Identifying fraudulent Taxpayers using Spectral Clustering

Team:
Puneet Mangla (CS17BTECH11029)
Yash Khasbarge (CS17BTECH11044)
Rushikesh Tammewar (CS17BTECH11041)

Objective:

- Here we implement Spectral clustering to detect fraudulent taxpayers . In spectral clustering we create graphs from a dataset , and the datapoint considered as a vertex is connected to the nearest datapoint . K means clustering only considers distance but spectral clustering takes in account both distance as well as connectivity.

- Created a graph from data set with edge weight represents gaussian similarity, With edge $W_{ij} = exp(-\left\|x_i - x_j\right\|^2)$ and each vertex is connected with its most similar 5 vertex . finally made sure that the created graph is undirected.

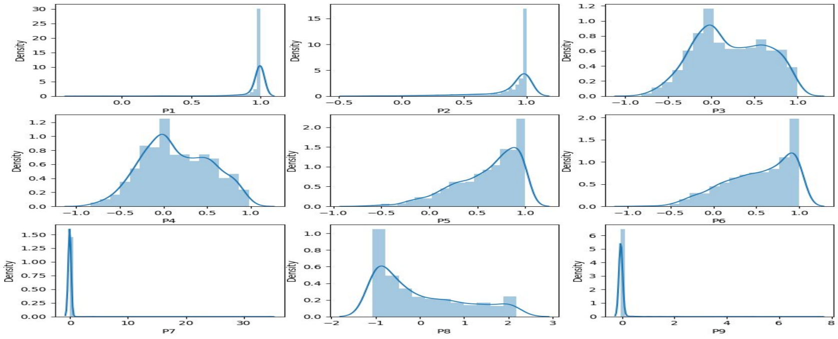
- As we have to find a cluster such that the cut is minimalized .We have to find a normalized min cut. We will use a normalised laplacian which is a positive semidefinite matrix.

- Eigenvalues of normalised laplacians shows how many clusters are in data . count of zero eigenvalues gives number of connected component

- We will find vertex in each component using k-means clustering on reduced eigenvector space. Here we have used a spectral biparition algorithm to find fraudulent taxpayers.

Dataset summary:

- We have 1163 sample with 9 feature and distribution of feature across the sample is :



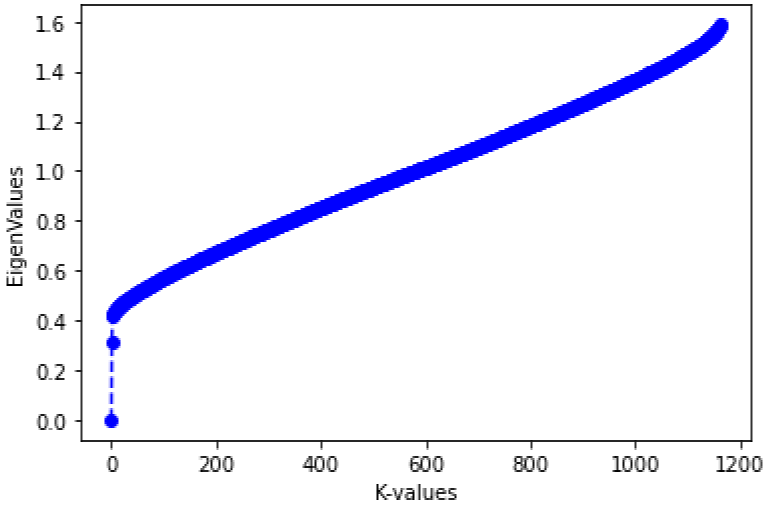
- Mean of sample is :
  [0.959524 , 0.85616583 , 0.2161109 , 0.14919403 ,0.60036861, 0.52484523,00098841, -0.00269966 ,-0.02701196]

- Variance of sample is :
  [0.01632525, 0.05948979 ,0.16666798, 0.15047038 ,0.11192182 ,0.14930412, 1.03006199,0.99289213, 0.08219917]

- As there are some outliers in the dataset we have to normalize the dataset .

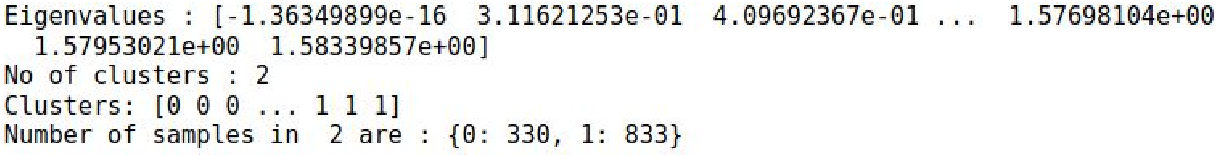
Experimentation results:

- Plot of eigenvalues of normalised laplacian:



- This plot shows that the graph is connected and there is only one complete connected component but the significance difference between the 1st and 2rd eigenvalue shows that we can classify data in 2 clusters. Value of K is 2.

- Output is:



```
Eigenvalues : [-1.36349899e-16  3.11621253e-01  4.09692367e-01 ...  1.57698104e+00
  1.57953021e+00  1.58339857e+00]
No of clusters : 2
Clusters: [0 0 0 ... 1 1 1]
Number of samples in  2 are : {0: 330, 1: 833}
```

- Finally we got 330 samples as fraudulent and clusters vector stores which sample is marked as fraudulent