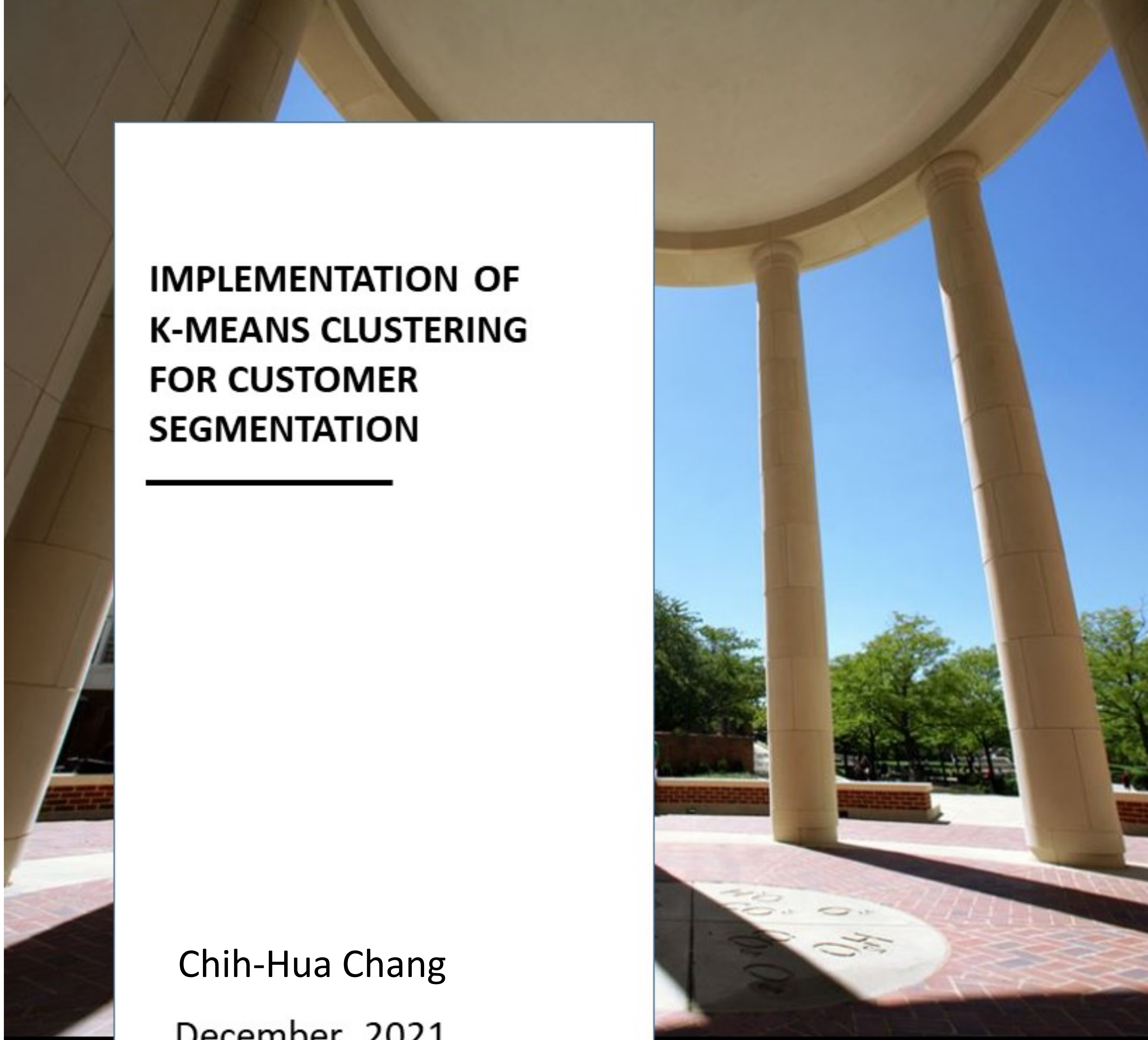


IMPLEMENTATION OF K-MEANS CLUSTERING FOR CUSTOMER SEGMENTATION

Chih-Hua Chang

December, 2021

Principles of Machine Learning
Machine Learning
Data Science and Analytics
University of Maryland at College Park

Introduction

My project is to find the clustering of the data by using k-means algorithm. The first part is preprocessing, which is to drop off the index, normalize the data and add a bias column. Second, split the data set into training part and testing part. Plot the graph between the costs of different clusters and plot the graph between different features. Decide to divide the data into five clusters and randomly select five points in the training part as the center of clusters. Calculate the distance between all the data and the each of the center point. Assign the data points to its nearest center and calculate new centers of the clusters. Repeat the process until the center points stay the same and cost stop changing. Plot the graph of different features in 2D to see how the clustering look like. Then, test the data. Calculate the distance between testing data and each of the final centroid, then assign the points to the groups with the same closest centroids. After assigning all the testing data, calculate the cost and compare it to the training data.

Background

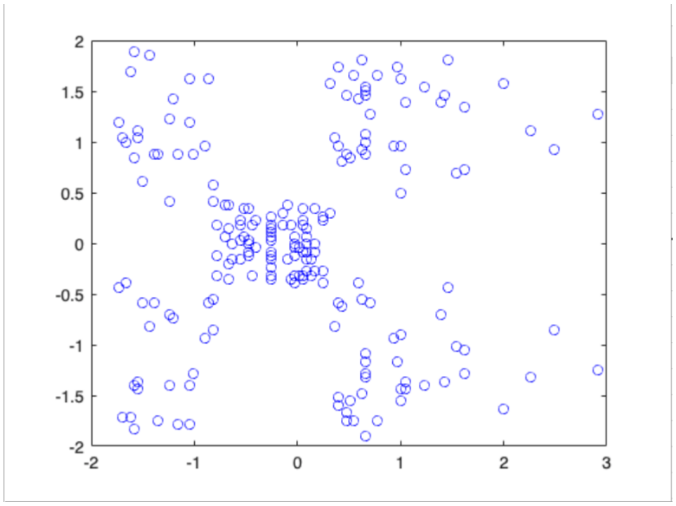
K-means algorithms is a non-supervised learning algorithm. The algorithm clusters n data to k clusters based on the distance between the data and centroids. To begin the clustering, first look at the cost between different clusters and find the elbow point of cost to decide the cluster number we are using or graph of the raw data and see if there is initial number of clustering group. Once decide the number of groups, which is k, set k random instances in the training data to be the centroids. Then, calculate the distance between all the data and each of the centroid. Assign each of the data point to its nearest centroid group. Next, calculate the means of each group and set them as the new centroids. Repeat calculating the distance between data points and new centroids until the center points stay same and the mean square error cost stop changing. Last, test the new data. Calculate the distance between new data and each of the final centroids and assign cluster the new data to k groups. Note that if the cluster number increase, the cost decrease, and as k gets larger and larger, it turns meaningless cause there are too many clusters in the data. Hence, select the elbow points in the graph of different k and cost to be the desire cluster number. Choose elbow points to prevent doing meaningless clusters. If the plot of the clustering does not look good, try different centroids for several times.

Implementation

There are four parts in my code, the first part is the preprocessing of the data, the second part is to see the cost between different numbers of clusters, the third part is to train the data and find the group centroids, and the last part is the testing the data.

(1) To run the code, first, open the script “data603_project2_preprocessing.m”. The first part is importing the data. Then drop of the index, normalize all the feature, and add bias column.

Next, visualized feature4 and feature5, which is annual income and spending scores, and find out there are about five clusters.



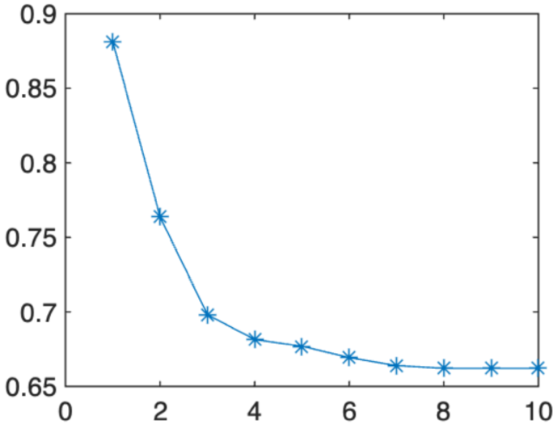
The graph shows that there are five clusters in the data.

Split the data to training part and testing part randomly and save the training data and testing data as “Train.csv” and “Test.csv” respectively.

(2) Open the script “data603_project2_findclusternumber.m”.

The script calculates the costs of different cluster number, k, in training data.

*Be careful: if the script shows error, try more times, because sometimes there won’t be instance in every group, and it will affect the calculation of centroid and cost.



Hence, select the elbow point in the graph, which is five, to be the clustering number.

(3) Open the script “data603_project2_train.m”.

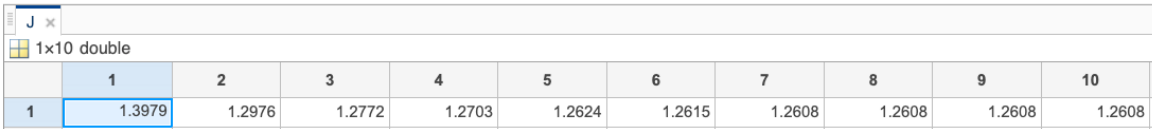
Import the training data, “Train.csv”. The Train size is [m*n]

Select five random instances in the training data and set them as initial centroids, which is the matrix “centroids”.

While iterations less then 10, for i=1:m, do the norm of the subtraction of i row in training data to each of the centroid to calculate the distance between data points and centroids. For example, for the first training row, subtract the first row of “centroids”, which is the first centroid, and do the norm function. Group the five results of distance for i row as an array, dists, and set the index of the minimum number in the dists as “group_id”. Group the data in a cell function by “group_id”.

Calculate the means of the five groups and set them as new “centroids”. Calculate the mean square error cost between the fives groups and their centroids (new centroids). Repeat the whole process until the cost converges and the centroids stay the same.

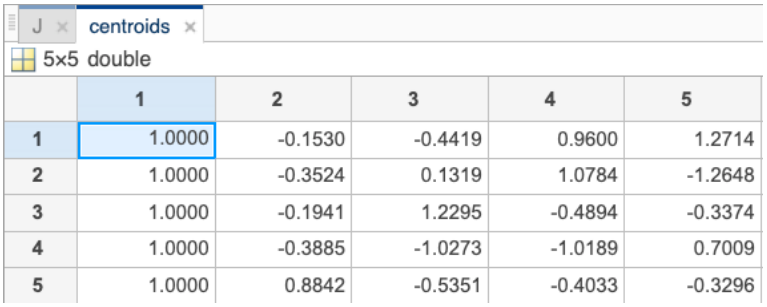
The cost for training data dropped from 1.3979 to 1.2608 in 7 iterations then converge.



1x10 double

	1	2	3	4	5	6	7	8	9	10
1	1.3979	1.2976	1.2772	1.2703	1.2624	1.2615	1.2608	1.2608	1.2608	1.2608

Five final centroids after the cost converged.

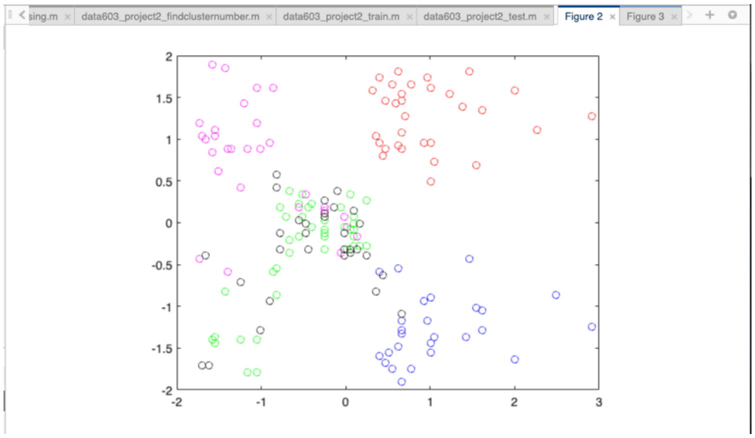

J ×
centroids ×

5x5 double

	1	2	3	4	5
1	1.0000	-0.1530	-0.4419	0.9600	1.2714
2	1.0000	-0.3524	0.1319	1.0784	-1.2648
3	1.0000	-0.1941	1.2295	-0.4894	-0.3374
4	1.0000	-0.3885	-1.0273	-1.0189	0.7009
5	1.0000	0.8842	-0.5351	-0.4033	-0.3296

Plot the grouping

2D graph of annual income in x axis vs spending score in y axis in training data

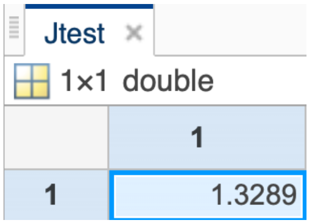


(4) Test the data.

Calculate the distance between the testing data and the five final centroids and assign the points to their nearest centroids.

Calculate the MSE cost of testing data.

The final test cost is 1.3289

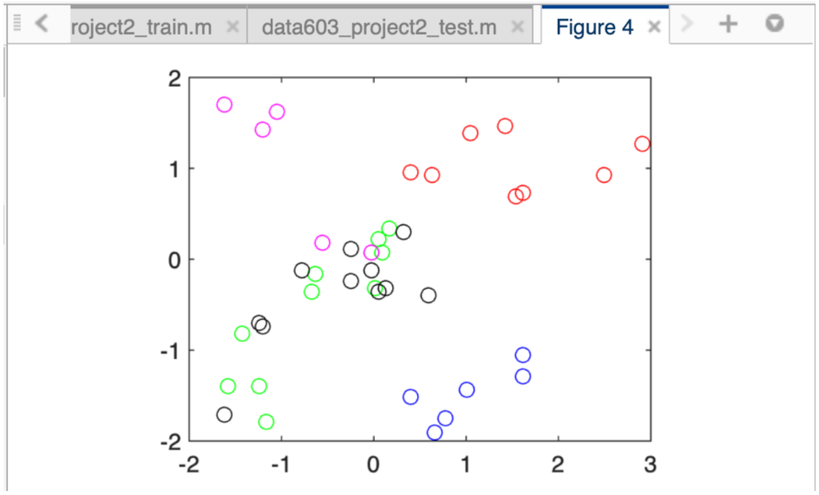


Jtest

1x1 double

	1
1	1.3289

2D plot of the test data.



Conclusions

In conclusion, there are five groups in the data, which are low-income with low-spending, low-income with high-spending, moderate income with moderate spending, high-income with low-spending, and high-income with high-spending.

Bibliography