ACMS 80770-03: Deep Learning with Graphs
Homework 3

Handed out: Friday, October 28, 2022

Due: Friday, November 11, 2022 10:00 PM

Notes:

- Software resources for this homework are available on course's GitHub repository. Only use the packages imported in the starter code for your implementations.
- **Don't fork the repository.** Download the code and add it to your current repository.
- Use the provided latex file to submit your written work.
- Submit your codes to Gradescope only through Bitbucket or GitHub.
- Upload your written work (PDF only) directly to Gradescope.
- Gradescope uses MOSS to detect similar submissions. Once detected, copying and pasting from classmates are regarded as a violation of the honor code.

Problem 1: QM9 database is a dataset of 133885 small, drug-like, stable organic molecules, that are constructed of a maximum of nine heavy atoms and consist of four heavy atom types, including carbon, oxygen, nitrogen, and fluorine. In the starter code, we are representing these molecules using hydrogen-depleted molecular graphs, in which hydrogen has been eliminated as its presence is determined using other atoms and bonds. Hence, only the four heavy atoms are labeled. For molecules with less than 9 atoms, we have added null nodes to the molecular graph, i.e. disconnected nodes that have a type "null".

In this problem, we aim to develop a GNN model that predicts chemical properties of the molecules. To that end, we use molecular graphs as input to the model and the one-hot vector of atom types as initial node attributes.

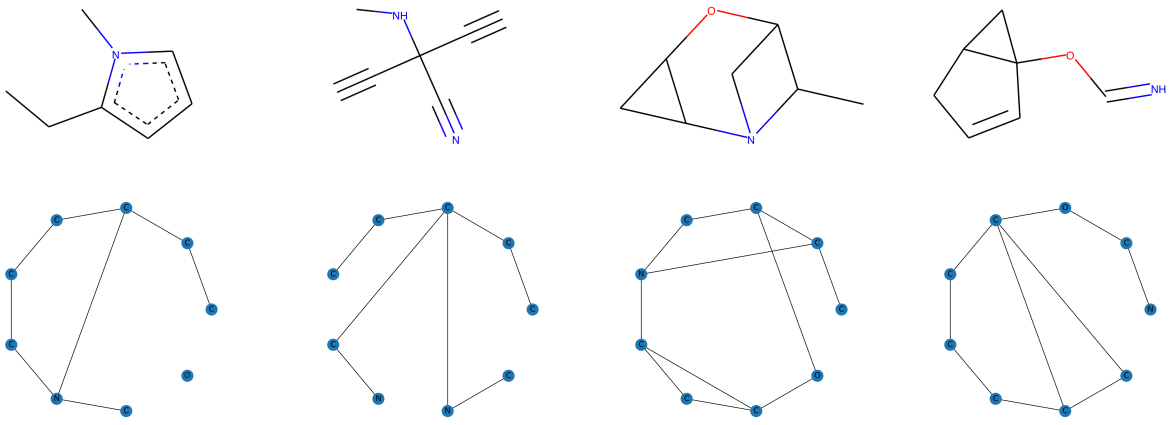

Figure 1: Examples of the Lewis representation of molecules and their corresponding molecular graph representations.

A. The Graph Convolutional Network (GCN) operates on the following propagation rule

$$\boldsymbol{h}_i^{(k)} = \sigma\left(\mathbf{W}^{(k-1)} \sum_{v_j \in N(v_i) \cup v_i} \frac{\boldsymbol{h}_j^{(k-1)}}{\sqrt{d_i d_j}}\right).$$

Rewrite this node-level rule using a graph-level equation.

B. Implement a python class that performs GCN propagation rule.

C. A multi-set pooling function f_p maps node embeddings $\boldsymbol{z}_i$ to a graph-level embedding $\boldsymbol{z}_G$,

$$\boldsymbol{z}_G = f_p(\{\{\boldsymbol{z}_i | \forall v_i \in V\}\}).$$

Implement a python class for graph pooling layer that uses sum as the pooling function.

D. Develop a neural network model to predicts Highest Occupied Molecular Orbital (HOMO) energy of the molecules. The model should contain a single layer of GCN, a pooling layer, and a fully connected layer. The GCN layer takes $\boldsymbol{h}_i^{(0)} \in \{0,1\}^5$ as input and returns $\boldsymbol{h}_i^{(1)} \in \mathbb{R}^3$ as output.

E. Train the model using 5000 training points using SGD. Change the batch size as needed. Plot the errors for 200 epochs (error on y-axis vs. epoch on x-axis).

F. Use 1000 test data points to evaluate the model. Show the results using a target (y-axis) vs. prediction (x-axis) scatter plot.

Problem 2: Recall that K iterations of the aggregation process in a neural message passing propagation rule uses an unfolded, height K subtree pattern rooted at node $v_i \in V$ to construct node embedding $\boldsymbol{h}_i^{(K)}$.

Consider the random graph in figure2.

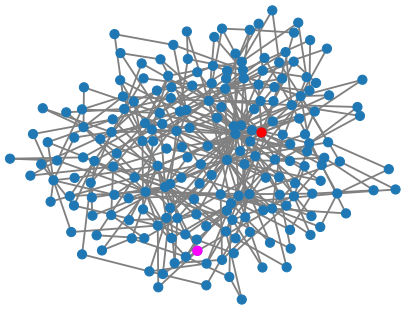

Figure 2: Random graph containing $|V| = 200$ nodes. Nodes v_{17} and v_{27} are shown in red and magenta, respectively.

A. Initialize node attributes with one-hot vectors of node indices.

B. Plot the effective range of nodes $v_{i=17}$ and $v_{i=27}$ for $K = 2, 4$, and 6 message passing layers.

C. Propagate the graph through 3 models with GCN layers (implemented in problem 1). The first model with dimensions 200-100, the second model with dimensions 200-100-50-20, and the third model has dimensions 200-100-50-20-20-20. For the output embeddings $\boldsymbol{h}_i^{(K)}$ of each model, compute the influence score I_K of nodes $v_j \in V$ on nodes $v_{i=17}$ and $v_{i=27}$,

$$I_K\left(v_j, v_i\right) = \overrightarrow{\mathbf{1}}^T \left(\frac{\partial \boldsymbol{h}_i^{(K)}}{\partial \boldsymbol{h}_j^{(0)}} \right) \overrightarrow{\mathbf{1}}.$$

D. In each setting, plot the graph with nodes colored with the influence score.

E. What do you conclude by comparing these plots?