International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 1 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/IJSRCSEIT

Analysis of K-Mean Algorithm

Vinaya Durga M[1], Ganapathi Sharma K[2]
[1]Assistant Professor, St Aloysius College, Mangalore, Karnataka, India
[2]Associate Professor, Shrinivas University, Mangalore, Karnataka, India

ABSTRACT

Clustering is one among the foremost common preliminary knowledge associates to analysis technique to get an intuition regarding the structure of the info. It is often outlined because the task of characteristic subgroups within the knowledge such knowledge points within the same subgroup (cluster) area unit are similar whereas knowledge points totally different in numerous clusters area different. There are several algorithms which deals with unsupervised learning. K means algorithm is one of such algorithm. Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as far as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid , that is $(x2-x1)^2+ (y2-y1)^2$

Keywords : Kmean, Cluster, Datamining, Hierarchical

I. INTRODUCTION

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Clustering is one among the foremost common preliminary knowledge associates to alysis technique to get an intuition regarding the structure of the info. It is often outlined because the task of characteristic subgroups within the knowledge such knowledge points within the same subgroup (cluster) area unit are similar whereas knowledge points totally different in numerous clusters area different. In different words, we tend to try and realize homogenized subgroups at intervals of the knowledge or the information such data points in every cluster square measure as similar as attainable per a similarity measure like Euclidean-based distance or correlation-based distance. the choice of that similarity live to use is application-specific. Clustering analysis are often done on the idea of options or on the idea of samples. On the idea of options wherever we tend to try and realize subgroups of samples supported options. On the idea of samples wherever we tend to try and realize subgroups of options supported samples Clustering is taken into account as associate degree unattended learning technique. within the unattended learning the bottom truth doesn't exist to match the output of the cluster algorithmic program to verity labels to guage its performance.

Kmeans Algorithm

There are several algorithms which deals with unsupervised learning.K means algorithm is one of such algorithm. Kmeans algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as far as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid , that is $(x2-x1)^2+ (y2-y1)^2$

II. METHODS AND MATERIAL

The steps that the k-means algorithm follows:

1. Indicate number of clusters *K.*
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
    - Keep iterating until there is no change to the centroids. Compute the sum of the squared distance between data points and all centroids.
    - Assign each data point to the closest cluster (centroid).
    - Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Applications

- Kmeans algorithm is very well established algorithm.it is used in a several applications such as image segmentation and image compression, market segmentation, document clustering, etc.
- Image segmentation is that the classification of a picture into totally different teams.
- several researches are tired the world of image segmentation mistreatment bunch.
- There are a unit totally different ways and one in every of the foremost fashionable ways is k-means bunch formula.
- K -means bunch formula is AN unattended formula and it's wont to phase the interest space from the background.
- however, before applying K -means formula, initial partial stretching improvement is applied to the image to enhance the standard of the image.
- subtractive bunch methodology is information bunch methodology wherever it generates the centre of mass supported the potential worth of the information points.
- therefore subtractive cluster is employed to get the initial centers and these centers area unit utilized in k-means formula for the segmentation of image.
- Then finally medial filter is applied to the metameric image to get rid of any unwanted region from the image.
- customer segmentation Customer Segmentation is that the subdivision of a market into separate client teams that share similar characteristics. client Segmentation may be a strong means that to spot unhappy client desires. mistreatment the on top of knowledge firms will then outdo the competition by developing unambiguously appealing product and services
- The most common ways that within which businesses phase their client base are:
  1. Demographic info, like gender, age, familial and legal status, income, education, and occupation.
  2. Geographical info, that differs betting on the scope of the corporate. For localized businesses, this information would possibly pertain to specific cities or counties. For larger firms, it would mean a customer's town, state, or perhaps country of residence.
  3. Psychographics, like people, lifestyle, and temperament traits.
  4. Behavioural knowledge, like payment and consumption habits, product/service usage, and desired advantages.

Advantages of client Segmentation

1. Verify applicable product rating
2. Develop bespoke selling campaigns.
3. Style AN best distribution strategy.
4. opt for specific product options for preparation.
5. rate new development efforts.

III. RESULTS AND DISCUSSION

Advantages of client Segmentation

1. verify applicable product rating.
2. Develop bespoke selling campaigns
3. style AN best distribution strategy
4. opt for specific product options for preparation.
5. rate new development efforts.

K Means Clustering Algorithm

1. K means suggests that agglomeration algorithmic program
2. Specify range of clusters K.
3. Initialize centroids by 1st shuffling the information set then arbitrarily choosing K data points for the centroids while not replacement.
4. Keep iterating till there's no modification to the centroids. i.e assignment of information points to clusters isn't dynamic.
5. Document agglomeration
6. Document agglomeration may be a basic operation utilized n unattended document organization, automatic topic extraction and knowledge retrieval. agglomeration involves dividing a group of objects into a specified range of clusters. The motivation behind agglomeration a group {of information of knowledge of information} is to search out inherent structure within the data and expose this structure as a group of teams. the information objects among every cluster ought to exhibit an outsized degree of similarity whereas the similarity among completely different clusters ought to be reduced Their area unit 2 major agglomeration techniques:

"Partitioning" and "Hierarchical". Most document agglomeration algorithms may be classified into these 2 teams. hierarchic techniques manufacture a nested sequence of partition, with one, wide cluster at the highest and single clusters of individual points at all-time low. The partitioning agglomeration technique seeks to partition a set of documents into a group of non-overlapping teams, thus on maximize the analysis worth of agglomeration. image segmentation image compression Kmeans agglomeration is one in every of the foremost common agglomeration algorithms and frequently the primary issue practitioners apply once resolution agglomeration tasks to urge an inspiration of the structure of the dataset. The goal of kmeans is to cluster information points into distinct non-overlapping subgroups. It will a awfully smart job once the clusters have a sort of spherical shapes. However, it suffers because the geometric shapes of clusters deviates from spherical shapes. Moreover, it additionally doesn't learn the amount of clusters from the information and needs it to be pre-defined. To be a decent professional person, it's smart to grasp the assumptions behind algorithms/methods in order that you'd have a fairly smart plan concerning the strength and weakness of every technique. this may assist you decide once to use every technique and beneath what circumstances.

Formally, compression is that the variety of information compression applied to digital pictures to scale back their price of storage or transmission. Before moving on to the implementation, let's bear the K-means agglomeration algorithmic program in short. K-means agglomeration is that the optimisation technique to search out the 'k' clusters or teams within the given set of information points. the information points area unit clustered along on the premise of some reasonably similarity. Initially, it starts with the

random initialisation of the 'k' clusters then on the premise of some similarity (like euclidian distance metric), it aims to reduce the gap from each datum to the cluster center in every clusters.

IV.CONCLUSION

Kmeans clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of kmeans is to group data points into distinct non-overlapping subgroups. It does a very good job when the clusters have a kind of spherical shapes. However, it suffers as the geometric shapes of clusters deviates from spherical shapes. Moreover, it also doesn't learn the number of clusters from the data and requires it to be pre-defined. Kmeans agglomeration is one amongst the foremost widespread agglomeration algorithms and frequently the primary factor practitioners apply once finding agglomeration tasks to induce a thought of the structure of the dataset. The goal of kmeans is to cluster knowledge points into distinct non-overlapping subgroups. It will a awfully smart job once the clusters have a sort of spherical shapes. However, it suffers because the geometric shapes of clusters deviates from spherical shapes. Moreover, it conjointly doesn't learn the quantity of clusters from the info and needs it to be pre-defined. To be a decent professional, it's smart to understand the assumptions behind algorithms/methods so you'd have a reasonably smart plan concerning the strength and weakness of every methodology. this can assist you decide once to use every methodology and below what circumstances. during this post, we tend to lined each strength, weaknesses, and a few analysis strategies associated with kmeans.

Cite this article as :

Vinaya Durga M, Ganapathi Sharma K, "Analysis of K-Mean Algorithm", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 1, pp. 133-136, January-February 2020.

Journal URL : http://ijsrcseit.com/CSEIT206126