Graph ML Final Project Report

Zahra Tebyanian

Professor: Dr. Zahra Taheri

Contents

0.1 Introduction

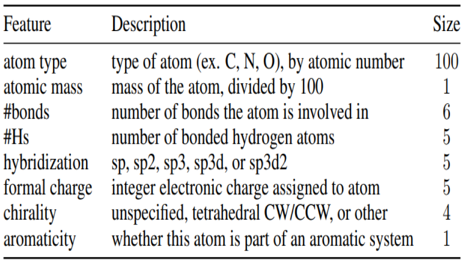
In this report, we investigated various neural network models to perform classification and regression on molecular data sets. Graph Neural Networks (GNN), which have been developed as a powerful candidate for modeling graph-structured data, are one of the hottest topics in recent years. A molecule can be naturally represented as a graph that In it, atoms (vertices) are connected by chemical bonds (edges).

0.2 Initial Atom, Bond, and Molecule-level Features of Datasets

We used BBBP dataset for classification and FreeSolv for regression. Each of these datasets have initial features.

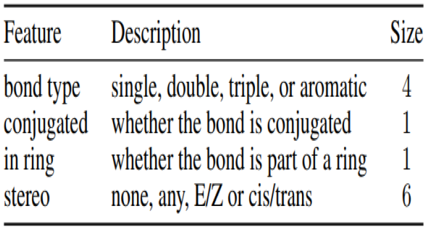
Three features are extracted and utilized for each molecule: atom, bond, and molecule-level features. All features are computed rapidly in silico using the open-source package RDKit.

The initial atom and bond features are listed in tables below. All features in these two tables are one-hot encodings except for atomic mass, a real number scaled to be on the same order of magnitude.



Feature	Description	Size
atom type	type of atom (ex. C, N, O), by atomic number	100
atomic mass	mass of the atom, divided by 100	1
#bonds	number of bonds the atom is involved in	6
#Hs	number of bonded hydrogen atoms	5
hybridization	sp, sp2, sp3, sp3d, or sp3d2	5
formal charge	integer electronic charge assigned to atom	5
chirality	unspecified, tetrahedral CW/CCW, or other	4
aromaticity	whether this atom is part of an aromatic system	1

Table 1: Initial Atom Features



Feature	Description	Size
bond type	single, double, triple, or aromatic	4
conjugated	whether the bond is conjugated	1
in ring	whether the bond is part of a ring	1
stereo	none, any, E/Z or cis/trans	6

Table 2: Initial Bond Features

Also 200 RDKit features are generated for the molecule-level features to capture the global molecular information. These RDKit features are concatenated to the output of the GNN's readout to go through a feed-forward neural network with two hidden layers utilized with the activation function ReLU to generate property predictions.

0.3 GNNs

To solve classification and regression problems, we implemented six different neural networks:

- model_1: A GCN with two layers.
- model_2: A GCN with four layers and batch normalizations and applying dropouts.
- model_3: A GraphSAGE with two layers.
- model_4: A GrapgSAGE with four layers and normalizations and applying dropouts.
- model_5: A Custom GNN with four layers and normalizations and applying dropouts.
- model_6: A Custom GNN with four layers and normalizations and applying dropouts, considering edge features in message passing.

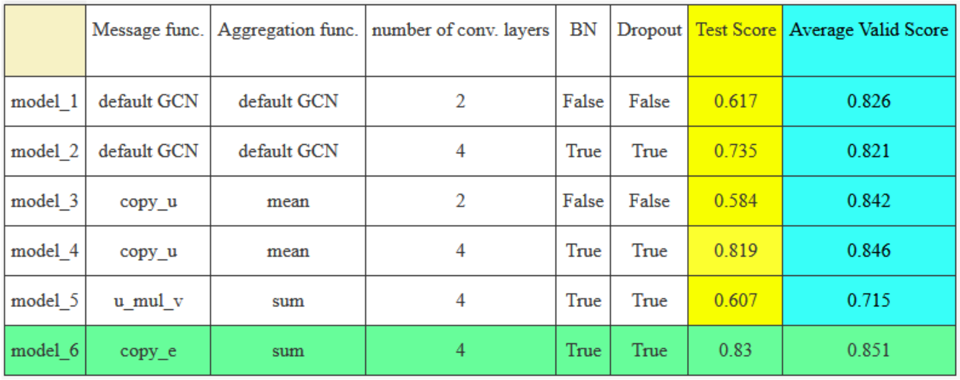
0.4 Graph Classification

Graph classification is the problem of determining the category or label of a graph. If we have a dataset consisting of a large number of input graphs, the problem is to classify each of the graphs into the correct target category or label.

We used BBBP dataset for classification task. BBBP (Blood–brain barrier penetration) dataset comes from a recent study on the modeling and prediction of barrier permeability. This dataset records whether a compound is permeable to the blood-brain barrier. The problem is as follows: predict whether a chemical compound is permeable to the blood-brain barrier or not. Therefore it is a binary classification problem.

0.4.1 Results

The results of each model based on the ROC-AUC metric are shown in the table below.



	Message func.	Aggregation func.	number of conv. layers	BN	Dropout	Test Score	Average Valid Score
model_1	default GCN	default GCN	2	False	False	0.617	0.826
model_2	default GCN	default GCN	4	True	True	0.735	0.821
model_3	copy_u	mean	2	False	False	0.584	0.842
model_4	copy_u	mean	4	True	True	0.819	0.846
model_5	u_mul_v	sum	4	True	True	0.607	0.715
model_6	copy_e	sum	4	True	True	0.83	0.851

Table 3: Classification Results

According to the table above, the best model is model_6.

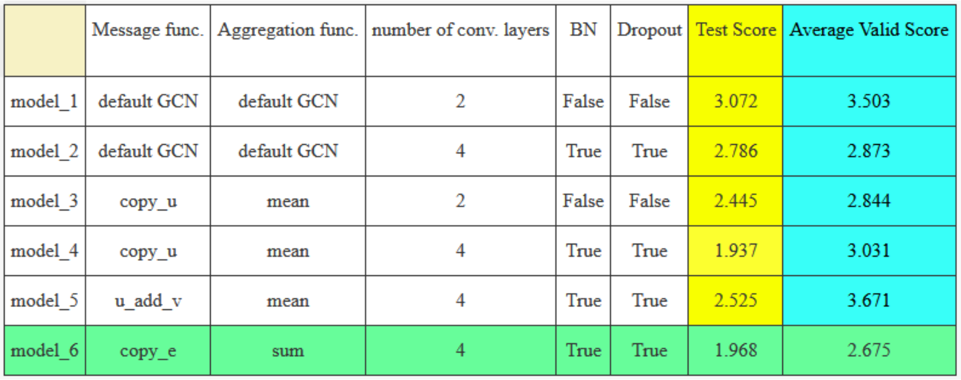
0.5 Graph Regression

Regression of graphs is similar to their classification and the difference is in the loss function and the performance metric.

We used FreeSolv dataset for regression task. FreeSolv is selected from the Free Solvation Database, which contains the hydration free energy of small molecules in water from experiments and alchemical free energy calculations. In fact, the goal is to predict the amount of hydration free energy of a small molecule in water.

0.5.1 Results

The results of each model based on the RMSE metric are shown in the table below.



	Message func.	Aggregation func.	number of conv. layers	BN	Dropout	Test Score	Average Valid Score
model_1	default GCN	default GCN	2	False	False	3.072	3.503
model_2	default GCN	default GCN	4	True	True	2.786	2.873
model_3	copy_u	mean	2	False	False	2.445	2.844
model_4	copy_u	mean	4	True	True	1.937	3.031
model_5	u_add_v	mean	4	True	True	2.525	3.671
model_6	copy_e	sum	4	True	True	1.968	2.675

Table 4: Regression Results

According to the table above, the best model is model_6.