k-Means Clustering

Given a set of observations $(x_1, x_2 \dots x_n)$ where each observation is a d-dimensional real-vector, k-means clustering aims to partition the n observations into k sets $S = \{S_1, S_2 \dots S_k\}$ that minimize the within cluster sum of squares.

Formally, we have to find

$$\min \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2 = \min \sum_{i=1}^{k} |S_i| \, \text{variance}(S_i)$$

where μ_i is the mean of the points in S_i

this is same as

$$\min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x, y \in S_i} |x - y|^2$$

because of the identity $\sum_{x \in S_i} |x - \mu_i|^2 = \sum_{x \neq y \in S_i} (x - \mu_i)(\mu_i - y)$

↳ This problem is NP-hard. Some heuristics based algorithms are known (that dont guarantee optimality)

Standard algorithm / naïve k-means / Lloyd's algo

Given an initial set of k-means $m_1, m_2 \dots m_k$, we iteratively perform 2-steps:-

Assignment Step → Assign each observation to the cluster with the nearest mean

$S_i^{(t)} = \{ x_p : |x_p - m_i|^2 \leq |x_p - m_j|^2 \; \forall j, 1 \leq j \leq k \}$

each x_p is assigned to exactly one S even if it could be assigned to 2 or more.

<u>Update step</u> → Recalculate means for observations assigned to each cluster.

$$m_i := \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j$$

<u>Initialization methods</u> → Forgy method → randomly choose k observations & use them as initial means.

Random Partition → Randomly assign a cluster to each observation & then proceed to update step. Take their mean as the mean points basically.

<u>Complexity</u> → $O(nkdi)$ where d is no. of dimensions & i is no. of iterations. i is usually small resulting in a linear time algorithm but is $2^{-\Omega(\sqrt{n})}$ in the worst case.

2 problems with k-means algo :-

1. worst-case running time is super polynomial.
2. approximation found can be arbitrarily bad.

K-means ++ → Solves second problem of K-means. Gives a guaranteed approximation ratio of $O(\log k)$ in expectation. It uses a different approach to select initial centres & then applies k-means iterations. Algorithm is this →

1. Choose one center uniformly at random among the data points
2. For each data point x not chosen yet, compute D(x), the distance b/w x & nearest centre to it.
3. Choose a new data point at random as a new center, using a weighted probability proportional to $D(x)^2$
4. Repeat 2 & 3 k-1 times
5. Proceed using standard k-means.

Explanation of K-means → In both assignment & update step, the objective function

$$\phi = \sum_{i=1}^{k} \sum_{x \in S_i} |x - c_i|^2$$

always decreases.

For assignment step, this is obvious, for update step, we use the fact that variance is minimized by using the mean as the centre. This is proven by the property,

$$\underline{\sum_{x \in S} |x - z|^2 - \sum_{x \in S} |x - CM(S)|^2 = |S| \cdot |CM(S) - z|^2}$$

Hence k-means algo just finds a local minima by iteration. Since, no ~~config~~ selection of clusters is ever repeated, this gives a naive bound of $O(k^n)$ on the number of iterations.

Competition ratio of k-means++ →

Theorem → $E[\phi] \leq 8(\ln k + 2)\phi_{OPT}$ just after the seeding, since k-means only decreases ϕ, it holds true for the full algo.

C_{opt} denotes optimal clustering centres, $\phi(A) = \sum_{x \in A} \min_{c \in C} |x-c|^2$, $C(A)$, centre of mass of A

Lemma 1 → Let A be an arbitrary cluster in C_{OPT} & let C be the clustering with just one center, chosen uniformly at random from A.

then $\underline{E[\phi(A)] = 2\phi_{OPT}(A)}$.

Proof → $$E[\phi(A)] = \frac{1}{|A|}\sum_{a_0 \in A}\sum_{a \in A}|a-a_0|^2 \quad \text{(by definition of } E\text{)}$$

$$= \frac{1}{|A|}\sum_{a_0 \in A}\left(\sum_{a \in A}|a-C(A)|^2 + |A| \cdot |a_0 - C(A)|^2\right)$$

$$= 2\sum_{a \in A}|a - C(A)|^2$$

Since A is a cluster of $C_{OPT} \Rightarrow \phi_{OPT}(A) = \sum_{a \in A}|a - C(A)|^2$

Hence proved.

Lemma 2 → Let A be an arbitrary cluster in C_{OPT} & let C be an arbitrary clustering. If we add a random center to C from A, chosen with D^2 weighting, then $E[\phi(A)] \leq 8\phi_{OPT}(A)$

Proof → Prob. of choosing a_0 as the center given that we are choosing from A = $\dfrac{D(a_0)^2}{\sum_{a \in A} D(a)^2}$

After choosing center a_0, each point will contribute exactly $\min(D(a), |a-a_0|)^2$ to the potential.

$$\Rightarrow \quad \mathcal{E}[\phi(A)] = \sum_{a_0 \in A} \frac{D(a_0)^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} \min(D(a), |a-a_0|)^2$$

$$D(a_0) \leq D(a) + |a-a_0| \qquad \text{(triangle inequality)}$$

$$D(a_0)^2 \leq 2D(a)^2 + 2|a-a_0|^2 \qquad \text{(power inequality)}$$

Sum over a_z to get $\quad D(a_0)^2 \leq \frac{2}{|A|} \sum_{a \in A} D(a)^2 + \frac{2}{|A|} \sum_{a \in A} |a-a_0|^2$

$$\Rightarrow \quad \mathcal{E}[\phi(A)] \leq \frac{2}{|A|} \sum_{a_0 \in A} \frac{\sum_{a \in A} D(a)^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} \min(D(a), |a-a_0|^2) + \frac{2}{|A|} \sum_{a_0 \in A} \frac{\sum_{a \in A} |a-a_0|^2}{\sum_{a \in A} D(a)^2} \cdot \sum_{a \in A} \min(D(a), |a-a_0|^2)$$

$$\mathcal{E}[\phi(A)] \leq \frac{4}{|A|} \sum_{a_0 \in A} \sum_{a \in A} |a-a_0|^2 = 8\phi_{OPT}(A)$$

Proved.

Lemma 3 → Let C be an arbitrary clustering. Choose $u > 0$ "uncovered" clusters from C_{OPT} & let X_u denote the set of points in these clusters. Also, let $X_c = X - X_u$. Now suppose we add $t \leq u$ random centers C chosen with D^2 weighting. Let C' denote the resulting clustering & let ϕ' denote the corresponding potential. Then,

$$\mathcal{E}[\phi'] \le \left(\phi(X_c) + 8\phi_{OPT}(X_u)\right)\cdot(1+H_t) + \frac{u-t}{u}\cdot\phi(X_u)$$

$$H_t = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} \cdots \frac{1}{t}$$

Proof → by induction on ~~$(t-1, u)$~~ $(t-1, u)$ & $(t-1, u-1)$ to show (t, u) is true with base case $t=0, u>0$ & $t=u=1$.

Finally, put $u = t = k-1$ in above lemma. Let A be the C_{OPT} cluster that the first center belonged to. Then,

$$\mathcal{E}[\phi] \le \left(\phi(A) + 8\phi_{OPT} - 8\phi_{OPT}(A)\right)(1 + H_{k-1})$$

Since $H_{k-1} \le 1 + \ln k$ & $\mathcal{E}[\phi(A)] \le 2\phi_{OPT}(A)$

$$\Rightarrow \boxed{\mathcal{E}[\phi] \le 8(2 + \ln k)\phi_{OPT}}$$

Hence proved

* This result is tight.