K-Means Clustering with Math

Common Unsupervised learning technique for data analysis

sampath kumar gajawada · Follow
Published in Towards Data Science
4 min read · May 13, 2019

Listen Share

Photo by Perry Grone on Unsplash

When we are working with huge volumes of data, it makes sense to partition the data into logical groups and doing the analysis. We can use Clustering to make the data into groups with the help of several algorithms like K-Means.

In this article, I will try to address

a. Clustering

b. K-Means and working of the algorithm.

c. Choosing the right K Value

Clustering

A process of organizing objects into groups such that data points in the same groups

K-Means clustering is a type of unsupervised learning. The main goal of this algorithm to find groups in data and the number of groups is represented by K. It is an iterative procedure where each data point is assigned to one of the K groups based on feature similarity.

Algorithm

K-Means algorithm starts with initial estimates of K centroids, which are randomly selected from the dataset. The algorithm iterates between two steps *assigning data points* and *updating Centroids.*

Data Assignment

In this step, the data point is assigned to its nearest centroid based on the squared Euclidean distance. Let us assume a Cluster with c as centroid and a data point x is assigned to this cluster, based on the distance between c,x. There are some other distance measures like Manhattan, Jaccard, and Cosine which are used based on the appropriate type of data.

Centroid Update

Centroids are recomputed by taking the mean of all data points assigned to a particular cluster.

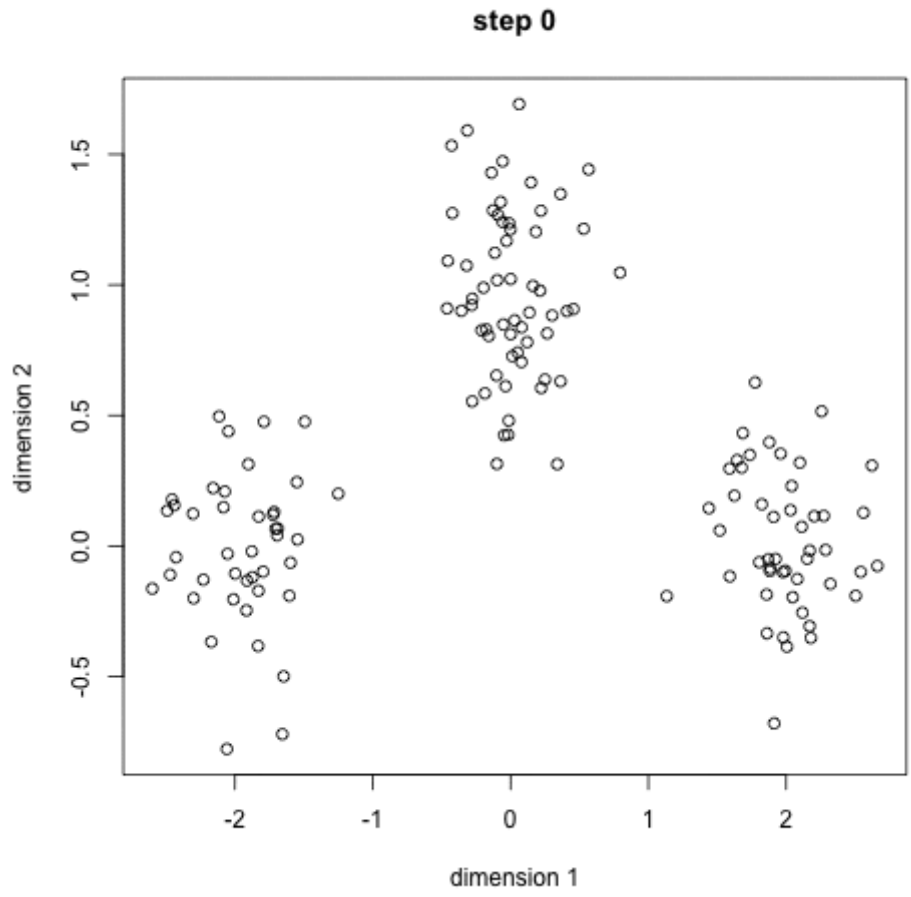


image by George Seif

Let us go through the above steps using the example below.

1. Consider 4 data points A,B,C,D as below

	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

Observations

2. Choose two centroids AB and CD, calculated as

> *AB = Average of A, B*
>
> *CD = Average of C,D*

	X1	X2
AB	4	2
CD	2	1

Two centroids AB, CD

3. Calculate squared euclidean distance between all data points to the centroids AB, CD. For example distance between A(2,3) and AB (4,2) can be given by $s = (2–4)^2 + (3–2)^2$.

	A	B	C	D
AB	5	5	9	5
CD	4	16	2	2

A is very near to CD than AB

4. If we observe in the fig, the highlighted *distance between (A, CD) is 4 and is less compared to (AB, A) which is 5. Since point A is close to the CD we can move A to CD cluster.*

5. There are two clusters formed so far, let recompute the centroids i.e, B, ACD similar to step 2.

> *ACD = Average of A, C, D*
>
> *B = B*

	X1	X2
B	6	1
ACD	2	1.67

New centroids B, ACD

6. As we know K-Means is iterative procedure now we have to calculate the distance of all points (A, B, C, D) to new centroids (B, ACD) similar to step 3.

	A	B	C	D
B	20	0	26	10
ACD	1.78	16.44	1.11	3.78

Clusters B, ACD

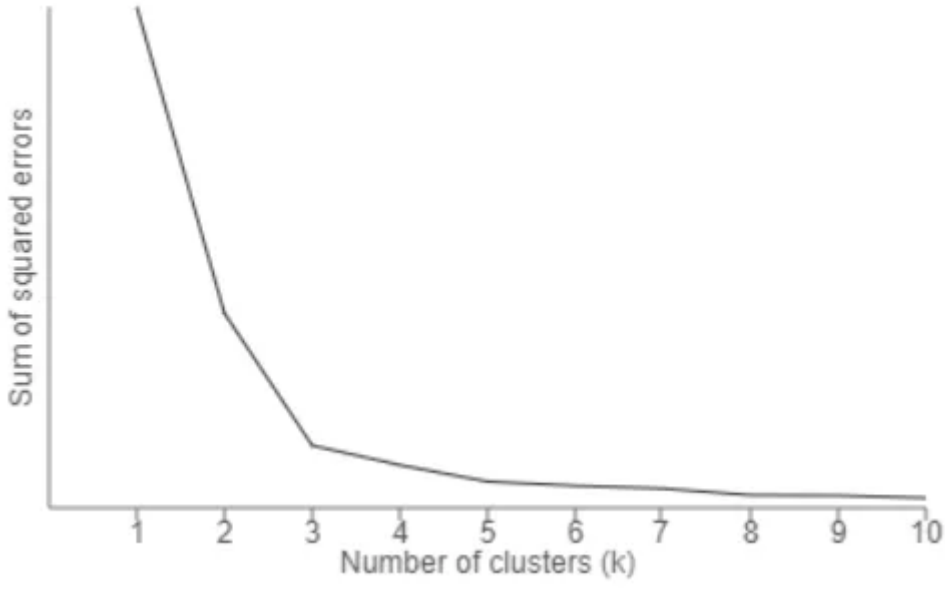
7. In the above picture, we can see respective cluster values are minimum that A is too far from cluster B and near to cluster ACD. All data points are assigned to clusters (B, ACD) based on their minimum distance. The iterative procedure ends here.

8. To conclude, we have started with two centroids and end up with two clusters, K=2.

Choosing K

One method of choosing value K is *the elbow method.* In this method we will run K-Means clustering for a range of K values lets say (K= 1 to 10) and calculate the Sum of Squared Error (SSE). SSE is calculated as the mean distance between data points and their cluster centroid.

Then plot a line chart for SSE values for each K, if the line chart looks like an arm then the elbow on the arm is the value of K that is the best.



Choose the Best K

Hope you enjoyed it!! Please do comment on any queries or suggestions.

Machine Learning | Data Science | Clustering | Mathematics | Unsupervised Learning

Follow

Written by sampath kumar gajawada

337 Followers · Writer for Towards Data Science

Machine learning Enthusiast | Analyst | Programmer | All I write my own | Linkedin: https://www.linkedin.com/in/sampath-kumar-gajawada/

More from sampath kumar gajawada and Towards Data Science