ASSIGNMENT PART – II

Question – 2 Clustering

a)

K-Means algorithm is the process of dividing the N data points into K groups or clusters. Here
the steps of the algorithm are:
1. Start by choosing K random points the initial cluster centres.
2. Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
3. For each cluster, compute the new cluster centre which will be the mean of all cluster members.
4. Now re-assign all the data points to the different clusters by using the new cluster centres.
5. Keep iterating through step 3 & 4 until there are no further changes possible.

Hierarchical Clustering proceeds a bit differently from the K-means Clustering method in the following ways
Given a set of N items to be clustered, the steps in the hierarchical clustering are:
1. Calculate the NxN distance (similarity) matrix, which calculates the distance of each data point from the other.
2. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item.
3. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
4. Compute distances (similarities) between the new cluster and each of the old clusters.
5. Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.
Generally, for large datasets, it is preferred to used K-means clustering whereas for smaller ones

As you can clearly see in hierarchical clustering, you don't need to specify the number of clusters as you do in K-means clustering. This is one of the key differences between K means and Hierarchical Clustering.

Generally, for large datasets, it is preferred to used K-means clustering whereas for smaller ones we use Hierarchical Clustering. The reason for this is that Hierarchical Clustering is computationally expensive. In each iteration, it runs on every cluster that has been formed previously and store them in the memory as well. Therefore, it uses a lot of RAM and if one has limited memory bandwidth, then it becomes a problem to get good clusters.

K- means, on the other hand, runs iteratively each time without any burden on the memory, since only the new K centroids need to be stored(given that K is available to use from the start).Since it would only compute new distances in each step(by assigning each point to its nearest cluster) and not store them in the memory for further comparisons, K- means is much faster to use.

b) Steps of K-Means Algorithm

K-Means algorithm is the process of dividing the N data points into K groups or clusters. Here
the steps of the algorithm are:
1. Start by choosing K random points, the initial cluster centres.
2. **Assignment Step**: Assign each data point to their nearest cluster centre, i.e. the cluster centre having the minimum Euclidean distance from the data point.
3. **Optimisation Step**: For each cluster, recompute the new cluster centres which will be the mean of all cluster members that were assigned in the previous step.
4. Now re-assign all the data points to the different clusters by using the new cluster centres.
5. Keep iterating through step 3 & 4 until there are no further changes possible.

c) The value of K can be chosen in 2 ways. Either by using Statistical Analyses or by using the business aspect. Both are explained below

Statistical Analyses

You can use either the Silhouette score or the elbow curve to find the optimal value of K. At this value, the most stable clusters are formed. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.
The elbow curve method uses the sum of square distances method to find the most optimal clusters at the point where there is not much change in the inertia, i.e. where distinct elbows are being formed.

Business Aspect

This is mostly oriented for situations where the business understanding of the problem dictates as to how many clusters need to be created. For example, if a food delivery company wants to create 5 hubs in a city, it needs to find the 5 most optimal points as to where it needs to be located so that the fastest delivery can happen. So in this case you need to create 5 clusters, irrespective of whether the statistical analyses give a different result.

d) Standardisation is the process of converting all the columns in any dataset to a comparable scale before applying the method of clustering. This is done to offset the effect of very large-scale variables on those having low scale.

For example, if you want to cluster a list of employees on the basis of the salary that they earn and the avg. hours that they work per day, then you’d find that the salary column would have a scale that is at least a thousand times more than the hours column. Now, clustering algorithms use Euclidean distance to compute the similarity between two points, this measure would be heavily influenced by the salary column only due to its larger scale. Therefore, to make sure that the employees are getting clustered properly using both the columns, we use standardisation - to bring the mean of each column to 0 and standard deviation 1.

e) **Types of Linkages in Hierarchical Clustering**

Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters

Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

Average Linkage: Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

Question 3 Principal Component Analysis

a) Applications of PCA

- Dimensionality reduction: The first and foremost application of PCA is to reduce the number of dimensions before applying any prediction model over it which leads to a faster execution.
- Data visualization and Exploratory Data Analysis: It can be difficult to visualise multidimensional data on a 2-D surface. Hence performing PCA will help in summarising the features and using the first 2 principal components, we can visualise the dataset.
- Create uncorrelated features/variables that can be an input to a prediction model: Uncorrelated features lead to a stable model as you don’t have to worry about Multicollinearity anymore.
- Uncovering latent variables/themes/concepts: PCA also helps in uncovering latent themes in the dataset. For example, in a dataset containing user ratings on thousands of movies, PCA can help in summarising those movies using certain themes like genres, etc.
- Noise reduction in the dataset: Since we generally remove the low variance components while performing PCA, it leads to significant reduction of noise and helps in identifying patterns in the dataset.

b) The two fundamental building blocks of PCA are:

  - **Basis Transformation**: The first fundamental building block is the idea of transforming the original standard basis that we have for our dataset to a whole new basis. Sometimes, it becomes easier to explain the data to others using a new basis. For example, polar coordinates are sometimes useful in communicating distances rather than the normal latitude/longitude version.

Additional Note: Also, we can choose a set of basis vectors in such a way that some of those directions would be explaining most of the variance/ information of the dataset and some would be explaining very less. This enables us to perform dimensionality reduction by eliminating those directions which are not explaining much information about the dataset.

- **Variance as information**: The information in a dataset is captured by the variance of the columns that we have present there. This enables us to find the directions which have more variance and as a result, more information. For example, a column having the same values throughout all the rows is of less significance since it would have a zero variance. Therefore, the most important, or the most informative columns in a dataset would be the ones with the most amount of variance.

c) PCA has the following shortcomings

- PCA is limited to linearity and therefore if a non-linear model produces a better solution then the latter method is more preferred.
- PCA needs the components to be perpendicular, though in some cases, that may not be the best solution.
- PCA assumes that columns with low variance are not useful, which might not be true in prediction setups (especially classification problem with class imbalance)