Seeded Center Initialization In k-Means Clustering

Masum Billal BILLALMASUM93@GMAIL.COM
SENIOR DATA SCIENTIST
Data Science Department
Shohoz
Bangladesh

Editor: To be submitted

Abstract

k-Means is a popular clustering algorithm that reduces sum of minimum distances from data points to the centers. Now a days most of the times seeded centers are used instead of choosing all uniformly at random. k-Means++ chooses the centers with a so called D^2 weighting. In this paper, we explore various dimensions of clustering. We introduce a way to choose the first center with probability instead of choosing uniformly at random for k-Means++, two new ways of initializing centers like k-Means++. Empirical evidence shows that our proposed algorithms perform better than available methods of center initialization and that our way of choosing first center for k-Means++ consistently performs better than original k-means++. We also explore various properties regarding a set of points that helps us understand how the centers should be chosen. These properties help us analyze the expected value of inertia function as well.

1. Introduction

Widely regarded as the most popular clustering techniques, k-Means remains a humble interesting topic in machine learning as well as computational geometry. Roughly the problem is: given a set of points $\mathbf{X}$ in $\mathbb{R}^d$. Find a set of centers $\mathcal{C}$ such that the function *inertia*

$$\mathcal{I} = \sum_{\mathbf{x} \in \mathbf{X}} \min_{\mathbf{c} \in \mathcal{C}} (\|\mathbf{c} - \mathbf{x}\|^2)$$

is minimum where $\|\cdot\|$ is the L_2 norm[1].

Default k-Means algorithm starts with random centers and then converge based on minimum distances of the centers from the data points. New centers are calculated based on the centroid. This is known as Lloyd's algorithm (Lloyd, 1982). We repeat this process until no more change is possible. Ostrovsky et al. (2006) and Arthur et al. (2007) take it one step further by choosing the initial centers carefully. Only the first one or two centers are chosen uniformly at random. Then the rest of the centers are chosen with a probability. We intend to introduce other ways of initialization.

1. $\|c - x\|$ or L_2 norm of $\mathbf{c} - \mathbf{x}$ is the distance between the center $\mathbf{c}$ and point $\mathbf{x}$ or the magnitude of the vector $\mathbf{c} - \mathbf{x}$.

2. Related Work

There has been multiple surveys on k-Means in the literature. Probably the most relevant work in this regard is done by Celebi et al. (2013). However, most of the algorithms used in that paper are not used practically very much. It was also noted by the authors themselves that k-Means++ and its greedy version work better than most. They also mention that probabilistic algorithms perform better than deterministic ones. Moreover, another highly influential center initialization algorithm (Ostrovsky et al., 2006) was not considered in their experiments. There is no mention of Ostrovsky's algorithm in their paper whatsoever. Surprisingly, there is no mention of Ostrovsky's algorithm in k-Means++ paper either even though it was published in 2007 whereas Ostrovsky's algorithm was published in 2006.

We would also like to point out that to our knowledge no surveys were done after removing linear dependency prior to running the experiments. This is a very important step if we are to get a meaningful clustering out of k-Means algorithm. Because this removes a big problem that k-Means is known to suffer from. PCA decomposes the existing variables into orthogonal[2] ones. If we use PCA (principal component analysis) before running a clustering algorithm, we can redefine the variables into linearly independent ones. For our experiment, we have used PCA on every dataset before running cluster algorithm.

3. Proposed Initialization Methods

For a set of points S and a point x, we use $\min(\|x-S\|)$ to denote the minimum of distances from x to the points of S that is $\min(\|x-S\|) = \min_{a\in S}(\|x-a\|)$. Set $D(\mathbf{x}) = \min(\|\mathbf{x}-\mathcal{C}\|)$ for a point x and a set of centers $\mathcal{C}$. Let us denote the centroid of S by μ_S that is $\mu_S = \frac{1}{|S|}\sum_{x\in S} x$.

3.1 Variance Based Seeding

Roughly our idea is that we want to minimize the variance of distances from the new center to existing ones. The intuition behind this is to make the clusters as balanced as possible, hence possibly eliminating some anomalies within a cluster and reducing inertia. However, since the new centers are taken based on variances, we need two initial centers. We chose to initialize the first two centers as in ORSS algorithm. Assume that for a new point $\mathbf{x}$ and a set of centers $\mathcal{C}_k = \{c_1, \cdots, c_k\}$, the variance of the squared distances $\{\|c_1 - \mathbf{x}\|^2, \|c_2 - \mathbf{x}\|^2 \cdots, \|c_k - \mathbf{x}\|^2\}$ is $\nu_{\mathbf{x}}(\mathcal{C}_k)$.

i Choose two centers $\mathbf{x}, \mathbf{y}$ with probability proportional to $\|\mathbf{x} - \mathbf{y}\|^2$. Set $\mathcal{C}_2 = \{\mathbf{x}, \mathbf{y}\}$.

ii For already existing set of i centers $\mathcal{C}_i = \{c_1, \cdots, c_i\}$, choose a new center $\mathbf{x} \in \mathbf{X}$ with probability $\rho(\mathbf{x}) = 1 - \frac{\nu_{\mathbf{x}}(\mathcal{C}_i)}{\sum_{\mathbf{y}\in\mathbf{X}} \nu_{\mathbf{y}}(\mathcal{C}_i)}$.

iii Repeat step (ii) until $i = k$.

iv For each $1 \leq i \leq k$, set $\mathcal{C}_i = \{\mathbf{x} \in \mathbf{X} : \|\mathbf{x} - c_i\| = \min(x - \mathcal{C})\}$.

2. It is well known that orthogonal vectors are linearly independent.

v Set $c_i = \mu_{\mathcal{C}_i}$.

vi Repeat (iv) and (v) until convergence or number of iteration is reached.

3.2 Centroid of Centers Based Seeding

In this method, we want to choose x with probability proportional to squared distance from centroid of the cluster centers. As we will show later, this is inspired from the idea of choosing x with probability proportional to $\sum_{c \in \mathcal{C}} \|x - c\|^2$.

i Choose a centers $\mathbf{x}$ uniformly at random. Set $\mathcal{C}_1 = \{\mathbf{x}\}$.

ii For already existing set of i centers $\mathcal{C}_i = \{c_1, \cdots, c_i\}$, choose a new center $\mathbf{x} \in \mathbf{X}$ with probability proportional to $\|\mathbf{x} - \mu_{\mathcal{C}_i}\|^2$.

iii Repeat step (ii) until $i = k$.

iv For each $1 \leq i \leq k$, set $\mathcal{C}_i = \{\mathbf{x} \in \mathbf{X} : \|\mathbf{x} - c_i\| = \min(x - \mathcal{C})\}$.

v Set $c_i = \mu_{\mathcal{C}_i}$.

vi Repeat (iv) and (v) until convergence or number of iteration is reached.

3.3 First center for k-Means++

In k-Means++ algorithm, the first center is chosen uniformly at random. However, not all points have the same contribution to inertia. We choose x as a first center in a way that is analogous to explained variance ratio by x.

i Choose $\mathbf{x}$ with probability $\dfrac{f(\mathbf{x})}{\sum_{\mathbf{y} \in \mathbf{X}} f(\mathbf{y})}$ where $f(x) = \sum_{\mathbf{y} \in \mathbf{X}} \|\mathbf{x} - \mathbf{y}\|^2$. Set $\mathcal{C}_1 = \{\mathbf{x}\}$.

ii Repeat the remaining steps in k-Means++ (Arthur et al., 2007, Section 2.2, Page 3).

4. Analysis

First, we will analyze the choosing of first center in k-Means++.

4.1 k-Means++ Improved

Consider a set of n points $\mathbf{X}$. Let us write $\sum_{\mathbf{x} \in \mathbf{X}} \mathbf{x}$ as $\mathcal{S}$ and the probability of $x \in \mathbf{X}$ being chosen as a center as $p(x)$. From the definition of variance, for a set of points S,

$$\sigma^2(S) = \frac{\sum_{x \in S} \|x - \mu_S\|^2}{|S|} - \|\mu_S\|^2$$

$$\sum_{x \in S} \|x - \mu_S\|^2 = |S|(\sigma^2 + \|\mu_S\|^2) \tag{1}$$

For any arbitrary point a and μ as the centroid of $\mathbf{X}$,

$$\begin{aligned}
\sum_{\mathbf{x}\in\mathbf{X}} \|\mathbf{x}-a\|^2 &= \sum_{\mathbf{x}\in\mathbf{X}} \|\mathbf{x}-\mu+\mu-a\|^2 \\
&= \sum_{\mathbf{x}\in\mathbf{X}} \left(\|\mathbf{x}-\mu\|^2 + 2\langle \mathbf{x}-\mu, \mu-a\rangle + \|\mu-a\|^2\right) \\
&= \sum_{\mathbf{x}\in\mathbf{X}} \|\mathbf{x}-\mu\|^2 + 2\left\langle \sum_{\mathbf{x}\in\mathbf{X}} \mathbf{x}-n\mu, \mu-a \right\rangle + n\|\mu-a\|^2 \\
&= n(\sigma^2 + \|\mu\|^2) + 2\langle n\mu - n\mu, \mu-a\rangle + n\|\mu-a\|^2 \\
&= n(\sigma^2 + \|\mu\|^2 + \|\mu-a\|^2) \qquad (2)
\end{aligned}$$

This is the inspiration we mentioned in centroid of centers seeding. Using equation (2), we have the following.

$$\begin{aligned}
\sum_{\mathbf{x}\in\mathbf{X}}\sum_{\mathbf{y}\in\mathbf{X}} \|\mathbf{x}-\mathbf{y}\|^2 &= \sum_{\mathbf{x}\in\mathbf{X}} n(\sigma^2 + \|\mu\|^2 + \|\mu-\mathbf{x}\|^2) \\
&= n(n\sigma^2 + n\|\mu\|^2 + \sum_{\mathbf{x}\in\mathbf{X}} \|\mu-\mathbf{x}\|^2) \\
&= n(n\sigma^2 + n\|\mu\|^2 + n(\sigma^2 + \|\mu\|^2)) \\
&= 2n^2(\sigma^2 + \|\mu\|^2) \qquad (3)
\end{aligned}$$

Using equation (3), the probability becomes

$$\begin{aligned}
p(x) &= \frac{\sum_{\mathbf{y}\in\mathbf{X}} \|\mathbf{x}-\mathbf{y}\|^2}{\sum_{\mathbf{y}\in\mathbf{X}}\sum_{\mathbf{x}'\in\mathbf{X}} \|\mathbf{x}'-\mathbf{y}\|^2} \\
&= \frac{n(\sigma^2 + \|\mu\|^2 + \|\mu-\mathbf{x}\|^2)}{2n(\sigma^2 + \|\mu\|^2)} \\
&= \frac{\sigma^2 + \|\mu\|^2 + \|\mu-\mathbf{x}\|^2}{2(\sigma^2 + \|\mu\|^2)} \\
&= \frac{1}{2} + \frac{\|\mu-\mathbf{x}\|^2}{2(\sigma^2 + \|\mu\|^2)}
\end{aligned}$$

Now it becomes more clear how this is related to explained variance ratio. However, to make things smoother, one can also choose to use the following as the probability of x being chosen as the first center.

$$p(x) = \frac{\|\mu-\mathbf{x}\|^2}{2(\sigma^2 + \|\mu\|^2)}$$

Also, notice that computationally this version of $p(x)$ is much cheaper than the one we mentioned above. Therefore, we considered $p(x)$ proportional to $\|x-\mu\|^2$ for choosing x as the first center in our experiments.

4.2 Centroid of Centers Seeding

We will now analyze the center initialization algorithms using inertia value. Since these algorithms are probabilistic, we are going to take a look at the expected value of inertia. For a set of points $\mathbf{X}$ and a set of centers $\mathcal{C}$, we denote the inertia by $\mathcal{I}_{\mathcal{C}}(\mathbf{X})$. For the optimal set of cluster centers $\mathcal{C}_{opt}$, we denote the corresponding inertia by $\mathcal{I}_{opt}(\mathbf{X})$.

$$\mathcal{I}_{\mathcal{C}}(\mathbf{X}) = \sum_{\mathbf{x}\in\mathbf{X}} \min_{\mathbf{c}\in\mathcal{C}}(\|c - x\|^2)$$

If the context is clear, we may omit $\mathcal{C}$ and $\mathbf{X}$. For a fixed set of points $\mathbf{X}$, if the probability of $\mathbf{x} \in \mathbf{X}$ being chosen to be a center is $p(x)$ with respect to a set of centers $\mathcal{C}$, then the expected value of inertia $E[\mathcal{I}(\mathbf{X})]$ is

$$E[\mathcal{I}(\mathbf{X})] = \sum_{\mathbf{x}\in\mathbf{X}} p(x) \sum_{\mathbf{y}\in\mathbf{X}} \min(D(x), \|y - x\|)^2$$

The following lemma was proven in (Arthur et al., 2007, Lemma 3.2).

Lemma 1 *Let A be an arbitrary cluster in $\mathcal{C}_{opt}$, and let $\mathcal{C}$ be the clustering with just one center, which is chosen uniformly at random from A. Then, $E[\mathcal{I}(A)] = 2\mathcal{I}_{opt}(A)$.*

Here, $\mathcal{I}_{opt}(A) = \sum_{x\in A} \sum_{b\in A} \|a - b\|^2$.

Now, we want to find a similar formula for the remaining centers using centroid of centers seeding. If we take $p(x)$ proportional to $f(x) = \|x - \mu_{\mathcal{C}}\|^2$, then using equation (2),

$$\begin{aligned} f(x) &= \|x - \mu_{\mathcal{C}}\|^2 \\ \sum_{x\in S} f(x) &= \sum_{x\in S} \|x - \mu_{\mathcal{C}}\|^2 \\ &= |S|(\sigma^2 + \|\mu_S\|^2 + \|\mu_S - \mu_{\mathcal{C}}\|^2) \end{aligned}$$

Then the probability of x being chosen as a center is

$$\begin{aligned} p(x) &= \frac{f(x)}{\sum_{y\in S} f(y)} \\ &= \frac{\|x - \mu_{\mathcal{C}}\|^2}{|S|(\sigma^2 + \|\mu_S\|^2 + \|\mu_S - \mu_{\mathcal{C}}\|^2)} \end{aligned}$$

Since $D(y) \leq \|y - \mu_{\mathcal{C}}\|$, we have the following

$$\begin{aligned} \sum_{y\in S} \|y - \mu_{\mathcal{C}}\|^2 &\geq D(y)^2 \\ \frac{1}{\sum_{y\in S} \|y - \mu_{\mathcal{C}}\|^2} &\leq \frac{1}{\sum_{y\in S} D(y)^2} \\ \frac{\sum_{y\in S} D(y)^2}{\sum_{y\in S} \|y - \mu_{\mathcal{C}}\|^2} &\leq 1 \end{aligned}$$

Our expected value of inertia would be

$$\begin{aligned}
E[\mathcal{I}(S)] &= \sum_{x \in S} p(x) \sum_{y \in S} \min(D(y), \|x - y\|)^2 \\
&= \sum_{x \in S} \frac{\|x - \mu_{\mathcal{C}}\|^2}{\sum_{y \in S} \|y - \mu_{\mathcal{C}}\|^2} \sum_{y \in S} \min(D(y), \|x - y\|)^2 \\
&\leq \sum_{x \in S} \frac{\|x - \mu_{\mathcal{C}}\|^2}{\sum_{y \in S} \|y - \mu_{\mathcal{C}}\|^2} \sum_{y \in S} D(y)^2 \\
&\leq \sum_{x \in S} \|x - \mu_{\mathcal{C}}\|^2
\end{aligned}$$

Using equation (2), we have

$$E[\mathcal{I}(S)] \leq n(\sigma^2 + \|\mu_S\|^2 + \|\mu_S - \mu_{\mathcal{C}}\|^2)$$

Now, using equation (3), we have

$$\begin{aligned}
\mathcal{I}_{opt}(S) &= \sum_{x \in S} \sum_{y \in S} \|x - y\|^2 \\
&= 2n^2(\sigma^2 + \|\mu_S\|^2) \\
\frac{\mathcal{I}_{opt}(S)}{2n} &= n(\sigma^2 + \|\mu_S\|^2)
\end{aligned}$$

Thus, we have the following theorem.

Theorem 2 *Let S be an arbitrary cluster chosen from $\mathcal{C}_{opt}$ and $\mathcal{C}$ be an arbitrary clustering. If a center is chosen from S using centroid of centers initialization, then*

$$E[\mathcal{I}(S)] \leq \frac{\mathcal{I}_{opt}(S)}{2n} + n\|\mu_S - \mu_{\mathcal{C}}\|^2$$

5. Experiment Setup

We ensure that all algorithms are run under the same conditions. All of them share the same environment and no special optimizations were made for any particular algorithm. Only CPU was used to determine the values we are interested in and no parallelism mechanism was in place for speeding up the process. This way, we can get an idea about the raw performance metrics of the algorithms involved.

Python is used as the programming language to write necessary codes. Some common auxiliary packages such as *scipy, scikit-learn, numpy* etc are used to help with the code. The algorithms are simply different methods of the same class, so they share the same fitting and prediction function. Only the initialization differs for different algorithm. It should be mentioned that even though some packages have native support for k-Means implementation, we did not use them to run the experiments. Not all algorithms we want to test are available in those packages. Therefore, in order to ensure same environment and optimizations for every algorithm, we wrote them all from scratch so that we could be sure they are tested under the same settings.

Algorithm	Average	Minimum	Time
k-Means	1603.09	1577.05	9.27s
k-Means++	1563.71	1518.9	9.4s
ORSS	1569.41	1530.46	9.44s
Var-based	1552.37	1510.91	9.69s

Table 1: Results on Cloud data set

The data sets we have used are some popular ones: Iris data set, Mall customers data set, Airline clustering data set, Wine testing dataset. We did not duel too much on using too many data sets. Miligan et al. (1988) shows that using z-score to standardize the data is not favorable for clustering because it loses between-cluster variation. Therefore, we did not use any sort of standardization or normalization lest it should lose variance. Instead, we have used PCA to remove linear dependency among variables.

While experimenting on such algorithms, it is of utmost importance to run the same experiment more than once under the same parameters and conditions. For example, assume that we want to compare k-Means++ and Forgy's algorithm (Forgy, 1965) for $k = 5$ clusters. We should run this experiment at least m times where $m > 1$ in order to eliminate bias and account for randomness. We run each experiment a total of 20 times and take the average and minimum values of inertia. The number of iterations was set to 300.

6. Results

We present the results on those four algorithms based on inertia and time required to run them. Instead of checking how many iterations it take for the centers to converge, we have let each algorithm run exactly the same number of times (300) which is more than enough for convergence. The time registered here is the time taken by each algorithm for those 300 iterations. As expected, default k-Means is the fastest algorithm. However, if we were to define convergence, that leads to a separate problem which we think other implementations suffer from. We also mentioned something similar before. This might also be one of the reasons why our results of k-Means++ differ from the authors implementation. Our argument is also complemented by the experiments. All algorithms seem to reach the same minimum inertia. Therefore, it is the average of inertia that we should focus on.

6.1 Comparison of Different Initialization

The comparison of inertia for different initialization is shown in tables 1, 2, 3.

6.2 k-Means++ vs k-Means++ Improved

The comparison of k-Means++ against k-Means++ improved is shown in tables 4, 5.

Notice that in all cases, we achieved lower minimum inertia for k-Means++ and yet k-Means++ improved version achieved lower inertia on average. Moreover, even though these are randomized algorithms, pretty much every time we ran the experiments, k-Means++ improved version performed better. Given the mathematical intuitive justification and

Algorithm	Average	Minimum	Time
k-Means	57.74	43.89	2.26s
ORSS	48.25	43.8	2.28s
k-Means++	49.19	43.83	2.28s
Var-based	47.05	44.79	2.32s

Table 2: Results on IRIS data set

Algorithm	Average	Minimum	Time
k-Means	180.37	174.92	0.95s
ORSS	183.13	174.89	0.96s
k-Means++	181.23	0.96	1.72s
Var-based	175.4	174.85	0.97s

Table 3: Results on mall customers data set

Algorithm	Average $\mathcal{I}$	Minimum $\mathcal{I}$	Time
k-Means++	73.3	57.23	0.74s
k-Means++ Improved	67.58	57.26	0.74s

Table 4: Results on Iris dataset, 5 clusters

Algorithm	Average $\mathcal{I}$	Minimum $\mathcal{I}$	Time
k-Means++	89866.67	78385.08	0.99s
k-Means++ Improved	88468.3	78392.42	0.99s

Table 5: Results on Mall customer dataset, 5 clusters

experimental results, we are inclined to say that this version works better than original k-Means++ consistently.

References

D. Arthur and S. Vassilvitskii. `k-means++`: The Advantages of Careful Seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. Biometrics, $21, 768 - 780, 1965$.

Stuart P. Lloyd. Least squares quantization in pcm. IEEE Transactions on Information Theory, $28(2) : 129–136, 1982$.

M. E. Dyer. A simple heuristic for the p-center problem. Operations Research Letters, Volume 3, February 1985, pp. $285 - 288$.

G. Milligan, M. C. Coope. A Study of Standardization of Variables in Cluster Analysis, Journal of Classification 5(2) (1988) 181–204.

R. Ostrovsky, Y. Rabani, Leonard J. Schulman, C. Swamy. The Effectiveness of Lloyd-Type Methods for the k-Means Problem. Proceedings of the 47th Annual Symposium on Foundations of Computer Science. 2006.

M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela. A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. Expert Systems with Applications, $40(1) : 200–210, 2013$.

Tom M. Apostol. Introduction to Analytic Number Theory. Springer, $10.1007/978 - 3 - 662 - 28579 - 4$, 1976.