K-Means Algorithm and Application in Data Compression using Python

Nguyen D.D, Nune

Information System Institute

The State University of Management

Ryazanskiy Prospekt, 99, Moscow, 109542

Email: [Dungduongthanhson@gmail.com](mailto:Dungduongthanhson@gmail.com)

[s122926@guu.ru](mailto:s122926@guu.ru)

*Abstract*—Clustering problem is a task of dividing a set of objects (also called members) into different groups (called clusters) based on object’s characteristics. Members of a group will have more similarities in comparison with those in other group. This report discusses a traditional clustering method called K-Means algorithm from mathematical perspective. Additionally, an experiment is provided to examine the algorithm in two-dimensional space then an application in image compressing.

1. Introduction

The goal of K-Means Algorithm is to correctly separating objects in a dataset into groups based on object’s properties.

For instance, objects could be house and their properties are size, number of floor, location, power consumption per year, etc. The goal is to classify house dataset into groups which are luxury, average, poor. In that case, all properties of houses have to be processed to turn into number to create a vector, this process is called vectorization.

***Chart, scatter chart

Description automatically generated***Chart, scatter chart

Description automatically generatedAnother example, take each point in a panel as a object and each object has two properties which are x-axis and y-axis location. And input K = 4. The algorithm correctly finds the cluster. (Fig. 1.a)

1. Input: N = 100, K = 4 (b) Output

Fig. 1: K-Means on 2-dimensional Points

1. Mathematical analysis
2. Input and Output

The K-Means Algorithm takes a set pf observations X = [x1, x2, …, xN] R­dxNwhere each observation is a d-dimensional vector. N is the number of observations (members) and the number of group (K, K<N) as two inputs. The algorithm outputs the center of K group [m1, m2, …, mk] R­dxN and the index or name of group that each member belonged to (label)

1. Lost Function and Optimization Problem

Suppose xi (i [1, N]) belong to cluster k (k [1,K], the lost value of observation xi is the distance from observation xi to center m­k in Euclidean space, defined by (xi – mk). Let’s yi = [yi1, yi2 …, yiK] be the label vector of each observation xi, yik = 1 if xi belongs to group k and y­ij = 0 j K

Label vector of each observation contains only one digit 1 because each observation belongs to only one group which leads to the following equation.

The objective is to minimize the within-cluster sum of squares (variance), also known as square errors of, where each square error of an observation xi from group mk is defined by:

||xi – mk||2 = yik ||xi – mk||2(2)

From the equation 1, sum of all elements in a label vector is equal 1. The square error of an observation is:

The square error of all observation is the sum of every square error of in the given set of observation. The goal is minimize the lost function, equation 4 where Y = [y1, y2, …, yN] be the matrix contains all label vector of N observation and M = [m1, m2, …, mk] be the center of K groups (clusters).

The objective is also to find the center and label vector of each observation which are Y and M, the two outputs that are mentioned in II-A.

Y, M = argminY,M (5)

1. Solving Optimization Problems

There are two variable in equation 5 which are center of each group of observation and label vector of each observation. The problem could be solved by fixed each variable then minimize the other variable.

*1/ Fixed M, center of observation group*: Because all centers (M) are constant, the objective is to correctly identify label vector which is identifying the group that each observation belonged to so that the square error in equation 4 is minimized.

yi = argminyi ||xi – mj||2 (6)

Retrieving from equation 1. Because only one element in vector yi, i [1, K] = 1. Equation 6 could be rewritten as:

j = argminj||xi – mj||2 (7)

The value of ||xi – m­j||2 is the square of distance from observation to center of group in Euclidean space. Concretely, when M is constant, equation 7 shows that minimizing the sum of square error could be achieved by choosing label vector so that the center are closest to observation.

*2/ Fixed Y, label vector of each observation*: When label vector (Y) is constant, the objective is to correctly identify the center so that the square error in equation 4 is minimized. In this case, the optimization problem in equation 5 could be rewritten by the following equation.

mj = argminmj ||xi – mj||2 (8)

The equation 8 is a convex function and differentiable for each i ∈ [1, N]. Hence equation 8 could be solved by finding the root of the partial derivative function. This approach will make sure that the root is the value that make the function reach a optimum.

Let’s g(mj) = ||xi – mj||2 (retrieving from equation 8 and take the partial derivative of g(mj):

(9)

The equation 9 is equal 0 is equivalent to:

mj  =

⬄ mj =

The value of yij = 1 when observation xi belongs to group mj. Hence, the denominator of equation 11 is the number of observations that belonging to group mjand the nominator is the sum of all observations belonging to group mj.

In other word, when Y is constant, the square errors could be minimize by assigning the centers to the means of observations in the groups that the observations belonging to.

1. Algorithm summary and Flowchart

*1/ Summary*: The algorithm can be done by continuously constative Y and M, one each a time as discussed in II-C1 and II-C2.

Step 1: Cluster the data into k groups where k is predefined.

Step 2: Select k points at random as cluster centers.

Step 3: Assign objects to their closest cluster center according to the Euclidean distance function.

Step 4: Calculate the centroid or mean of all objects in each cluster.

Step 5: Repeat steps 2.

*2/ Flowchart*: The following chart describe K-Means Algorithm

START

Number of group (K), Set of observations (X)

True

Distance from observation (Xi) to group center (M)

Center being Moved

False

End

Grouping based on minimum distance in Euclidean space

Fig. 2: K-Means Algorithm Flowchart

1. Discussion

*1/ Convergence*: The algorithm will stop after a certain number of iteration because the square error function is a strictly decreasing sequence and the square error is always greater than 0. But this algorithm will not make sure that it will find a global optimum because solving the equation 8 by finding the root when the partial derivative is equal 0 will only return the value for local optima but not make sure that local optima will be a global minimum

The following figure describe a case where poorly seeding leads to a local optimum.

Chart, scatter chart

Description automatically generated

Fig. 3: Poorly Seeding K-Means

2/ *Sensitiveness to initial cluster*: K-Means algorithm requires careful seeding, which means the final result is very sensitive to the initial value of cluster. Numerous efforts have been made to improving K-Means clustering algorithm due to its drawbacks

1. Application in Data Compression

An experiment will be reperformed where K-Means algorithm is applied to reduce the size of image and outputs a new image without the smaller number of color as compared to the original one. Each pixel of a image contains three elements which are red, green, blue (RGB) value. Let’s each pixel be the observation (X) then the number of pixel in an image be the number of observations. Each observation has three properties which are RGB value. In this case, K-Means algorithm is applied to identify K main colors in that image.

A picture containing grass, tree, outdoor, plant

Description automatically generated

A tree in a field

Description automatically generated with medium confidenceA tree in a field

Description automatically generated with low confidencea/ original b/ K = 32

c/ K = 16 d/ K = 8

The table below shows how file size of the original image is reduced.

|  |  |
| --- | --- |
| K (Color) | File size (kB) |
| Original image | 2550 |
| 32 | 1560 |
| 16 | 895 |
| 8 | 279 |

1. Conclusion

K-Means Algorithm could be very simple and quick to be implemented, the clustering problems where all clusters are centroids and separated can be solved by the algorithms. However, it will not be effective when the dataset and clusters are more complex.

This report doesn’t come with new idea to improve the effectiveness of the algorithm, the aim of the report to introduce the reader to a basic, entry level clustering methods with some visual examples on 2-dimensional and 3-dimensional dataset.

1. Further Research

The algorithm is simple to implement, however, the sensitivity to initial centroids, and strict structure of dataset, etc. are those drawbacks of the algorithm which are undeniable. Further research could be made to improve the value of initial centroids. The traditional algorithm is depended on randomness, research could also be made to discover a way to make a fixed initial centroid. Furthermore, on a large dataset, the algorithm could be very slow to converge. Research could be spent to make the iteration stops earlier.

1. Acknowledgment

The authors would like to thank Prof. Andrew Ng for inspirational machine learning course on Coursera

1. References
2. Joaquin Perez Ortega, Ma. Del, Rocio Boone Rojas, and Mara J. Somodevilla “*Research issues on k-means algorithm: An experimental trial using MATLAB*”
3. Arthur, D., Vassilvitskii, S., 2016. k-Means++: “*The Advantages of Careful Seeding*”. Technical Report, Stanford.