Name – Bbiswabasu Roy

Roll – 19EC30008

Assignment – 3

Preprocessing data:

The features in the dataset can be broken into 3 categories:

- Binary categorical features
- Non-binary categorical features
- Numerical features

Among them numerical features can be left as it is because they represent scalar values of features. Binary categorical features can also be left as they are because similarities between 0 and 0 is more than that between 0 and 1 and similarities between 1 and 1 is more than that between 1 and 0.

However, the non-binary categorical values cannot be used directly to build the model and hence one hot encoding was applied to convert them to binary categorical features. For example, the feature "Education" takes values {0, 1, 2, 3} and after applying one hot encoding 4 binary features were added namely "Education_0", "Education_1", "Education_2", "Education_3" and the column "Education" was removed. If the "Education" column had value of 2, then "Education_2" was assigned value 1 while others were assigned value 0. Same principle was applied on other non-binary categorical features as well.

Also, the ID column in the original dataset was useless and hence removed.

Finally, normalization was applied on all columns to scale the values appropriately using the formula:

$$data[i,j] = \frac{data[i,j] - \min_i(data[i,j])}{\max_i(data[i,j]) - \min_i(data[i,j])}$$

Where, $data[i,j]$ is the value in j-th column of i-th row.

Clustering Algorithm:

Following is the brief description of the algorithm used:

1. Choose some k (number of clusters)
2. Choose initial centroids to be random k points from the dataset
3. For each point in dataset do the following:
4. Find the centroid at minimum distance from this point
5. Assign this point to same bin as the centroid at minimum distance
6. Take the mean of all points in each of the k bins
7. These k mean points form the new set of centroids
8. If the average distance moved by the k centroids is very small, then terminate
9. Else, go to step 3

Euclidean distance was chosen as distance metric, both for forming the clusters and for performance evaluation.

Performance evaluation:

There are several metrics to quantify the performance of clustering algorithm like Inertia, Silhouette score, etc. For this assignment scatter index was used to measure the performance which is defined as:

$$Scatter\ index = \frac{Average\ Intracluster\ distance}{Average\ Intercluster\ distance}$$

Since calculating distances between all pair of points was found to be computationally expensive, centroids were considered to be representative of each cluster and following approximations were used:

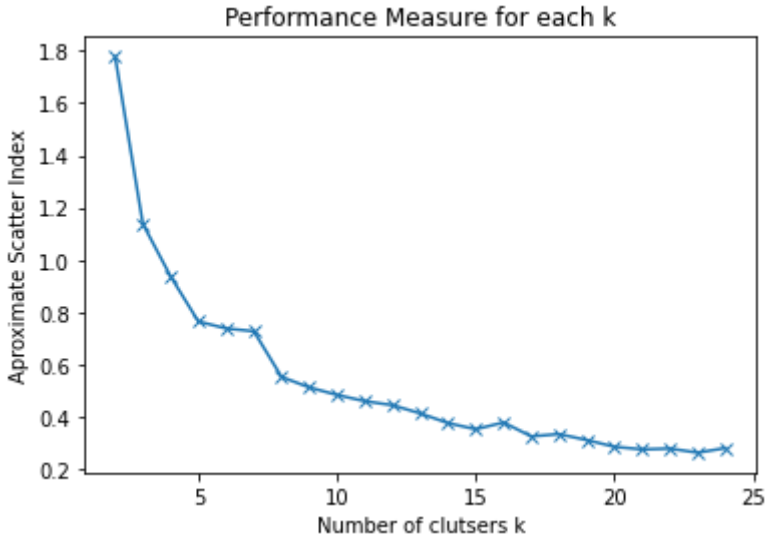
$$Average\ Intracluster\ distance = \frac{1}{N}\sum_{c=1}^{k}\sum_{data[i]\in c} \|data[i] - centroid[c]\|$$

Essentially, instead of computing distance between each point in each cluster, it computes the distance of each point in data set to the centroid of its own cluster and takes the average. Euclidean norm was taken for distance computation. Similarly,

$$Average\ Intercluster\ distance = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} ||centroid[i] - centroid[j]||$$

Instead of computing distance between all points of different clusters, distance between centroids of each pair of clusters were added and then their average was taken.

This led to the following plot which had sharp decrease for small values of k and very less decrease after certain value of k. Since the decrement of scatter index was not significant after k=15, hence 15 was chosen to be the number of clusters.



Conclusion:

- Euclidean distance (L_2 norm) was used as distance metric. However other distance metrics could also have been used and might have yielded somewhat different results.
- It was found that with increasing k, scatter index decreased. However, after certain point the decrement was not significant and that point was chosen to be optimal k.
- There are various clustering quality performance measures but in this problem, the scatter index as defined above was used. Other measures might have yielded somewhat different value of optimal k.