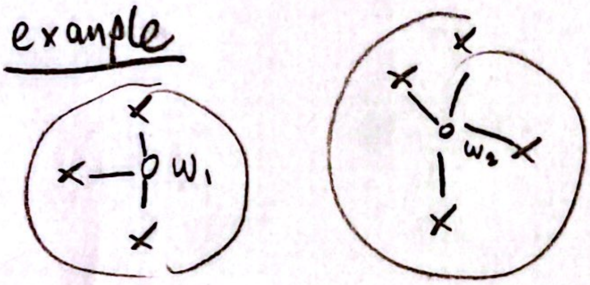
On k-means

$$L(w_1, \dots, w_k) = \sum_{i=1}^{n} \min_{j \in [k]} \|x_i - w_j\|^2$$

example



C_1 C_2

cluster centers
induce
partition of points

k-means is really a combinatorial problem over all n^k partitions of the n points into k groups

Fix a partition $C_1, \dots, C_k$. Then automatically

$$w_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

we induce a partition

Equivalent (relaxed) problem

$$\min_{\substack{W \subseteq \mathbb{R}^d \\ |W| = k}} \sum_{i=1}^{n} \min_{w \in W} \|x_i - w\|^2 \longleftrightarrow \min_{\bigsqcup_{i=1}^{k} C_i = X} \sum_{i=1}^{k} \frac{1}{|C_j|} \sum_{x,y \in C_j} \|x-y\|^2$$

$|W| = k$

Note that: $\mathbb{E}(\|X-Y\|^2) = 2\mathbb{E}(X^2)$ for X, Y iid! ②

$$\frac{1}{|C|^2} \sum_{x,y \in C} \|x-y\|^2 = \frac{1}{|C|} \sum_{x \in C} \|x - \mu_C\|^2$$

<u>Pf</u>: $\mathbb{E}(X^2 - 2XY + Y^2) = 2\mathbb{E}(X^2) - 2\cancel{\mathbb{E}(XY)}$ independent $\Rightarrow$ uncorrelated

This is just uniform distributions. □

The natural "k-means algorithm"

Lloyd's method

① Randomly initialize centre $w_1, .., w_k$

corresponding clusters $C_1, .., C_k$

② Let $w_i \leftarrow \frac{1}{|C_i|} \sum_{c \in C_i} c$

③ Repeat until w_i's no longer change

<u>FACT 1</u>: monotone decreasing loss (given any partition, the optimum is the mean — minimizer)

<u>FACT 2</u>: unbounded approximation ratio

Re: **FACT 1**

$$\min_{x \in \mathbb{R}^d} \sum_{s \in S} \|x - s\|^2 = \frac{1}{|S|} \sum_{s \in S} s$$

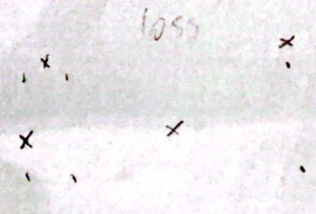
Show it by differentiation: $\frac{\partial L}{\partial x} = 0$

$$\sum_{s \in S} x - s = 0$$

$$|S| \cdot x - \sum_{s \in S} s = 0 \qquad \boxed{x = \frac{1}{|S|} \sum_{s \in S} s}$$

Variation on this argument shows that Lloyd's algorithm is effectively a gradient-based technique (Newton's method)

Re fact 2 $\quad \frac{\text{cost(Lloyd)}}{\text{cost(OPT)}} \to \infty$



Other facts

$(9+\varepsilon)$-approximation (Kanungo et al)

$(1+\varepsilon)$ but exponential dependence on k or d (Matousek)

NP-hard (Dasgupta)

but 1D easy using DP

A simple way to improve k-means: (4)

probabilistic farthest-first choice of centers

(1) Pick w_1 uniformly from $x_1, \dots, x_n$ $W = \{w_1\}$

(2) For $i \in \{2, \dots, k\}$

KMEANS++

pick $x \in X$ w.p.

$$\frac{\mathrm{DIST}^2(W, x)}{\sum_{z \in X} \mathrm{DIST}^2(W, z)} = \frac{D^2(x; W)}{\sum_z D^2(z; W)}$$

$W \leftarrow x$
append.

Thm Let W_{++} be output of kmeans++

Then $\mathbb{E}(\phi_{W_{++}}(X)) \leq O(\log(k))\, \phi_{OPT}(X)$
(loss)

(in fact, $O(1)$ for nice settings)

Scanned with CamScanner

For sake of time, omitting full proof.
But we do two lemmas.

$\boxed{\text{Lemma 1}}$ Let A be cluster from C_{OPT}.

$c \sim U(A)$.

$$\mathbb{E}_c\left[\phi_{\{c\}}(A)\right] = 2\,\phi_{OPT}(A)$$

uniform choice

$\boxed{\text{Pf}}$ $$\sum_{c \in A} \mathbb{P}(\text{pick } c) \cdot \phi_{\{c\}}(A) = \sum_{c \in A} \frac{1}{|A|} \sum_{v \in A} \|c-v\|^2$$

$$= \frac{1}{|A|} \sum_{c,v \in A} \|c-v\|^2$$

$\left(\mathbb{E}(|X-Y|^2) = 2\mathbb{E}(X^2)\right)$ $$= 2 \sum_{v \in A} \|v-\mu_A\|^2$$

$\left(\text{mean is L2 minimizer}\right)$ $$= 2\,\phi_{OPT}(A)$$

$\boxed{\text{Lemma 4}}$ Pick $A \sim C_{OPT}$ arbitrary.
(cluster in optimal)

Let C be arbitrary set of cluster centers.

Pick $c \sim A$ according to <u>farthest-first weighting</u>

Then $\mathbb{E}_c\left(\phi_{C \cup \{c\}}(A)\right) \leq 8 \cdot \phi_{OPT}(A)$

$\boxed{\text{Pf}}$

$$\mathbb{E}_c\left(\phi_{C\cup\{c\}}(A)\right) = \sum_{c \in A} \mathbb{P}(\text{we pick } c) \cdot \phi_{C\cup\{c\}}(A)$$

$$= \sum_{c\in A} \frac{D^2(c;C)}{\sum_{a\in A} D^2(a;C)} \cdot \sum_{a \in A} \underbrace{D^2(a; C \cup \{c\})}_{\min_{b \in C\cup\{c\}} \|b-a\|^2}$$

$$= \sum_{c\in A} \frac{D^2(c;C)}{\sum_{a\in A} D^2(a;C)} \sum_{a\in A} \min\left(D(a;C)^2, \|a-c\|^2\right)$$

$$D(c;C) \underset{\uparrow}{\leq} D(a;C) + \|a-c\| \underset{\uparrow}{\leq} 2D(a;C)^2 + \|a-c\|^2$$

Δ Inequality   power-mean

$$\Rightarrow |A| \cdot D(c) \leq 2\sum_{a\in A} D(a;C)^2 + 2\sum_{a\in A}\|a-c\|^2$$

$$\leq \frac{2}{|A|}\sum_{c\in A} \frac{\sum_{a\in A} D^2(a)}{\sum_{a\in A} D^2(a)} \sum_{a\in A} \min(\,''\,)$$

$$+ \frac{2}{|A|}\sum_{c\in A} \frac{\sum_{a\in A}\|a-c\|^2}{\sum_{a\in A} D^2(a)} \sum_{a\in A} \min(\,''\,) \leq \frac{4}{|A|}\sum_{a,c\in A}\|a-c\|^2$$

$$= 8\,\phi_{OPT}(A) \quad \square$$