Homework #5

MPS6 - Lab G

April 7, 2025

Instructions

Complete all five questions. Show your work for mathematical derivations and explain your reasoning clearly.

Question 1. [Easy] Graph Representation Basics:

Consider the molecule of Propionic acid (CH_3CH_2COOH).

(a) Draw the graph representation of Propionic acid, with atoms as nodes and bonds as edges. Label each node with its element symbol.

(b) Write out the adjacency matrix for this graph.

(c) List the node and edge features you would include for each atom and edge to effectively represent this molecule for a graph neural network.

Question 2. [Medium] Message Passing and Equivariance:

(a) Explain the concept of message passing in graph neural networks. How does information flow between nodes during this process?

(b) A key property of molecular GNNs is equivariance, specifically permutation equivariance. Explain what permutation equivariance means in the context of molecular representation learning and why it is important.

(c) Consider the molecule 1-propanol ($CH_3CH_2CH_2OH$). write out one iteration of a simple message passing operation where each node's feature is updated based on the sum of its neighbors' features (with added self-connections). Assume each node's initial feature is just the atomic number.

Question 3. [Easy] PyTorch Geometric Implementation:

(a) Write the PyTorch Geometric code to load the HIV dataset from MoleculeNet.

(b) How many molecules are available in this dataset?

(c) How would you extract the molecular features of a single compound from this dataset? What are the node features for each molecule?

(d) Implement a simple PyTorch transform to normalize the node features across the dataset.

Question 4. [Medium] Graph Convolutional Network Layer:

(a) Derive the mathematical formulation for a Graph Convolutional Network (GCN) layer.

(b) Implement a custom GCN layer in PyTorch that takes a graph with node features $X \in \mathbb{R}^{N \times F_{in}}$ and adjacency matrix $A \in \mathbb{R}^{N \times N}$ and outputs updated node features $X' \in \mathbb{R}^{N \times F_{out}}$. Can you further modify the implementation with residual connections?

(c) Explain how your implementation handles different graph sizes in a batch. What provisions would you need to make for molecules of different sizes?

Question 5. [Hard] Graph Neural Network Molecular Property Prediction:

For this problem, you'll design and evaluate a GNN for molecular property prediction.

(a) Design a GNN architecture for predicting aqueous solubility (logS) of drug molecules. Specify your choices for:

- Node features
- Edge features (if any)
- Type of GNN layers
- Number of message passing steps
- Readout function to aggregate node features
- Prediction head

(b) Write the full PyTorch training loop for your model, including:

- Loss function choice and justification
- Optimizer configuration
- Training/validation split approach
- Early stopping criteria

(c) Discuss two potential limitations of your model for this task and propose solutions to address them.

Submission Guidelines

- Submit your solutions as a single PDF file.
- Include code snippets where requested and explain your code.
- For mathematical derivations, ensure your steps are clearly shown.
- Due date: [11 April]