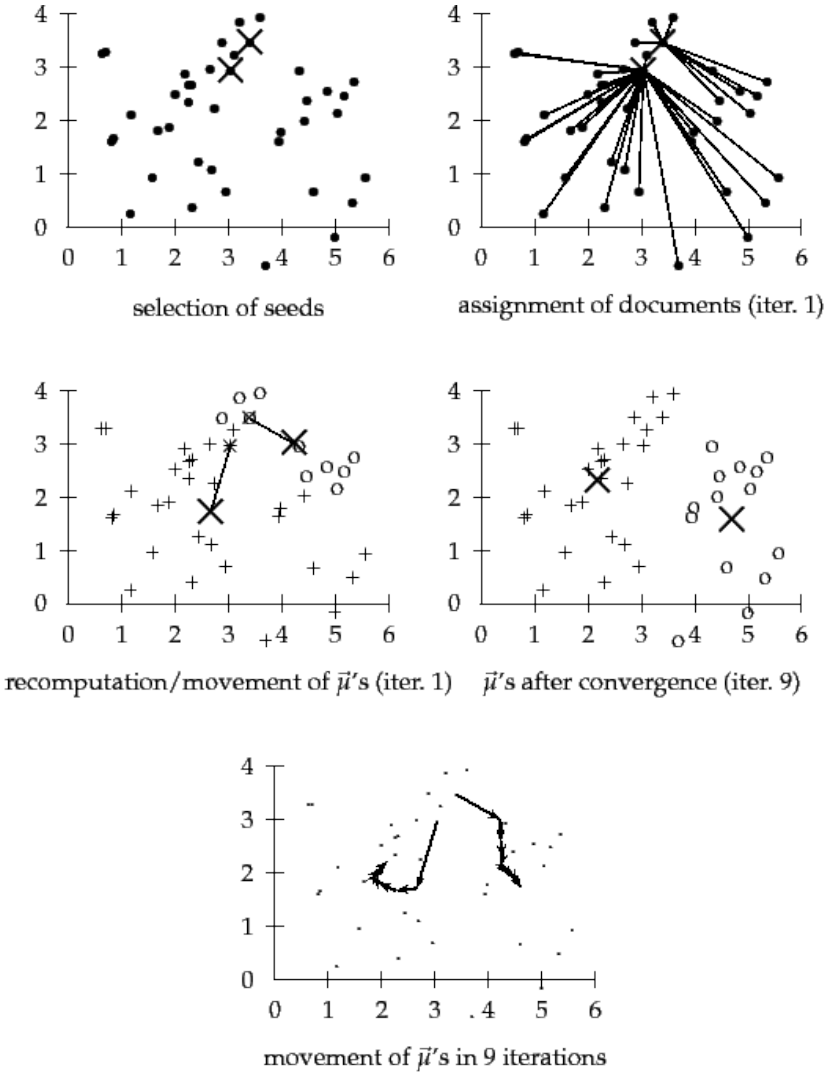
K-Means Plus Plus

Pragat Wagle

October 2021

1 Kmeans++ Algorithm

The k-means algorithm is used to partition a given set of observations into a predefined amount of k clusters. The algorithm as described by [?] starts with a random set of k center-points (μ). During each update step, all observations x are assigned to their nearest center-point (see equation 1). In the standard algorithm, only one assignment to one center is possible. If multiple centers have the same distance to the observation, a random one would be chosen.



► **Figure 16.3** A K-means example for $K = 2$ in $\mathbb{R}^2$. The position of the two centroids ($\vec{\mu}$'s shown as X's in the top four panels) converges after nine iterations.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - \mu_i^{(t)} \right\|^2 \leq \left\| x_p - \mu_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\} \tag{1}$$

This figure was pulled from https://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html.

The first plot displays the selection of the seeds or centers, which in the case of kmeans ++ takes one center uniformly at random from all the data points. Using this center computes the sum of squared distances of all the data points using the euclidean distance and then chooses new centers $x \in X$ such that $\frac{D(x')2}{\sum_{x \in X} D(x')2}$ where the denominator is the sum of all squared distance for the data points. The the computed center is used to recompute the new centers until convergence. It can be seen in the last plot the movement of the centers through 9 iterations.

1.1 Steps

1) Choose an initial uniformly center at random, the next center is chosen with $\frac{D(x')2}{\sum_{x \in X} D(x')2}$. Then repeat until we have k centers and then proceed as we normally would with k-means algorithm. This is the process of choose the k centers.

1a. Take one center c1, chosen uniformly at random from X .
1b. Take a new center ci $\frac{D(x')2}{\sum_{x \in X} D(x')2}$ where D(x) denotes the shortest distance from the data point x to its closest, already chosen centroid.
1c. Repeat Step 1b. until we have taken k centers altogether.
2-4. Proceed as with the standard k-means algorithm

2 Theorem 3.1

COPT denotes the optimal clustering and OPT the corresponding potential. Given a clustering C with potential , we also let (A) denote the contribution of A X to the potential (i.e., (A) = P aA mincCkx ck 2). If constructed with k-means++ the potential function ϕ satisfies

E| ϕ |$\leq$ 8(ln k + 2)$_{\phi \mathrm{OPT}}$

The chosen centers will satisfy this function where k is the number of clusters.

3 Lemma 3.2 and Lemma 3.3

Let A be an arbitrary cluster in COPT, and let C be the clustering with just one center, which is chosen uniformly at random from A. Then, $E[\phi(A)] = 2\phi_{OPT}(A)$.

A is a arbitrarily chosen cluster from COPT which are optimal clusterings. The c(A) below represents the center of mass for A and the result of equation corresponds to the corresponding potential for that calculated center of mass.

E| ϕA |= $2\sum_{a \in A}$ ||a-c(A)||^2.

If we add a center to C from A, arbitrarily chosen cluster with D2 weighting, then $E \mid \phi A \mid \leq 8_{\phi OPT}(A)$

The probability that we choose a fixed center with D2 weighting is $\frac{D(x')2}{\sum_{x \in X} D(x')2}$.

With a competitive seeding technique error in general at most is O(log k).

3.1 Lemma 3.4

Let C be an arbitrary clustering. Choose u $>$ 0 "uncovered" clusters from COPT, and let Xu denote the set of points in these clusters. Also let Xc = X - Xu. Now suppose we add t $\leq$ u random centers to C, chosen with D2 weighting. Let C' denote the the resulting clustering, and let ϕ' denote the corresponding potential. Then,

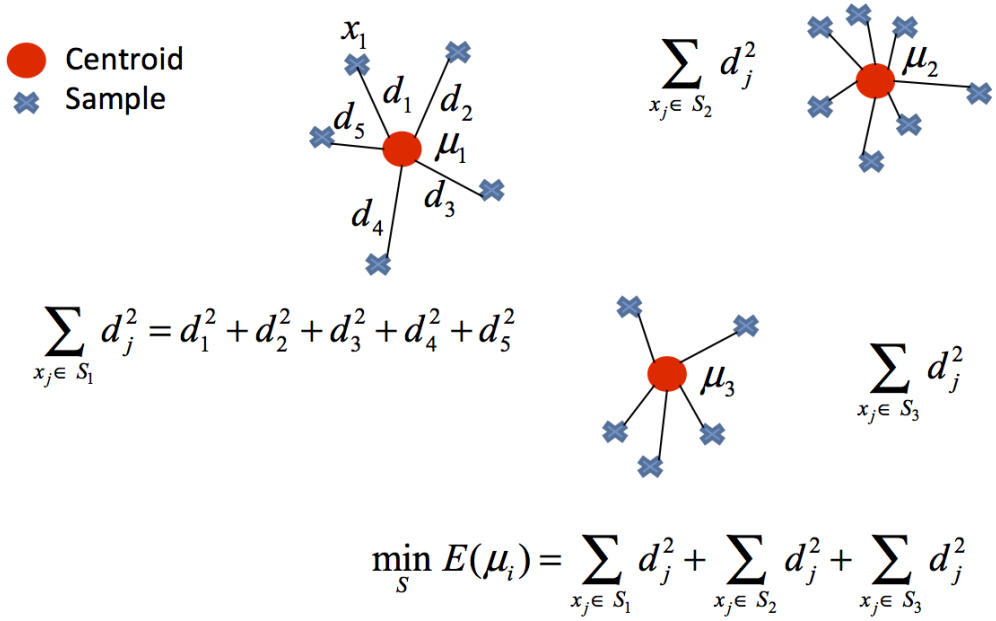
$$E \mid \phi A \mid \leq (\phi(Xc) + 8_{\phi OPT}(Xc) * (1 + H_{\mathrm{t}}) + \frac{u-t}{u} * \phi_{(}X_{\mathrm{u}}) \qquad (2)$$

where $\mathrm{H_t}$ denots the harmonic sum of $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{t}$

The idea here is that given u clusters chose from our optimal cloisterings. All the data points X - $\mathrm{X_u}$ is equal to $\mathrm{X_c}$, where C' denotes the arbitrary clustering of C with t random centers denoted by C'. This is used to meet the above lemma.

3.2 Summary

K means builds groups where the sum of the distances of the objects to its centroid is minimized within each group k, on each centroid update, from the mathematical point of view, we impose the extreme (minimum, in this case) necessary condition to the function. Once we find the minimum we know the center or point that gives the minimum potential or sum of squared distance to the our points.



The link to this figure is at https://www.unioviedo.es/compnum/labs/new/kmeans.html