Name: Tahereh Farivarnahid

Student ID: 401522010

Professor: Dr. Taheri

Course Project: Graph Mining

Learning molecular representation through graph-based approaches is a novel method designed for representing molecules in vector space.

In the implementation of this exercise, the PyTorch and DGL packages have been used for deep learning on graphs. Additionally, the Adam optimizer has been utilized for updating the model parameters. The Dropout method is employed to prevent overfitting issues. Throughout training, hyperparameter tuning, and testing, a batch size of 64 is used. Models are trained for 200 epochs and Early Stopping is considered with a patience of 10. The data is split into training, validation, and test sets using three identical random seeds.

The datasets used in the experiments are divided into two categories: classification and regression. The subsets of each category are as follows:

BBBP: A recently created dataset for modeling and predicting blood-brain barrier permeability (classification).

Lipophilicity: A dataset derived from the ChEMBL database, focusing on an important property of drug molecules that affects membrane permeability and solubility (regression).

In this exercise, four graph neural network architectures, namely GIN, GraphSAGE, GCN, and GAT, with two and three layers, are employed. Increasing the number of layers in the desired graph neural network improves performance. ReLU activation function is used in all four architectures. Furthermore, a feed-forward neural network with two and three hidden layers and ReLU activation function is used to generate feature predictions.

The performance evaluation of classification models is done through ROC-AUC (Area Under the Receiver Operating Characteristic curve), while regression models are evaluated using RMSE (Root Mean Squared Error).

Finally, the table below presents the model errors with different architectures for both classification and regression models:

Test score in a classification model:

GNN / Nomber Layer	GCN	GraphSAGE	GIN	GAT
2	0.634	0.727	0.612	0.530
3	0.585	0.616	0.605	0.464

Test score in a regression model:

GNN / Nomber Layer	GCN	GraphSAGE	GIN	GAT
2	1.372	1.323	1.389	1.328
3	1.370	1.294	1.301	1.310