PCA&CLUSTERING – DIBYALOK BHUTIARAI

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.
b) Briefly explain the steps of the K-means clustering algorithm.
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
d) Explain the necessity for scaling/standardisation before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

a. Compare and contrast K-means Clustering and Hierarchical Clustering.

Ans:

- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

b. Briefly explain the steps of the K-means clustering algorithm.

Ans:
K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps :

1. Specify the desired number of clusters K : Let us choose k=2 for these 5 data points in 2-D space.

2. Randomly assign each data point to a cluster.

3. Compute cluster centroids

4. Re-assign each point to the closest cluster centroid

5. Re-compute cluster centroids : Now, re-computing the centroids for both the clusters.

6. Repeat steps 4 and 5 until no improvements are possible : Similarly, we'll repeat the 4th and 5th steps until we'll reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

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:
Elbow method is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of *K* increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the **elbow point**.

Clearly if elbow is forming at K=3. So the optimal value will be 3 for performing K-Means.

d. Explain the necessity for scaling/standardisation before performing Clustering.

Ans:

scaling/standardisation refers to the process of rescaling the values of the variables in your data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis. Standardization is important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

When you are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

e. Explain the different linkages used in Hierarchical Clustering.

Ans:

We use 3 linkages in Hierarchical Clustering:

Complete Link : In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest **maximum** pairwise distance).

Single Link: In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest **minimum** pairwise distance).

Average Link: Average Link is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.
b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.
c) State at least three shortcomings of using Principal Component Analysis.

a)Give at least three applications of using PCA.

PCA is predominantly used as a dimensionality reduction technique in domains like:

1. Facial recognition,
2. Computer vision
3. Image compression.

It is also used for finding patterns in data of high dimension in the field of finance, data mining, bioinformatics, psychology, etc

.

b)Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

1. Building Blocks of PCA - Basis of Space: A set of elements (vectors) in a vector space is called a basis, if every element of V can written in a unique way as a linear combination of elements of B . The coefficients of this linear combination are referred to as coordinates on B of the vector. The elements of a basis are called basis vectors. Equivalently B is a basis if its elements are linearly independent and every element of V is a linear combination of elements of B . A vector space can have several bases, however all the bases have the same number of elements, called the dimension of the vector space.

2. Building Blocks of PCA - Basis Transformation A basis for a vector space of dimension is a set of vectors (a1, a2…..an), with the property that every vector in the space can be expressed as a unique linear combination of the basis vectors.

Since it is often desirable to work with more than one basis for a vector space, it is of fundamental importance to be able to easily transform coordinate-wise representations of vectors and operators taken with respect to one basis to their equivalent representations with respect to another basis. This process of converting the information from one set of basis to another is called basis transformation.

c) State at least three shortcomings of using Principal Component Analysis.

1. Independent variables become less interpretable: After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.

2. Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.

For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.

3. Information Loss: Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.