**WEEK 11&12**

1. Cluster analysis has been recognized as a significant data mining task with many applications. Our textbook presents some specific examples, organized by whether the clustering's purpose is understanding or utility. Give one application example (other than the ones given in our textbook) for each of the cases: a) Clustering for Understanding and b) Clustering for Utility.

Clustering for Understanding :

To understand the outer space better, scientists have based mathematical formulas and spectral configurations to classify and understand the phenomenon. Gamma Ray bursts are used to determine the classes of events taking place in the cosmos. Depending upon the Gamma Ray bursts we can know if a star is born or dying , dying in a supernova or a hypernova, if the sudden burst is a quasar or a black hole etc. To obtain the most success in the discovery, effort is put in to find the dissimilarity in the events to be able to group them efficiently and if there is something unusual than the previous occurrences group them separately or subset it.

Clustering for Utility:

Clustering analysis can be used along with the ordination techniques to find the hidden pattern in the data. Ordination techniques arrange the similar datapoints closer together and the dissimilar datapoints farther apart. They use multiple linear least squares regression method. However when used along with the clustering algorithm , the closer datapoints are grouped together to find the hidden patterns in the dataset

1. One of the required parameters for the K-means algorithm is a number of clusters (K value). Discuss how you would come up with the K value other than a random number.

There are several methods to choose the value of “k” instead of assigning a random k-value. Elbow method is the most common method, here we compute the sum of squared error for some of the random values of k and then plot graphs between these values of k and their respective SSE’s. At the point where there is a abrupt decrease in the value of SSE for the corresponding value of “k”, that value is selected as the k value for your algorithm initialization.

Another method is Silhouette method, however it is more of a quality measurement of the clustering and is used in combination with the Elbow Method. Here we calculate the Silhouette Coefficient using the mean intra-cluster distance and mean nearest-cluster distance of each sample and the values vary between -1 to 1 with 1 being the best value and -1 being the worst. Similar to Elbow method a graph is computed between the Average Silhouette coefficient and corresponding “k” values. The optimal K-value selected is the one where it shows higher Average Silhouette coefficient to the number of clusters in the graph.

The above K methods are Direct methods, in addition to these there is third method to compute the k-value known as Gap Statistic Method. This method assumes null hypothesis of the dataset i.e dataset has no obvious clustering and then graphs the user-defined k-values to the expected-values. Gap is computed by subtracting the different k value with the expected value. The Expected k-value is computed through the bootstrapping method.

1. The K-means algorithm minimizes the total squared distance from instances to their cluster centers. Unfortunately, it's a local minimum, not a global minimum. How can you minimize this problem?

Local minima is the minimum within the clusters and is usually computed using the larger clusters. Global minima is the minimum value globally for all clusters. This problem cannot be eliminated however to reduce these optimization issues we can use the Gradient Descent method as mentioned in the book. The Gradient Descent technique finds the direction in the hyperplane which provides the steepest descent in the value of the loss function. To perform descent along the direction of the Gradient is known as Gradient Descent. The Gradient Descent function then decides the learning rate i.e the speed it should move in the direction it decided to find the actual minimum. Depending on the speed if we overspeed , we could miss the minima and if we drag then the training of the dataset will take far too-long to complete and could get stuck at local minima. Thus after deciding on the speed and direction we continue our descent in the gradient , however we never reach the exact minima. We then continue to run these iterations in hope of reaching the minima but we keep oscillating around the predicted area. Hence after multiple repetitions if the minima value hasn’t changed much the process is stopped depending upon the number of folds decided previously.

1. Suppose you have decided to get a K value (a number of clusters) for the K-means algorithm from a dendrogram generated by an Agglomerative Hierarchical Clustering algorithm. How would you cut the dendrogram to pick an optimal K value? Discuss why.

Agglomerative Hierarchical Clustering uses bottom to top approach and each dataset is considered a cluster in itself. Measuring the similarity between these objects it combines these clusters into a single cluster and continues to do so until all the objects are a one big cluster. This method does not directly compute the cluster value instead it uses the cluster distance cutoff value to find the clusters. The default value is 1.0 . The Sequence number of the fusion know as Linkage is used with its corresponding distance value to calculate the number of clusters. For example if the Linkages 5-6 in a dendrogram have a distance value of approximately ~ 1.0 then we see the corresponding value of the resulting clusters . This is the optimal k-value that would be used if the distance is 1.0 or higher.