding term II, summands for $z \in C_2$

- Summation over $z \in C_2$: $\displaystyle\sum_{z \in C_2} \sum_{x \in C_2} \frac{\|z - \widehat{\mu}_1\|^2}{g(\mathcal{X}, \widehat{\mu}_1)} \cdot d\big(x, \{\widehat{\mu}_1, z\}\big)^2$
    - In this case, $\widehat{\mu}_2$ sampled from C_2
    - But the term $\|z - \widehat{\mu}_1\|$ can be large since $z \in C_2$
- Claim $(\star)$: $\|z - \widehat{\mu}_1\|^2 \leq \dfrac{2}{|C_2|}\big(g(C_2, z) \; + \; g(C_2, \widehat{\mu}_1)\big)$

Proof: Continuing above assuming $(\star)$ holds

$$
\begin{aligned}
&\sum_{z\in C_2}\sum_{x\in C_2}\frac{\|z-\widehat{\mu}_1\|^2}{g(\mathcal{X},\widehat{\mu}_1)}\cdot d\big(x,\{\widehat{\mu}_1,z\}\big)^2 \\
&\quad\le \frac{2}{|C_2|}\sum_{z\in C_2}\sum_{x\in C_2}\left(\frac{g(C_2,z)}{g(\mathcal{X},\widehat{\mu}_1)} + \underbrace{\frac{g(C_2,\widehat{\mu}_1)}{g(\mathcal{X},\widehat{\mu}_1)}}_{\le 1}\right)\underbrace{d\big(x,\{\widehat{\mu}_1,z\}\big)^2}_{\le\min\{\|x-\widehat{\mu}_1\|^2,\|x-z\|^2\}} \\
&\quad\le \frac{2}{|C_2|}\sum_{z\in C_2}\sum_{x\in C_2}\frac{g(C_2,z)}{g(\mathcal{X},\widehat{\mu}_1)}\|x-\widehat{\mu}_1\|^2 + \frac{2}{|C_2|}\sum_{z\in C_2}\sum_{x\in C_2}\|x-z\|^2 \\
&\quad\le \frac{2}{|C_2|}\sum_{z\in C_2}g(C_2,z)\cdot\underbrace{\frac{g(C_2,\widehat{\mu}_1)}{g(\mathcal{X},\widehat{\mu}_1)}}_{\le 1} + \frac{2}{|C_2|}\sum_{z\in C_2}g(C_2,z) \\
&\quad\le \frac{4}{|C_2|}\sum_{z\in C_2}g(C_2,z) = 4\cdot\mathbb{E}_{z\sim\mathrm{Uniform}(C_2)}\left[g(C_2,z)\right] = \underbrace{8\cdot g(C_2,\mu_2)}_{\mu_2\text{ is center of }C_2}
\end{aligned}
$$

Proof of Claim $(\star)$

- Bounding $\|z - \widehat{\mu}_1\|$

$$\begin{aligned} \|z - \widehat{\mu}_1\|^2 &\leq \big(\|y - z\| \;+\; \|y - \widehat{\mu}_1\|\big)^2 && \ldots \text{ triangle inequality} \\ &\leq 2\big(\|y - z\|^2 \;+\; \|y - \widehat{\mu}_1\|^2\big) && \ldots\ (a+b)^2 \leq 2(a^2 + b^2) \end{aligned}$$

  - Above holds for every $y \in C_2$
  - We can take upper bound as average over all $y \in C_2$

$$\begin{aligned} \|z - \widehat{\mu}_1\|^2 &\leq \frac{2}{|C_2|} \sum_{y \in C_2} \big(\|y - z\|^2 \;+\; \|y - \widehat{\mu}_1\|^2\big) \\ &\leq \frac{2}{|C_2|} \big(g(C_2, z) \;+\; g(C_2, \widehat{\mu}_1)\big) \end{aligned}$$

Proof: Combining all previous steps

$$\mathbb{E}_{\widehat{\mu}_2 \mid \widehat{\mu}_1}\left[G(\widehat{\boldsymbol{\mu}})\right] = \underbrace{\mathbb{E}_{\widehat{\mu}_2 \mid \widehat{\mu}_1}\left[\sum_{x \in C_1} d(x, \widehat{\boldsymbol{\mu}})^2\right]}_{\leq g(C_1, \widehat{\mu}_1)} + \underbrace{\mathbb{E}_{\widehat{\mu}_2 \mid \widehat{\mu}_1}\left[\sum_{x \in C_2} d(x, \widehat{\boldsymbol{\mu}})^2\right]}_{\text{split into } \underbrace{\sum_{z \in C_1}}_{\leq g(C_1, \widehat{\mu}_1)} \text{ and } \underbrace{\sum_{z \in C_2}}_{\leq 8 g(C_2, \mu_2)}}$$

$$\leq 2 \cdot g(C_1, \widehat{\mu}_1) + 8 \cdot g(C_2, \mu_2)$$

Hence, $$\mathbb{E}_{\widehat{\boldsymbol{\mu}}}\left[G(\widehat{\boldsymbol{\mu}})\right] = \mathbb{E}_{\widehat{\mu}_1}\left[\mathbb{E}_{\widehat{\mu}_2 \mid \widehat{\mu}_1}\left[G(\widehat{\boldsymbol{\mu}})\right]\right] \leq \underbrace{2\, \mathbb{E}_{\widehat{\mu}_1}\left[g(C_1, \widehat{\mu}_1)\right]}_{= 2 g(C_1, \mu_1)} + 8 \cdot g(C_2, \mu_2)$$

$$\leq 8 \cdot \big(g(C_1, \mu_1) + g(C_2, \mu_2)\big) = 8 \cdot G(\boldsymbol{\mu})$$

Different questions about k-means problem

- k-means problem: $$\underset{\boldsymbol{\mu}\,:\,|\boldsymbol{\mu}|\le k}{\text{minimise}} \; G(\boldsymbol{\mu}) = \sum_{j=1}^{k} \sum_{x\in C_j} \|x-\mu_j\|^2$$

- Approximation guarantee:
  - Finding optimal k-means solution is NP-hard
  - Compare $G(\widehat{\boldsymbol{\mu}})$ for an efficient / poly-time algorithm to G_{opt} (mentioned earlier)

- Consistency:
  - Assume $x_1, \ldots, x_m \sim_{iid} \mathcal{D}$
  - What happens to G_{opt} or $G(\widehat{\boldsymbol{\mu}})$ as $m \to \infty$?

Consistency

- Assume $S = \{x_1, \ldots, x_m\} \sim \mathcal{D}^m$ . . . distribution only over x (no label)
- $G_{S,k}(\boldsymbol{\mu}) = k$-means cost on data S when the centers are $\boldsymbol{\mu}$
- Let $\boldsymbol{\mu}_S^* =$ centers corresponding to optimal k-means cost for S
- What happens to cost $G_{S,k}(\boldsymbol{\mu}_S^*)$ as $m \to \infty$?

Preparing for consistency theorem

- For $\boldsymbol{\mu} = (\mu_1, \ldots \mu_k)$,

$$G_{S,k}(\boldsymbol{\mu}) = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2$$

$$= \sum_{x \in S} d(x, \boldsymbol{\mu})^2 \qquad \ldots\ d(x, \boldsymbol{\mu}) = \min_{\mu_j \in \boldsymbol{\mu}} \|x - \mu_j\|$$

- Let $x_1, \ldots, x_m \sim_{iid} \mathcal{D}$, and fix $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$

- What happens to $\frac{1}{m} G_{S,k}(\boldsymbol{\mu})$ as $m \to \infty$? (law of large numbers)

$$\frac{1}{m} \cdot G_{S,k}(\boldsymbol{\mu}) = \underbrace{\frac{1}{m} \sum_{i=1}^{m} d(x_i, \boldsymbol{\mu})^2}_{\text{avg of independent terms}} \xrightarrow{m \to \infty} \mathbb{E}_{x \sim \mathcal{D}} \left[\min_{\mu \in \boldsymbol{\mu}} \|x - \mu\|^2 \right]$$

Consistency of k-means

Theorem kmeans.5 (Strong consistency of k-means (Pollard, 1981))

Let $S = \{x_1, \ldots, x_m\} \sim \mathcal{D}^m$ *and* $\boldsymbol{\mu}_S^*$ *denote centers corresponding to optimal* k*-means cost.*

$$\lim_{m\to\infty} \left[\frac{1}{m} \cdot G_{S,k}(\boldsymbol{\mu}_S^*) \right] = \min_{\boldsymbol{\mu}\,:\,|\boldsymbol{\mu}|\leq k} \mathbb{E}_{x\sim\mathcal{D}} \left[\min_{\mu\in\boldsymbol{\mu}} \|x-\mu\|^2 \right] \qquad \textit{with probability } 1$$

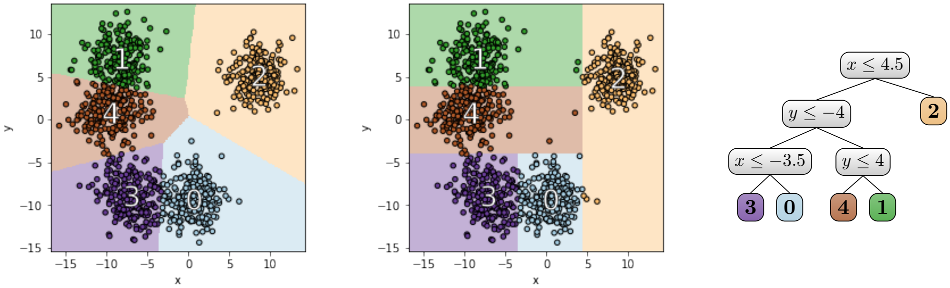
- One key idea of proof: Let $\mathcal{H} = \{\boldsymbol{\mu} \mid \boldsymbol{\mu} \text{ contains at most } k \text{ distinct centers}\}$

$$\sup_{\boldsymbol{\mu}\in\mathcal{H}} \left| \frac{1}{m} \cdot G_{S,k}(\boldsymbol{\mu}) - \mathbb{E}_{x\sim\mathcal{D}} \left[\min_{\mu\in\boldsymbol{\mu}} \|x-\mu\|^2 \right] \right| \to 0 \text{ as } m \to \infty$$

- Pollard (1982) show central limit theorem for $\boldsymbol{\mu}_S^*$ ... multivariate normal dist.

Explainable approximation of k-means

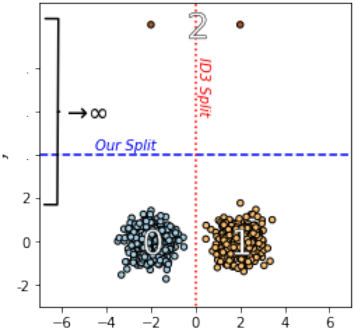
- k-means produces simple (linear) decision boundaries
- Linear boundaries less interpretable if they are not axis-aligned
- If k-means solution is approximated by a decision tree, then clusters are interpretable



(a) Optimal 5-means clusters (b) Tree based 5-means clusters (c) Threshold tree

Naïve approach

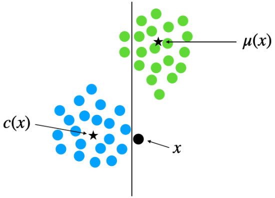
- Perform k-means to obtain clusters.
  Label the clusters.
- Learn decision tree with above defined labels
  - Decision learners greedily find axis-aligned cuts
  - Example: Iterative Dichotomiser 3 (ID3), at each step, splits along feature to maximally reduce entropy
- Issue: Splits may form clusters with very high k-means cost
  - ID3: Maximum reduction in entropy if large clusters are separated first



IMM (Dasgupta et al, ICML 2020)
Finds split where fewer points get separated from their k-means centers

Iterative Mistake Minimisation (IMM; Dasgupta et al, ICML 2020)

- Let $\mu_1, \ldots, \mu_k$ be the k centres from k-means
  - Every point x is associated with one of the k centres
  - Initially every x associated with its k-mean centre $c(x)$



- **Mistake** happens if point x is separated from $c(x)$ and assigned to different centre $\mu(x)$ due to axis-aligned split

- IMM: Builds tree iteratively
  - Each iteration separates two centres μ_i, μ_j
  - Find axis and cut that leads to minimum mistakes
  - k iterations result in k leaves / clusters, each with one centre $\ldots O(kpn)$ runtime

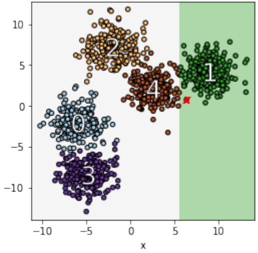
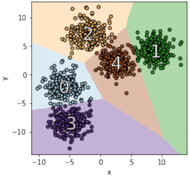
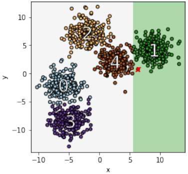


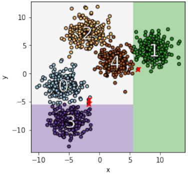
Illustration of IMM



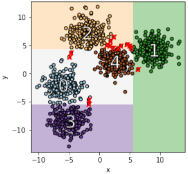
(a) Optimal 5-means clusters



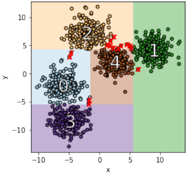
(b) 1st split: 1 mistake caused by this split (1 total mistake)



(c) 2nd split: 2 mistakes caused by this split (3 total mistakes)



(d) 3rd split: 12 mistakes caused by this split (15 total mistakes)



(e) 4th split: 0 mistakes caused by this split (15 total mistakes)

Figure 3: Figure 3(a) presents the optimal 5-means clustering. Figures 3(b)–3(e) depict the four splits of the IMM algorithm. The first split separates between cluster 1 and the rest, with a single mistake (marked as a red cross). Next, the IMM separates cluster 3 with 2 additional mistakes. The third split separates cluster 2, and this time the minimal number of mistakes is 12 for this split. Eventually, clusters 0 and 4 are separated

Approximation guarantee for IMM

Theorem kmeans.6 (Approximation guarantee for IMM)

Let $C_1, \ldots, C_k$ be clusters returned by k-means with cost $G(C_1, \ldots, C_k)$

IMM returns a threshold tree with k leaves corresponding to k clusters $\widehat{C}_1, \ldots, \widehat{C}_k$ such that

$$G(\widehat{C}_1, \ldots, \widehat{C}_k) = O(k^2) \cdot G(C_1, \ldots, C_k)$$

- Better upper bound: There is a randomised algorithm that returns clusters $\widehat{C}_1, \ldots, \widehat{C}_k$ with expected cost [Gupta et al., arXiv:2304.09743]
$$\mathbb{E}[G(\widehat{C}_1, \ldots, \widehat{C}_k)] = O(k \cdot \ln\ln k) \cdot G(C_1, \ldots, C_k)$$
- Lower bound: There is a set of points for which best threshold tree has cost
$$\Omega(k) \cdot G(C_1, \ldots, C_k)$$
[Gamlath et al., NeurIPS 2021]

Proof of approximation guarantee for IMM

- For every point x, define $c(x), \mu(x) \in \{\mu_1, \ldots, \mu_k\}$ as
  - $\mu(x)$ is centre that lies in same leaf of threshold tree as x
  - $c(x)$ is centre assigned by k-means (if $c(x) \neq \mu(x)$, then mistake happened as some node)
- Let T be threshold tree, U denotes an internal node in T

$$\begin{aligned}
G(\widehat{C}_1, \ldots, \widehat{C}_k) &\leq \sum_{j=1}^{k} \sum_{x \in \widehat{C}_j} \|x - \mu(x)\|^2 && \ldots \mu(x) \text{ is not mean of } \widehat{C}_j \\
&\leq \sum_{j=1}^{k} \sum_{x \in \widehat{C}_j} 2\|x - c(x)\|^2 + 2\|c(x) - \mu(x)\|^2 && \ldots \|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2) \\
&= 2G(C_1, \ldots, C_k) + 2 \sum_{U \in T} \sum_{\substack{x \text{ gets separated from} \\ c(x) \text{ in split at } U}} \|c(x) - \mu(x)\|^2
\end{aligned}$$

Proof (contd.)

- Key quantities:
  - $\boldsymbol{\mu}_U$ = set of centres that reach node U
  - $\mathcal{X}_U^{cor}$ = set of points x that reach U along with their their true centres $c(x)$ (that is, $c(x)$ lies in region defined by U)
  - t_U = #mistakes at U, that is, all $x \in \mathcal{X}_U^{cor}$ that get separated from $c(x)$ due to split at U
- Observe $\displaystyle\sum_{\substack{x \text{ gets separated from} \\ c(x) \text{ in split at } U}} \|c(x) - \mu(x)\|^2 \leq t_U \cdot \max_{a,b \in \boldsymbol{\mu}_U} \|a - b\|^2$

Proof (contd.)

- Let a^i denote i-th coordinate of $a \in \mathbb{R}^p$

  We will show: $t_U \cdot \max_{a,b \in \boldsymbol{\mu}_U} (a^i - b^i)^2 \;\leq\; 4k \cdot \sum_{x \in \mathcal{X}_U^{cor}} (x^i - c^i(x))^2$ $\qquad \ldots (\star)$

- Proof assuming above result:

$$
\begin{aligned}
G(\widehat{C}_1, \ldots, \widehat{C}_k) &\leq 2G(C_1, \ldots, C_k) + 2 \sum_{U \in T} t_U \cdot \max_{a,b \in \boldsymbol{\mu}_U} \|a - b\|^2 \\
&\leq 2G(C_1, \ldots, C_k) + 2 \sum_{U \in T} \sum_{i=1}^{p} t_U \cdot \max_{a,b \in \boldsymbol{\mu}_U} (a^i - b^i)^2 \\
&\leq 2G(C_1, \ldots, C_k) + 2 \sum_{U \in T} 4k \cdot \underbrace{\sum_{x \in \mathcal{X}_U^{cor}} \|x - c(x)\|^2}_{\leq G(C_1, \ldots, C_k)} \\
&\leq 2G(C_1, \ldots, C_k) + 8k^2 G(C_1, \ldots, C_k) \qquad \ldots \text{there are } k \text{ internal nodes}
\end{aligned}
$$

Proof of bound in ($\star$): First part

- Let k' centers remain at node U, and their i-th coordinates be $z_1 \le z_2 \le \ldots \le z_{k'}$

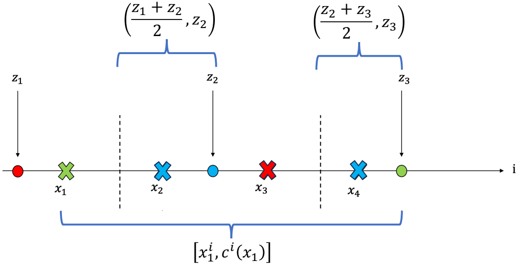
- Consider thresholds mid-way between two consecutive centers
  $\theta_j = \frac{z_{j-1}+z_j}{2}, j = 2, \ldots, k'$

- **Claim:** For every j, the split $x^i \gtrless \theta_j$
  makes at least t_U mistakes **(why?)**

- If the splits at $\theta_j, \ldots, \theta_{j'}$
  all separate x from $c(x)$, then

$$\left|x^i - c^i(x)\right| \ge \sum_{a=j}^{j'} \frac{z_a - z_{a-1}}{2}$$

$$\implies \quad \left(x^i - c^i(x)\right)^2 \ge \sum_{a=j}^{j'} \left(\frac{z_a - z_{a-1}}{2}\right)^2$$



Proof of bound in $(\star)$: Final part

$$\sum_{x \in \mathcal{X}_U^{cor}} \left(x^i - c^i(x)\right)^2 \geq \sum_{x \in \mathcal{X}_U^{cor}} \frac{1}{4} \sum_{\substack{a\,:\,\text{split at } \theta_a \\ \text{separates } x, c(x)}} \left(z_a - z_{a-1}\right)^2$$

$$= \frac{1}{4} \sum_{j=2}^{k'} \left(z_j - z_{j-1}\right)^2 \times \underbrace{\#(x, c(x))\text{-pairs separated by split at } \theta_j}_{\geq t_U}$$

$$\geq \frac{t_U}{4} \sum_{j=2}^{k'} \left(z_j - z_{j-1}\right)^2$$

$$\geq \frac{t_U}{4} \cdot \frac{1}{k'} \underbrace{\left(\sum_{j=2}^{k'} z_j - z_{j-1}\right)^2}_{=(z_{k'} - z_1)^2 = \max_{a,b \in \boldsymbol{\mu}_U} (a^i - b^i)^2} \qquad \ldots \text{using } \underbrace{\left(\sum_{i=1}^{n} a_i\right)^2 \leq n \cdot \sum_{i=1}^{n} a_i^2}_{\text{by Cauchy-Schwarz inequality}}$$