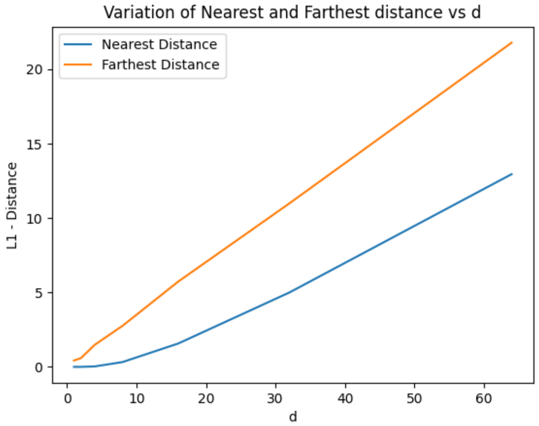
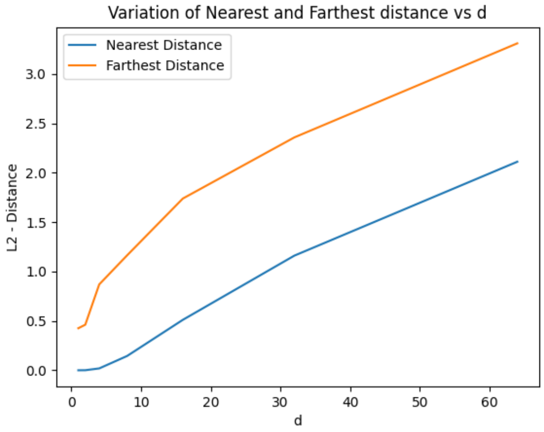
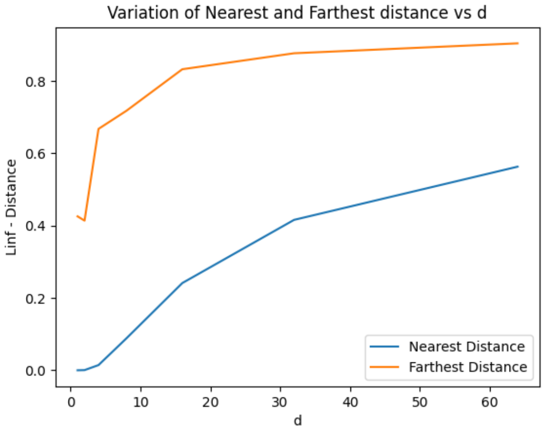
COL761-Data Mining HW3

Vonteri Harshith Reddy(2019CS50450)
Chapala Sriram Varma(2019CS50426)
Somisetty Harsha Vardhan(2020CS10390)

Q1.







From the curse of dimensionality, we know that with increasing dimensions the data points start to spread out, leading to data sparsity. In higher dimensions, the distance between every pair of points is almost the same. L1 distance measures the absolute sum of distances in every dimension. Hence, with increasing dimensions L1 distance increases. Linf is the maximum distance in any dimension. Hence, the distance in a dimension becomes constant as no of dimensions increases. L2 distance lies between these two extremes. We can observe this in the above plots.

Q2.

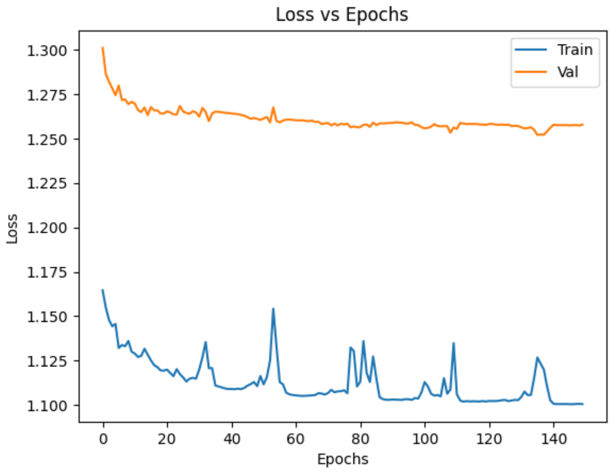
Task - 1: Classification:

Implementation Details:

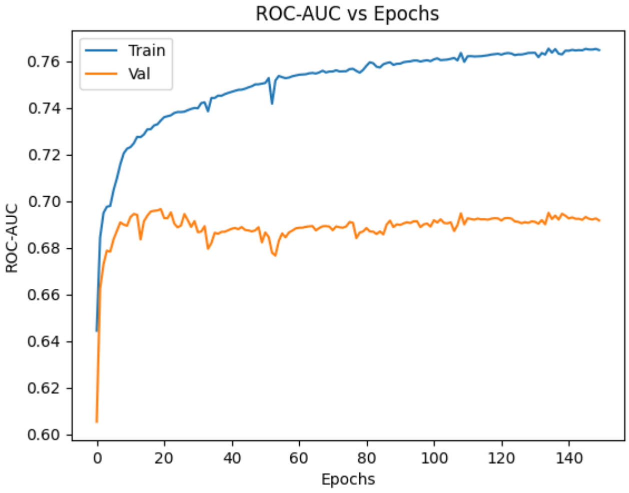
The nodes and edges are encoded with the encoder provided. Nodes are encoded to 16-dim space and edges are encoded to 2-dim space. We used 3 GAT Layers, each with hidden dimension 32. Then the embeddings for each node in each of 3 layers are concatenated. The embedding of the graph is obtained by applying a mean pool on the embeddings of each node. A linear layer is applied at the end to get a real value, whose sigmoid is the probability of belonging to class 1. We used BCE Loss and Adam Optimizer with a learning rate 0.01. As the given dataset is biased with 0 class graphs, we used loss function with weights to ensure effective learning. We saved the model with the best validation ROC-AUC score.

Learning Curve:

1. BCE Loss vs Epochs:



2. ROC-AUC Vs Epochs:

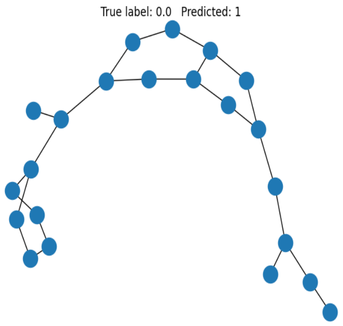


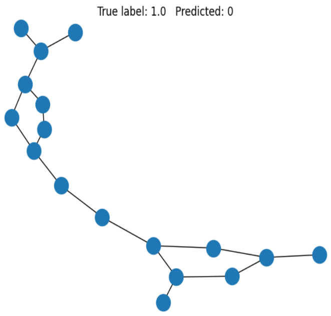
Comparison with Baselines:

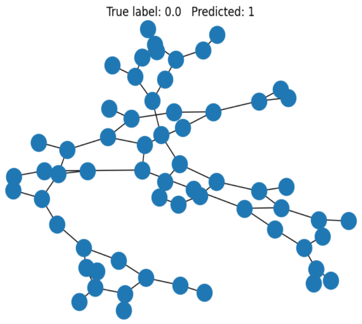
	Train ROC-AUC	Val ROC-AUC
GNN - Best Model	0.735	0.697
Logistic Regression	0.668	0.624
Random	0.504	0.469

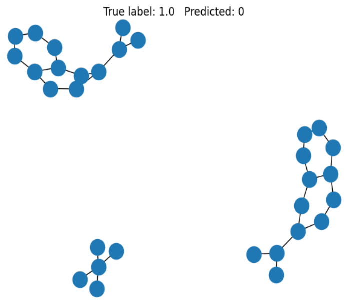
Analysis: We can see that our method performed better than the other baselines. From the learning curves, we can see that there is a steady increase in ROC-AUC with no of epochs.

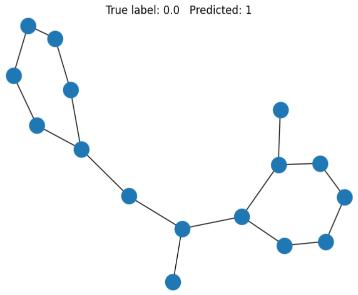
Misclassifications:

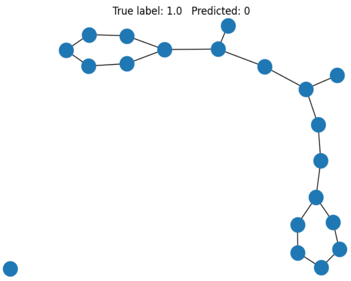












Reason: The above graphs contain multiple clusters of nodes which are connected with few or no links. Our model is not able to predict the true class for these types of graphs.

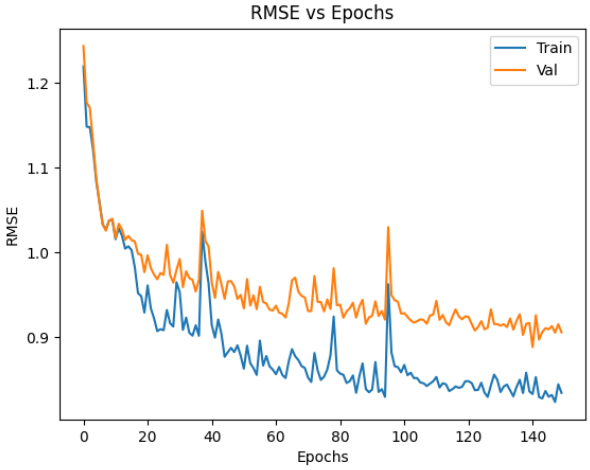
Task - 2: Regression:

Implementation Details:

The nodes and edges are encoded with the encoder provided. Nodes are encoded to 16-dim space and edges are encoded to 2-dim space. We used 3 GAT Layers, each with hidden dimension 32. Then the embeddings for each node in each of 3 layers are concatenated. The embedding of the graph is obtained by applying a mean pool on the embeddings of each node. A linear layer is applied at the end to get a prediction. We used MSE Loss and Adam Optimizer with a learning rate 0.01. We saved the model with the best validation RMSE score.

Learning Curve:

1. RMSE Vs Epochs:



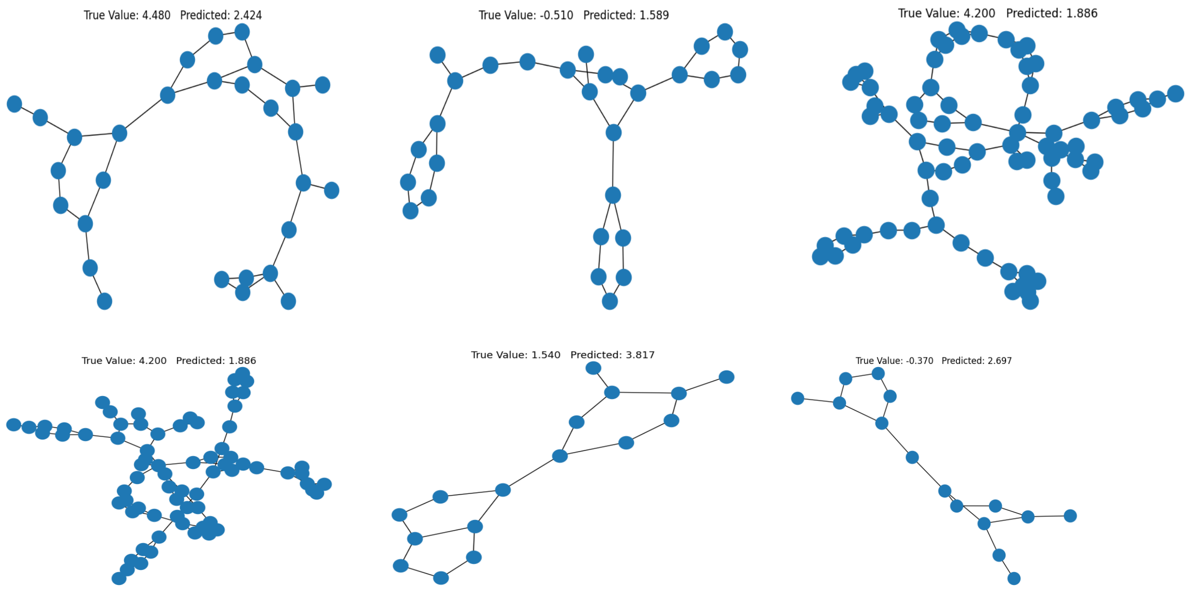
Comparison with Baselines:

	Train RMSE	Val RMSE
GNN - Best Model	0.834	0.905

Linear Regression	1.159	1.20
Random	2.07	2.12

Analysis: We can see that our method performed better than the other baselines. From the learning curves, we can see that there is a steady decrease in RMSE with no of epochs.

Mispredictions:


True Value: 4.480 Predicted: 2.424
True Value: -0.510 Predicted: 1.589
True Value: 4.200 Predicted: 1.886
True Value: 4.200 Predicted: 1.886
True Value: 1.540 Predicted: 3.817
True Value: -0.370 Predicted: 2.697

Reason: The above graphs contain multiple clusters of nodes which are connected with few or no links. Our model is not able to predict the true class for these types of graphs.

Contribution:

Vonteri Harshith Reddy - 33%
Sriram Varma - 33%
Somisetty Harsha Vardhan - 33%