PCA Assignment

Q1) **Assignment summary**

A) Problem statement:-

In the problem statement it states that there is a NGO organisation named Internal Humanitarian NGO organisatinon. This organisation commited to fight poverty and providing people of backward countries with basic needs and relief during natural disasters and calamities. We have to analyse strategically the name of five countries which are requires aid essentially from the NGO organisation.

B) Tasks:-

To achive this task Performed clustering of countries. Because by doing clustering analysis we can get the countries which are essentially need of aid by analysing the features of dataset.

To achive the clustering I have done Principal Component Ananlysis for dimensionality reduction. Which will help the clustering methods for better analysation and representation.

By analysing screeplot I conclude to use five numbers of principal components for further analysis. Because by considering only five principal components we can describe 95% of data so our aim to use as much as possible less number of dimension. So from first five principal components we can achivw our requiremnets.

i) K-Means Clustering:

In K-Means cluster I used two methods i.e Elbow Curve method and Sillhoutte Score method I obtained the 4 number of cluster for clustering of dataset.

ii) Hierarchical Clustering:

In hierarchical clustering I have used singale linkage method and complete method and obtain 4 number of clusters for clustering of data set.

For both of the clusters we get same country names so from this results we can say that both the clusters have produced better results as per our expectetions.

Q2) Clustering

A) Comparing and contrast K-Means clustering and hierarchical clustering?

Ans. K-Means Clustering :-

In K-Means clustering I have used 2 types of methds

i) Elbow curve method

ii) Sillhoutte score method

In elbow curve method I have plot a graph between the values of cost function for each cluster and concider the optimal number of K at the elbow bend of the curve.

In sillhoutte score it shows the measure of how an object is to similar to its own cluster.

Hierachical Clustering:

In hierarchical clustering I have used aglomorative hierarchical clustering . And use single linkage and complete linkage.

If we compare two types of clustering methods we can find that K-means clustering can handle a large size dataset because it works linearly, but hierarchical clustering have a complex algorithm so K-Means clustering can be use for every size of dataset.

But K-means clustering hav eone drawback that we have to predefine the optimal number of clusters.

B) Briefly explain the steps of the K-Means clustering algorithm?

Ans).   K-Means clustering is a linearly performed clustering algorithm. To perform this clustering we have to follow some procedure

Step-1:

Data preparation. In this step we have to prepare dataset like removing outliers, handling missing values, scaling of data and PCA on data set.

Step-2:

Cheking clustering tandency meance the clusterablity of the data set. By the help of Hopekins method we can check the clustering tandency. This Hopekins method bsicaly works on KNN(K Nearest Neighbour) process.

Step-3:

Applying of different methods to obtain the optimal number of K (number of cluster). Here I have used Elbow curve method and sillhoutte score method.

Basically in the K-Means algorithm there is two steps

I) Assignment step:

In this step we are taking randomly the centroids as the number of predefined number of clusters. And then assigne tha data points centroids by using euclidean distance according to their property.

II) Optimization step:

In this step we are calculating mean of each cluster then reassign centroids to the mean of clusters.

Above two process are inner loop of algorithm which will continue till the centroid comes same with its prevoius centroid.

Before these process to obtain number of clusers we have to use K++ algorithm this will help us to calculate the optimal number of K by using Costfunction.

C) How is the value of “K” chosen in K-Means clustering? Explain both the statistical as well as the business aspect of it?

Ans. By statistical Aspects:

In K-Means clustering to obtain the optimal number of K we are using K++ algorithm this will us costfunction to choose the optimal number of “K”. But to use this K++ algorithm there are some methods which will give the optiaml number of ”K”. They are

I) Sillhoutte Score Method:

In this method it measures a datapoint how much same to its own cluster by using cohession and separation. Cohession means distance between datapoint and its own cluster centroid and separation means distance between datapoint and its neighbour cluster centroid.

II) Elbow Curve Method:

In this method it plots the graph between the sum of each costfunction of each cluster between those clusters and in that plot some elbow bends will be form

By Business Aspects:

To choose the number of cluster from the business aspects we have to plot average sillhoutte score againist the nuber of cluster. Because it shows the likelyness of datapoints to its own cluster and neighbour cluster we can choose the number of cluster which have highest average sillhoutte score.

D) Explain the necessity fro scaling/standardisation before perormimg clustering?

Ans. We know that a data set have a number of features which are having values at different scales if we use this dataset for clustering then the clustering model form a very bad clustering model with outliers and cluster will not form correctly. Result of this problem will impact on datapoints which will cluster in another cluster but having diffferent properties.

E) Explain the different linkages used in Hierarchical clustering?

Ans. There are 3 types of linkages present in hierarchical clustering

I) Single Linkage:

In single linkage method but it considers minimum euclidean distance between each datapoints from each other data points with minimum distance will form cluster first.

II) Complete Linkage:

In average linkage method but it considers maximum euclidean distance between each datapoints from each other data points with minimum distance will form cluster first.

III) Average Linkage:

In average linkage method but it considers average euclidean distance between each datapoints from each other data points with minimum distance will form cluster first.

Q3) Principal components analysis

A) Give atleast three applications of PCA?

Ans. Because of PCA will perform dimensionality reduction so some applications of PCA(principal component analysis) :-

I) Image Compression
II) Face recognisation
III) Hand written digit or letter recognisation
B)

B)Briefly discuss the 2 important building blocks of PCA- Basic transformation and variances as information.

Ans. The two important building blocks of PCA are

I) Chemometrics:

It consists of building blocks of chemometrics. Chemometrics is involving data analysis and analytical chemistry it consists of some mathematical solutions and formulas or tools which help to interpret the chemical structures and formulas.

II) Hand written digit or letter Recognisation:

It consists of some classifier technique with PCA analysis which will predict the hand written letter and digits.

C) State at least three shortcomings of using Principal Components Analysis?

Ans. Three main shortcomings of using Principal Components Analysis are

I) There maybe a chance of information loss
II) It needs data standardisation before perofrming PCA
III) The independant variable will be less interpretable after performing PCA