

Data Clustering Algorithms

 Search this site


"As knowledge increases,
wonder deepens"

Quick Links

[Introduction](#)

[k-means clustering algorithm](#)

[Fuzzy c-means clustering algorithm](#)

[Hierarchical clustering algorithm](#)

[Gaussian\(EM\) clustering algorithm](#)

[Quality Threshold \(QT\) clustering algorithm](#)

[MST based clustering algorithm](#)

[Density based clustering algorithm](#)

[kernel k-means clustering algorithm](#)

[Clustering Algorithm Applications](#)

[FAQ](#)

[References](#)

[My Concise CV](#) [click here](#)

Reach Me

Email: [click here](#)

Facebook: [click here](#)

[My Blog](#) [click here](#)

Interesting Links

Sixth Sense: [Part 1](#) [Part 2](#) [Part 3](#) [Next Generation Mobile Phones](#)
[TED Conferences](#)

k-means clustering algorithm

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^c \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

where,

' $\|x_i - v_j\|$ ' is the Euclidean distance between x_i and v_j .

' c_i ' is the number of data points in i^{th} cluster.

' c ' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \dots, v_c\}$ be the set of centers.

- 1) Randomly select ' c ' cluster centers.
- 2) Calculate the distance between each data point and cluster centers.

Fun Gallery
[Baby Dance Video](#) [Puzzle Game](#)
Miscellaneous
<https://www.kaggle.com/>
[Interesting article to read](#)

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$\mathbf{v}_i = (1 / c_i) \sum_{j=1}^{c_i} \mathbf{x}_j \quad \text{where, 'c}_i\text{' represents the number of data points in } i^{th} \text{ cluster.}$$

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

Advantages

- 1) Fast, robust and easier to understand.
- 2) Relatively efficient: $O(tknd)$, where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, $k, t, d \ll n$.
- 3) Gives best result when data set are distinct or well separated from each other.

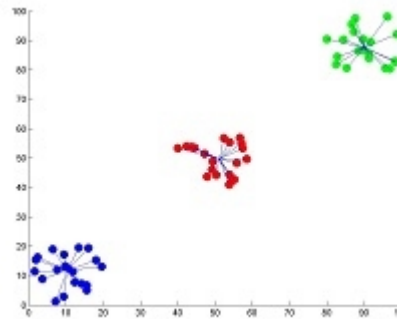


Fig I: Showing the result of k-means for ' N ' = 60 and ' c ' = 3

Note: For more detailed figure for k-means algorithm please refer to [k-means figure](#) sub page.

Disadvantages

- 1) The learning algorithm requires apriori specification of the number of cluster centers.

- 2) The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
- 3) The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data we get different results (data represented in form of cartesian co-ordinates and polar co-ordinates will give different results).
- 4) Euclidean distance measures can unequally weight underlying factors.
- 5) The learning algorithm provides the local optima of the squared error function.
- 6) Randomly choosing of the cluster center cannot lead us to the fruitful result. Pl. refer **Fig.**
- 7) Applicable only when mean is defined i.e. fails for categorical data.
- 8) Unable to handle noisy data and outliers.
- 9) Algorithm fails for non-linear data set.

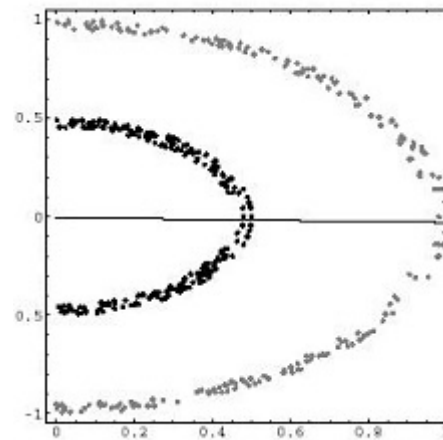


Fig II: Showing the non-linear data set where k-means algorithm fails











References

- 1) An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, David M. Mount,
Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.
- 2) Research issues on K-means Algorithm: An Experimental Trial Using Matlab by Joaquin Perez Ortega, Ma. Del
Rocio Boone Rojas and Maria J. Somodevilla Garcia.
- 3) The k-means algorithm - Notes by Tan, Steinbach, Kumar Ghosh.
- 4) http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

5) k-means clustering by ke chen.

Subpages (2): [k-means algorithm figure](#) [k-means initial cluster center selection](#)



 Ref-1_k-means.pdf (1930k)	Azad Naik, May 11, 2010, 1:51 AM	v.1	
 Ref-2_k-means.pdf (764k)	Azad Naik, May 11, 2010, 1:52 AM	v.1	
 Ref-3_k-means.pdf (361k)	Azad Naik, May 11, 2010, 1:53 AM	v.1	
 Ref-5_k-means.ppt (498k)	Azad Naik, May 14, 2010, 1:21 AM	v.1	
 Ref-6_k-means.ppt (784k)	Azad Naik, May 16, 2011, 6:19 AM	v.1	

Comments

You do not have permission to add comments.