A3 – COL761

Team Details:

Team Name: Data_Voyagers

Github link: https://github.com/rajasekhar108/data_voyagers

Members:

Name	Entry Number	Contribution
Bogam Sai Prabhath	2023AIB2079	33.33
Mikshu Bhatt	2023AIB2067	33.33
Nallavadla Rajasekhar Reddy	2023AIB2066	33.33

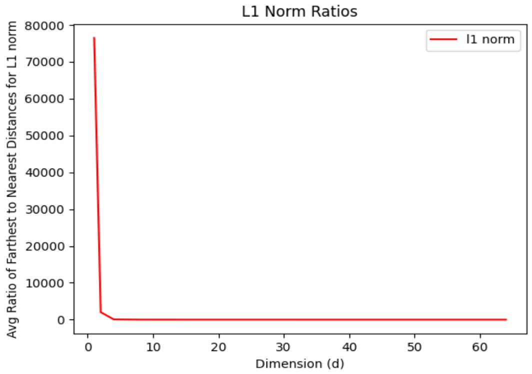
Q1) 1 Million random points are generated in the range of [0,1] which were distributed uniformly. For each query point selected from the generated points L1, L2, L-infinite norms are found out.

Observed Trends:

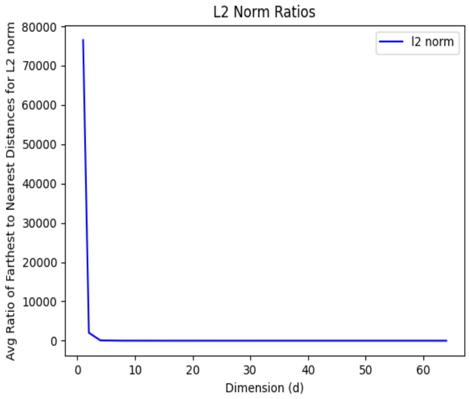
- Initially, when the dimension is 1 all norms will be same. As dimensions are increased , the data points will become more sparse (distance increases). This can be understood as follows: for 1st dimension all the points will be on a single line and the variation will only be in 1 dimension. When it is increased to 2, there will be variation in the values w.r.t 2 dimensions and so the distance between them increases. When we keep increasing the dimension, the distance keeps increasing between the data points as the volume increases w.r.t new increased dimension and also and the number of points still remain the same.
- This implies in low dimensions the points will be nearer to each other and after increasing the dimensions the distance also increases more on average.
- The n-dimensional volume of a Euclidean ball of radius R in n-dimensional Euclidean space is:

$$V_n(R) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} R^n$$

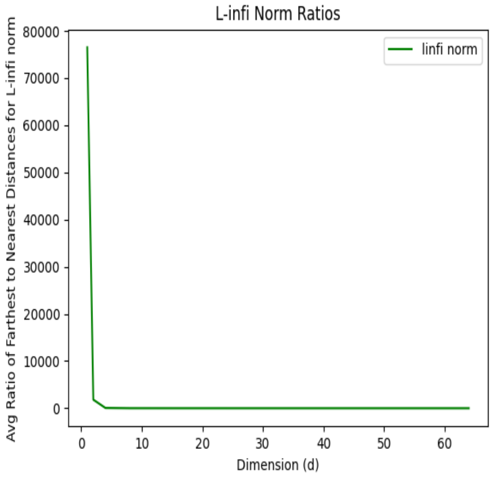
- In a hypercube of unit edge length, the number of points in a hyper sphere of radius r and dimension d will be proportional to (2r^d * π^d/2)/(Γ(d/2))
- The density of the points decreases exponentially with d the farthest distance will get increases only as power of ½.
- From the above theory we can deduce that, with increasing dimension the ratio of farthest to nearest points will decrease exponentially which can also be observed from the plots observed after running the code.
- The plots are as shown below:
- Plot for L1 Norm:



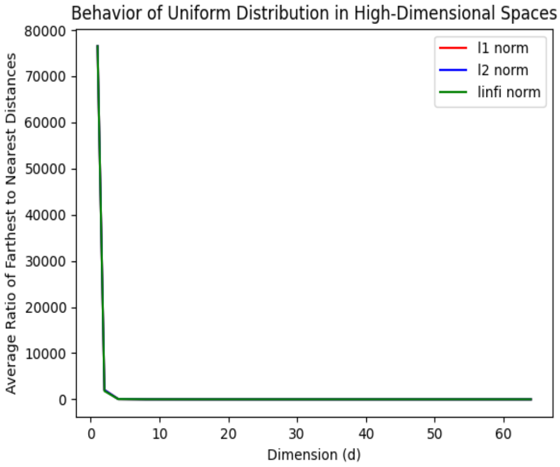
- Plot for L2 Norm:



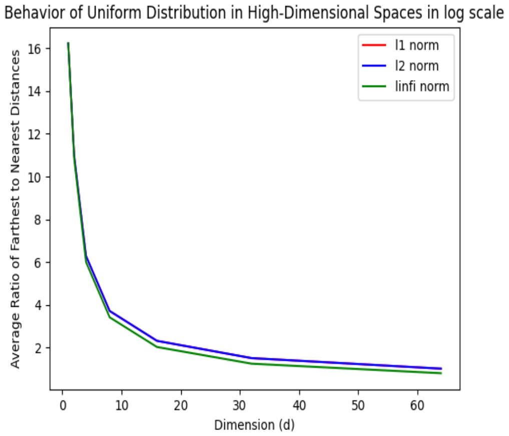
- Plot for L-infinity Norm:



- Plot for L1,L2,L-infinity norms combined:



- Plot for L1,L2,L-infinity norms in Log Scale(for better visualization):



Q2) When applying GNN for linear or logistic regression on large graph data it takes into account of the graph topology and other important features while baseline model doesn't take consider these features.

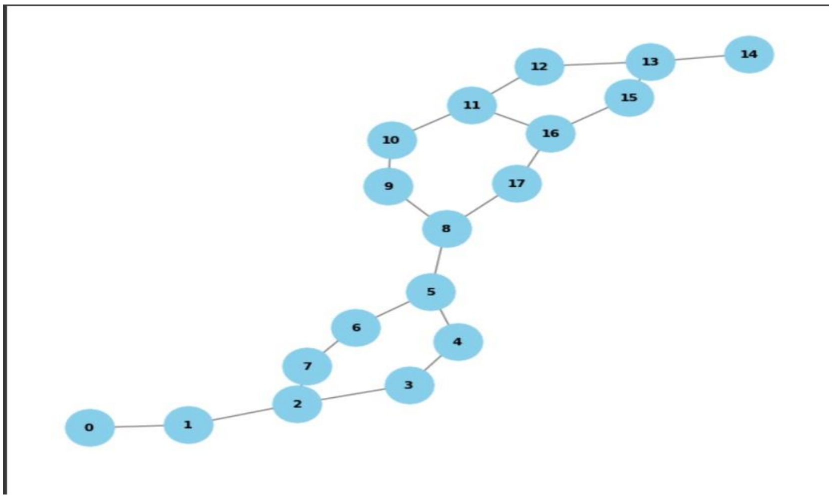
Using GNNs have several advantages:

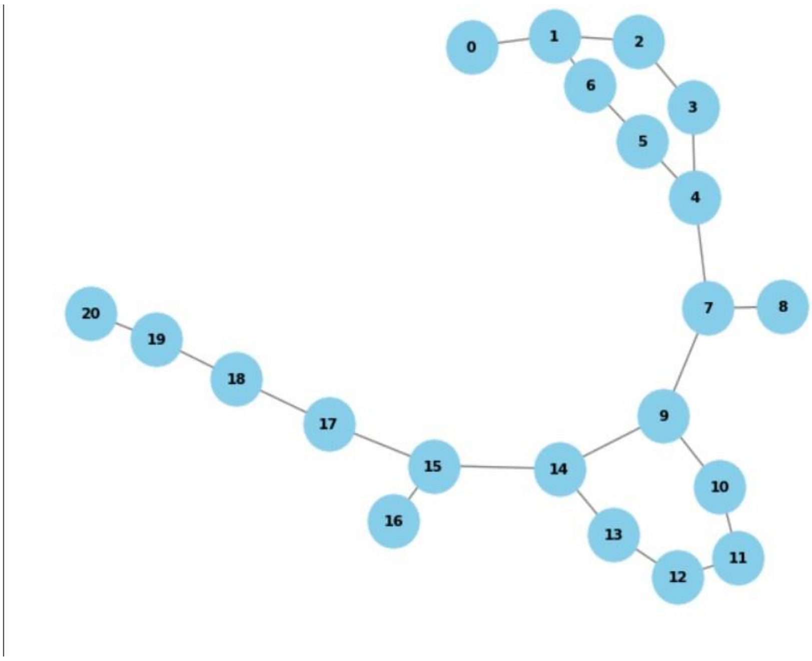
1. It incorporates graph structure: its topology and message passing.
2. Node embeddings
3. Neighbourhood information
4. End to End learning framework in GNN

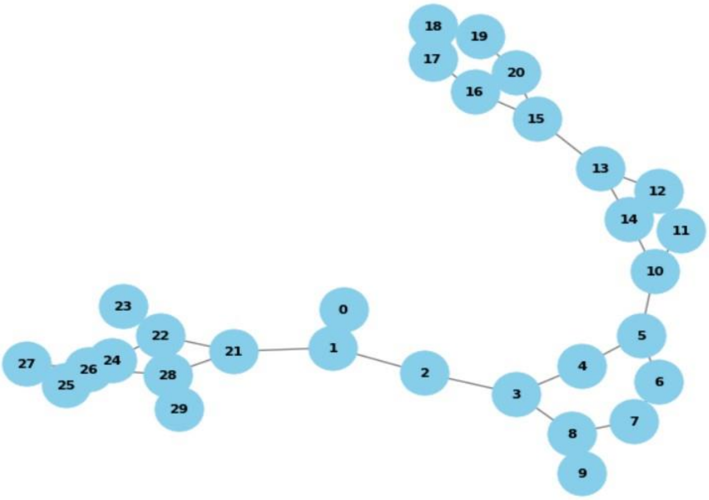
The GAT model has been used for both linear regression and classification task.

Classification:

In classification task the following graphs were observed to give maximum error when predicted using GNN.



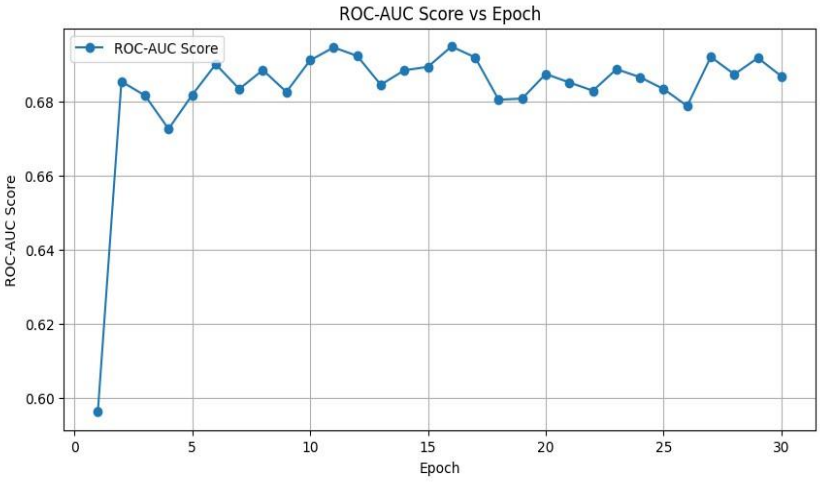


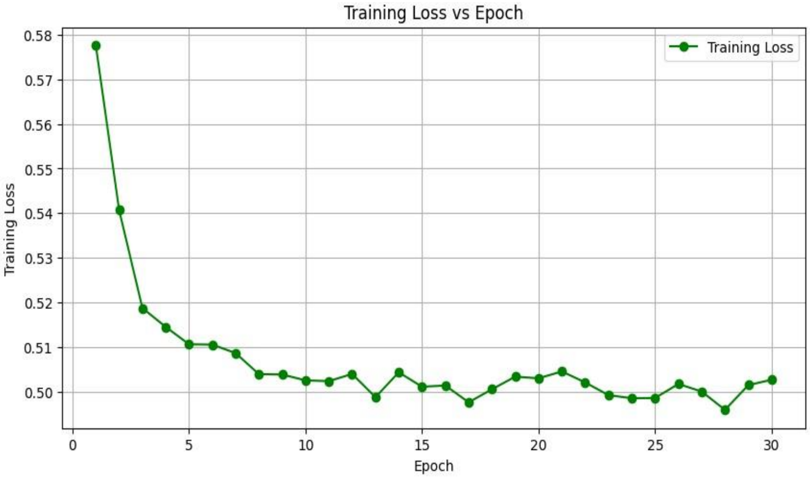


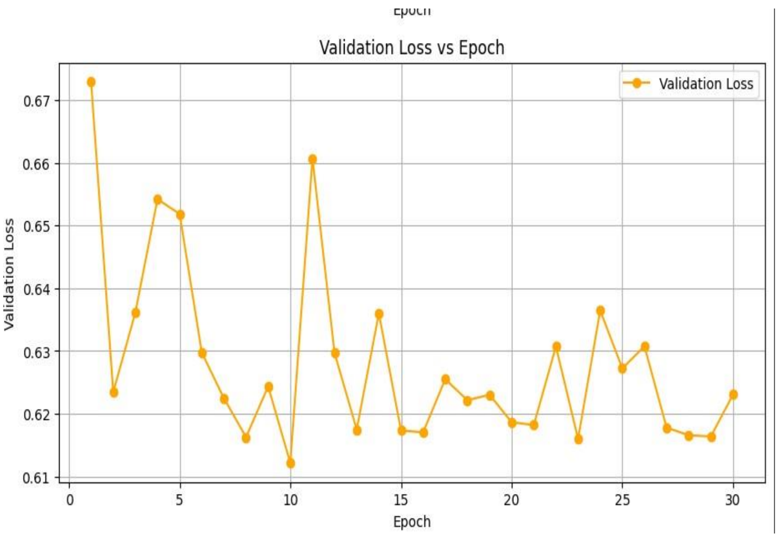
The following might be the reasons for misclassification of the above graphs:

- Irregular graph structure and graph size
- Irregular graph quality and topology
- Disconnected graphs

The plot for the loss during training is as below



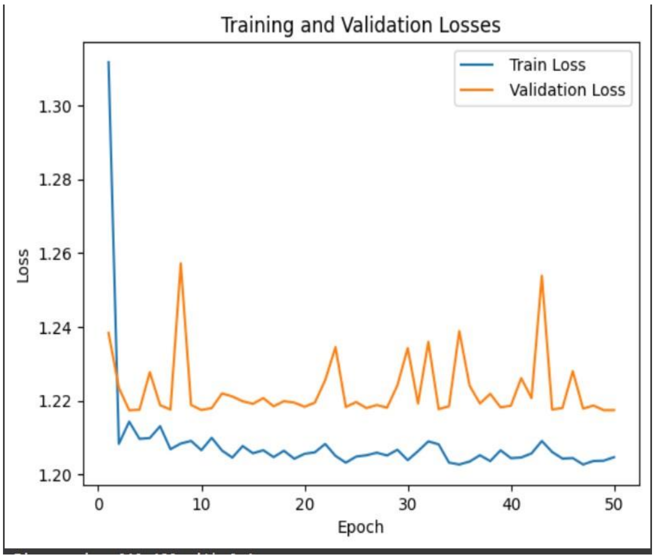


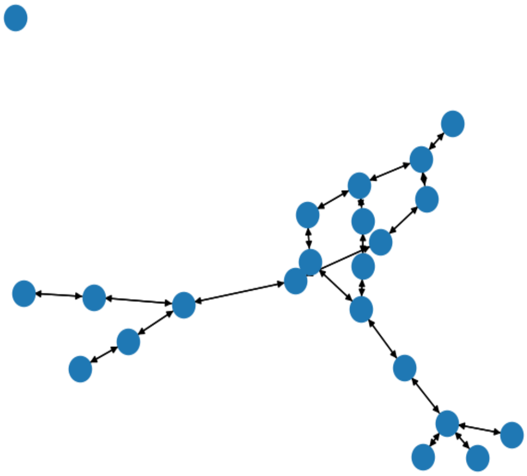


ROC-AUC for classification using GNN : 0.74

ROC-AUC for classification using Baseline model:0.53

Similarly, loss curve for regression





The model made significant errors while predicting for the following kinds of graphs:

- Irregular graph topology
- Disconnected graphs