An Efficient K-Means Clustering Algorithm

Presented By :

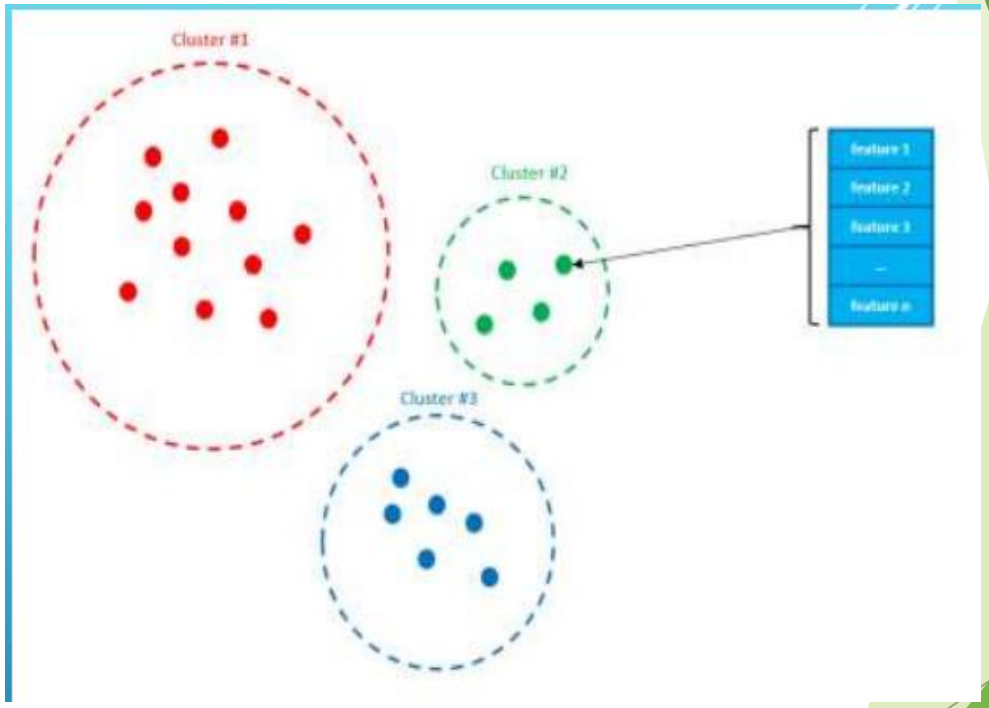
Koushik Veldhi

Sai Charan Kodadi

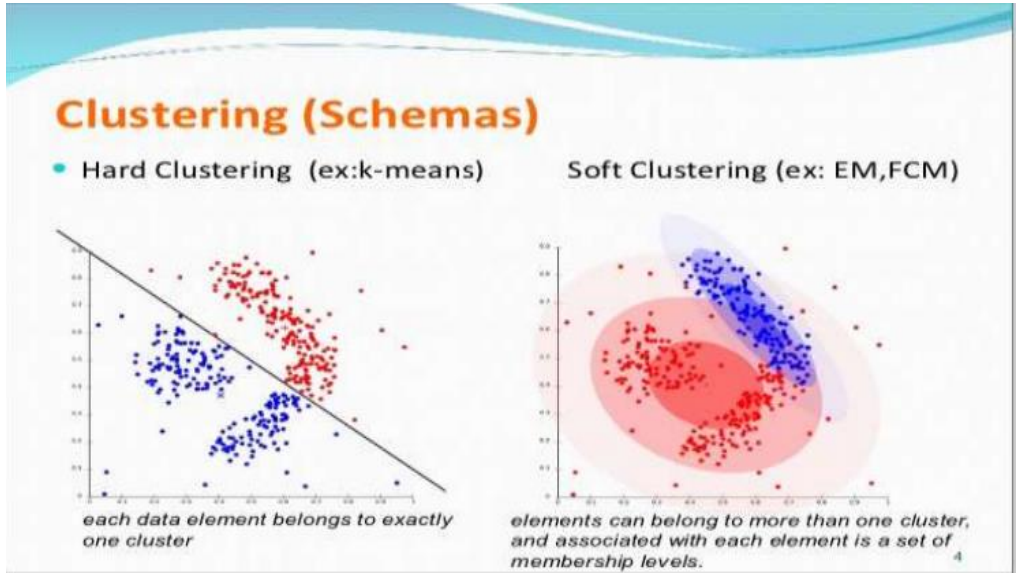
Pavan Kumar Reddy Alavala

Contents:

Clustering basic concept



- CLUSTERING Clustering is traditionally viewed as an unsupervised method for data analysis. Clustering is the task of the population or data points into a number of groups such that data points in the same groups are more to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

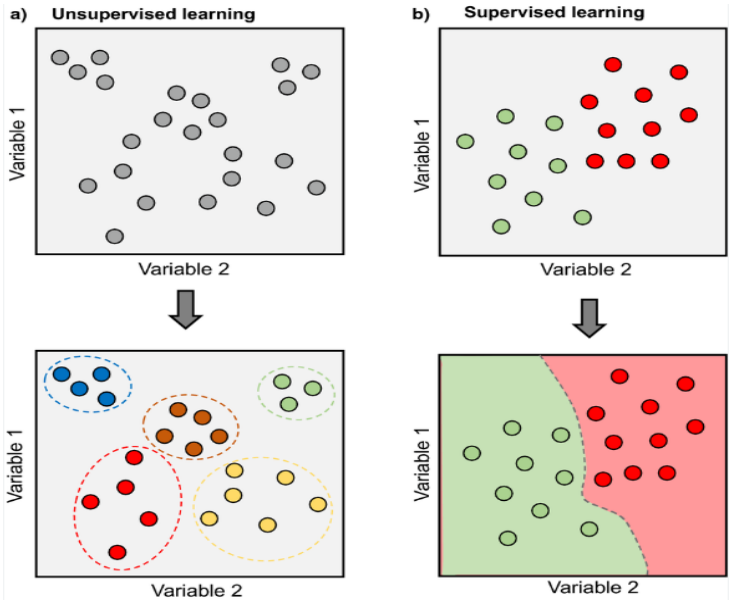


Types of clustering

- Broadly speaking, clustering can be divided into two subgroups
- HARD CLUSTERING: In hard clustering, each data point either belongs to a cluster completely or not. As an instance, we want the algorithm to read all of the tweets and determine if a tweet is a positive or a negative tweet.
- SOFT CLUSTERING: In the soft clustering method, each data point will not completely belong to one cluster, instead, it can be a member of more than one cluster it has a set of membership coefficients corresponding to the probability of being in a given cluster. As an instance, if you are attempting to forecast the rating changes for the counterparties who you trade with. . The algorithm can create clusters for each rating and indicate the likelihood of a counterparty to belong to a cluster.

types of clustering

Supervised Classification	Unsupervised Clustering
• known number of classes • based on a training set • used to classify future observations	• unknown number of classes • no prior knowledge • used to understand (explore) data



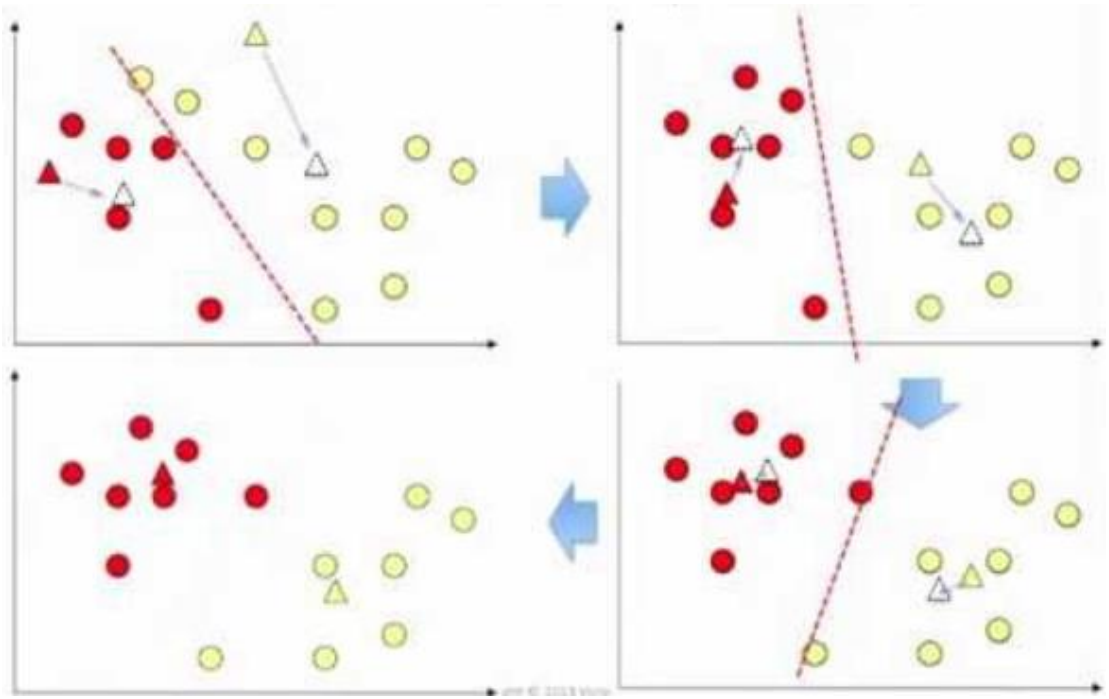
- Is clustering typically …?
- A. Supervised
- B. Unsupervised

K MEANS CLUSTERING

- K-means clustering (Macqueen, 1967) is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster. K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.E., Data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K ($N \geq K$). The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the k-means clustering algorithm are: 1.the centroids of the K clusters, which can be used to label new data. 2.labels for the training data (each data point is assigned to a single cluster)

K means clustering algorithms

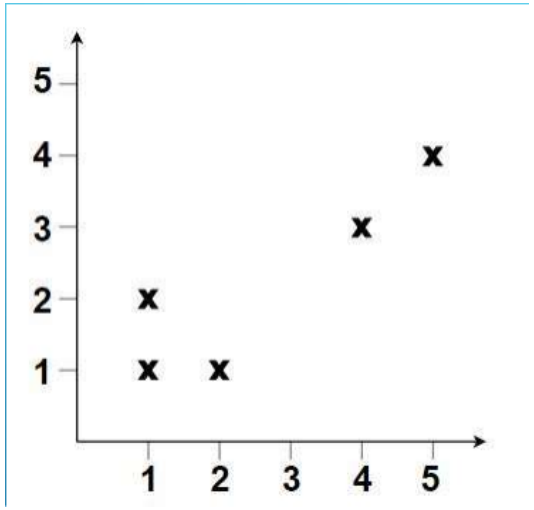
- AS, YOU CAN SEE, K-MEANS ALGORITHM IS COMPOSED OF 3 STEPS:
- STEP 1: INITIALIZATION The first thing k-means does, is randomly choose K examples (data points) from the dataset as initial centroids and that's simply because it does not know yet where the center of each cluster is. (A centroid is the center of a cluster).
- STEP 2: CLUSTER ASSIGNMENT Then, all the data points that are the closest (similar) to a centroid will create a cluster. If we're using the Euclidean distance between data points and every centroid, a straight line is drawn between two centroids, then a perpendicular bisector (boundary line) divides this line into two clusters.
- STEP 3: MOVE THE CENTROID Now, we have new clusters, that need centers. A centroid's new value is going to be the mean of all the examples (data points) in a cluster. We'll keep repeating step 2 and 3 until the centroids stop moving, in other words, k-means algorithm is converged.



K means clustering algorithms

K means clustering algorithm

- CLUSTER ANALYSIS – EXAMPLE We will work with a real-number example of the well-known k-means clustering algorithm. We will try to find clusters in the below dataset, consisting of 5 points.



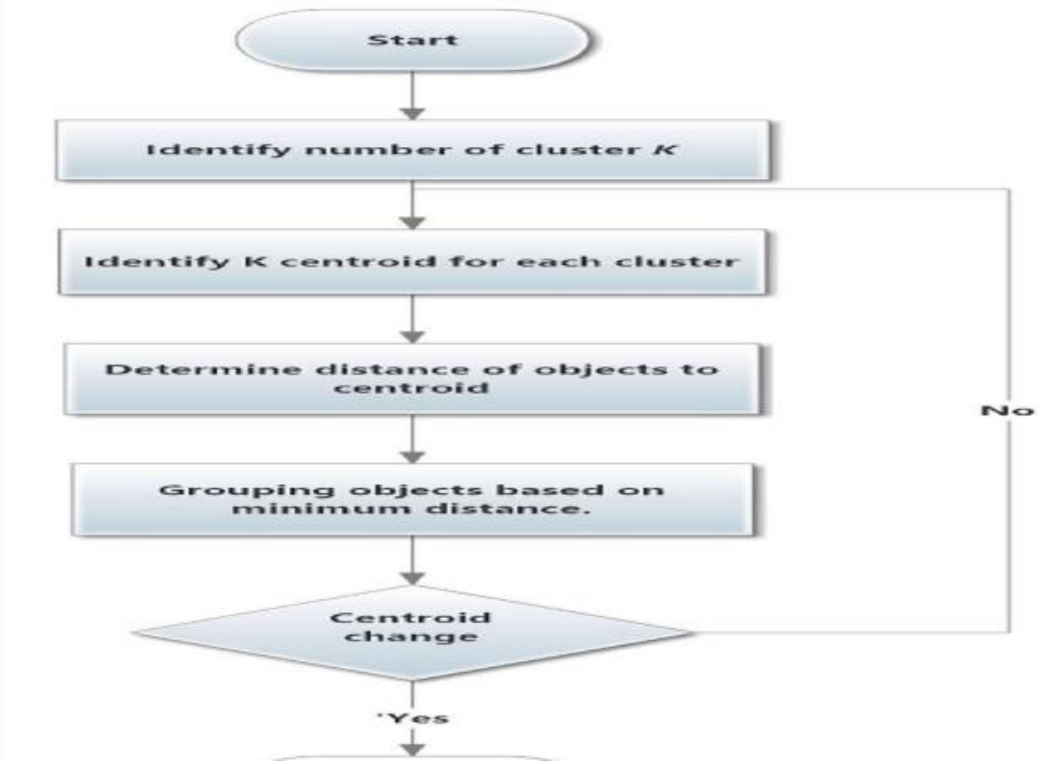
i	X	Y
A	1	1
B	2	1
C	1	2
D	4	3
E	5	4

K means clustering algorithms

- STEP 1: SET CLUSTER QUANTITY The k-means algorithm requires you to set a number of clusters k beforehand. Here, we take k=2(the data look like there clusters – one on the bottom left and one on the top right).

- STEP 2: ASSIGNMENT OF DATA POINTS In the assignment step, each data point gets assigned to the nearest cluster centroid. The cluster centroids can be seen as centers of gravity within each cluster. To start with, we chose random points as centroids. Here, we take point A(1,1) Instead of taking actual data points, we could have taken completely random points as well. To calculate the nearest cluster centroid for each data point, you need a distance measure. There is a large number of available metrics doing the job. We will work with the ordinary Euclidian distance

K means clustering algorithms



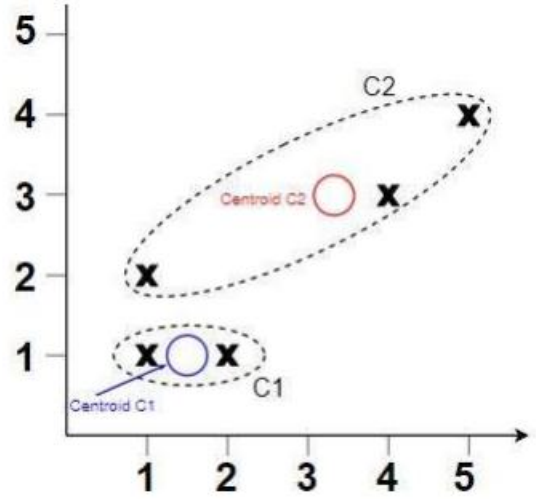
- STEP 3: MOVE THE CENTROID Now, we have new clusters, that need centers. A centroid's new value is going to be the mean of all the examples in a cluster. We'll keep repeating step 2 and 3 until the centroids stop moving, in other words, k-means algorithm is converged.

K means clustering algorithms

To calculate the nearest cluster centroid for each data point, you need a distance measure. There is a large number of available metrics doing the job. We will work with the ordinary Euclidian distance: $d(a,b) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$.

Distances to centroids	$(1,1)$	$(1,2)$
A $(1,1)$	**0**	1
B $(2,1)$	**1**	1.4
C $(1,2)$	1	**0**
D $(4,3)$	3.6	**3.2**
E $(5,4)$	5	**4.5**

For example, the distance between centroid $(1,1)$ and A is 0:
$d((1,1),(1,1)) = \sqrt{(1-1)^2 + (1-1)^2} = 0$ (A is the centroid) and the distance between centroid $(1,2)$ and E is 4.5:
$d((1,2),(5,4)) = \sqrt{(1-5)^2 + (2-4)^2} \approx 4.5$



requirements

- Requirements of clustering in data mining:-
- 1. Scalability - we need highly scalable clustering algorithms to deal with large databases.
- 2. Ability to deal with different kind of attributes - algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- 3. Discovery of clusters with attribute shape - the clustering algorithm should be capable of detect cluster of arbitrary shape. The should not be bounded to only distance measures that tend to find spherical cluster of small size.
- 4. High dimensionality - the clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
- 5. Ability to deal with noisy data - databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters. 6. Interpretability - the clustering results should be interpretable, comprehensible and usable.

Applications

- HERE ARE 7 EXAMPLES OF CLUSTERING ALGORITHMS IN ACTION.
- 1. IDENTIFYING FAKE NEWS : The way that the algorithm works is by taking in the content of the fake news article, the corpus, examining the words used and then clustering them. These clusters are what helps the algorithm determine which pieces are genuine and which are fake news. Certain words are found more commonly in sensationalized, click-bait articles. When you see a high percentage of specific terms in an article, it gives a higher probability of the material being fake news.
- 2. SPAM FILTER : k-means clustering techniques have proven to be an effective way of identifying spam. The way that it works is by looking at the different sections of the email (header, sender, and content). The data is then grouped together. These groups can then be classified to identify which are spam. Including clustering in the classification process improves the accuracy of the filter to 97%. This is excellent news for people who want to be sure they're not missing out on your favorite newsletters and offers.

Applications

- 3. ASTRONOMY: It helps to find groups of similar stars and galaxies.
- 4. GENOMICS: It can be used to derive plant and animal taxonomies, categorize genes with similar functionality and gain insight into structures inherent in populations.
- 5. CLASSIFYING NETWORK TRAFFIC : k-means clustering is used to group together characteristics of the traffic sources. When the clusters are created, you can then classify the traffic types. The process is faster and more accurate than the previous auto class method. By having precise information on traffic sources, you are able to grow your site and plan capacity effectively.
- 6. IDENTIFYING FRAUDULENT OR CRIMINAL ACTIVITY : By analysing the GPS logs, the algorithm is able to group similar behaviors. Based on the characteristics of the groups you are then able to classify them into those that are real and which are fraudulent.
- 7. DOCUMENT ANALYSIS : Hierarchical clustering has been used to solve this problem. The algorithm is able to look at the text and group it into different themes. Using this technique, you can cluster and organize similar documents quickly using the characteristics identified in the paragraph.
- 8.CALL RECORD DETAIL ANALYSIS: A call detail record (CDR) is the information captured by telecom companies during the call, SMS, and internet activity of a customer.

K-means advantages and disadvantages

- Advantages of k-means :
- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples (data points).
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.
- Disadvantage of k-means :
- Choosing k manually being dependent on initial values.
- For a low k, you can mitigate this dependence by running k-means several times with different initial values and picking the best result.
- As k increases, you need advanced versions of k-means to pick better values of the initial centroids (called k-means seeding).
- Clustering data of varying sizes and density.
- Clustering outliers and Scaling with number of dimensions

Conclusion

- Conclusion: K means algorithm is useful for undirected knowledge discovery and is relatively simple. K mean has found wide spread usage in lot of field raging from unsupervised learning of neural ,Pattern recognitions, classification analysis, Artificial intelligence ,Image processing and many others