Data Mining

Second Assignment

Kourosh Parand
Computer Science, Shahid Beheshti University
November 11, 2022

For the second assignment of data mining course, you should use your previous preprocessed dataset to cluster by several methods, find best parameters for clustering and write down your observations and characteristics of each cluster. First let’s review dataset features here:

ID	Unique ID of each customer
Year_Birth	Customer’s year of birth
Education	Customer's level of education
Marital_Status	Customer's marital status
Income	Customer's yearly household income in USD
Kidhome	Number of small children in customer's household
Teenhome	Number of teenagers in customer's household
Dt_Customer	Date of customer's enrollment with the company
Recency	Number of days since the last purchase
MntWines	The amount spent on wine products in the last 2 years
MntFruits	The amount spent on fruits products in the last 2 years
MntMeatProducts	The amount spent on meat products in the last 2 years
MntFishProducts	The amount spent on fish products in the last 2 years
MntSweetProducts	Amount spent on sweet products in the last 2 years
MntGoldProds	The amount spent on gold products in the last 2 years
NumDealsPurchases	Number of purchases made with discount
NumWebPurchases	Number of purchases made through the company's website
NumCatalogPurchases	Number of purchases made using a catalog (buying goods to be shipped through the mail)
NumStorePurchases	Number of purchases made directly in stores

NumWebVisitsMonth	Number of visits to the company's website in the last month
AcceptedCmp3	1 if customer accepted the offer in the third campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the fourth campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the fifth campaign, 0 otherwise
AcceptedCmp1	1 if customer accepted the offer in the first campaign, 0 otherwise
AcceptedCmp2	1 if customer accepted the offer in the second campaign, 0 otherwise
Complain	1 If the customer complained in the last 2 years, 0 otherwise
Response	1 if customer accepted the offer in the last campaign, 0 otherwise

First thing you should do is using your previous assignment preprocessed dataset.

After loading your preprocessed dataset, you should do each task in list below separately.

List of Tasks:

A) Drop columns Year_Birth, Dt_Customer, day, Complain, Response, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Marital_Status, Status, Kids, Education, Kidhome, Teenhome, Income, Age, Family_Size

B) Plot heat-map of data correlation

C) Scale data using standard scaler

D) Fit T-SNE model with 2 components and 35 as perplexity and random_state=1

E) Apply PCA with random state=1

F) Apply K-means clustering method with 2,3,4,...,9 clusters and add its' distortions to a list and then use elbow method to decide which cluster size is optimum

G) Calculate and print silhouette score for 3, 4, 5, 6 clusters

H) Apply K-means on PCA data with cluster_size=3 and plot a figure which visualize data segmentation

I) Use describe function to describe number of each cluster in your clustering method

J) Plot box-plot of each cluster for each column and write a complete observation over each clusters and indicate special characteristics of each cluster

K) Do tasks H, I and J for 5 clusters with random_state=0

L) Now use K-Medoids method with number of clusters=5 and random_state=1 and do all tasks H, I and J

M) Draw dendrogram of data with single, complete and average linkage with euclidean, chebyshev, mahalanobis and cityblock distance metrics

N) Cluster data with agglomerative method with 3 clusters, affinity=euclidean and linkage=ward and repeat all tasks H, I and J

O) Cluster data and data_pca with DBSCAN and Gaussian Mixture Model methods with optimized parameters and repeat tasks H, I and J

P) Write and conclusion and recommendation on all methods and parameters you use and describe them as detailed as you can. You must answer these questions: 1)What are the most meaningful insights from the data relevant to the problem? 2)How do different techniques perform? Which one is performing relatively better? Is there scope to improve the performance further? 3)What model do you propose to be adopted? Why is this the best solution to adopt?

Good Luck