Graph classification and Graph regrssion

Sadaf Fatollahy

Contents

Abstract

To accurately predict molecular properties, it is important to learn expressive molecular representations. Graph neural networks (GNNs) have made significant advances in this area.Inthis project we focus on graph classification and graph regression.

1 introduction

In the context of molecular graphs, a graph is typically represented as a set of atoms (nodes) and bonds (edges) between them, where the atoms represent the elements that make up the molecule (e.g., carbon, hydrogen, oxygen), and the bonds represent the covalent bonds between the atoms. Graph neural networks (GNNs) can be used to learn expressive molecular representations from these graph structures.

In this project we have 2 dataset ,one of them is for classification task and the other is for regression task.first we need to know what they are!

2 Graph Classification

Graph classification is a type of machine learning task in which a model is trained to classify graphs based on their structural properties. Unlike traditional machine learning tasks that operate on fixed-dimensional data, such as images or text, graph classification involves working with data that is represented as a graph, which is a collection of nodes (vertices) and edges that connect the nodes.

In graph classification, each graph is typically represented as an adjacency matrix or an edge list, which captures the pairwise relationships between the nodes in the graph. The goal of the model is to learn a function that maps the graph representation to a categorical label, indicating the class to which the graph belongs.

Graph classification has a wide range of applications, including chemical compound classification, protein function prediction, social network analysis, and citation network analysis, among others. It is a challenging task due to the complexity of graph data and the difficulty in extracting meaningful features from the graph structure. However, recent advances in deep learning and graph neural networks have led to significant improvements in graph classification performance, making it an active area of research in machine learning and artificial intelligence.

3 Chemical compound classification

Chemical compound classification is an important application of graph classification in which machine learning models are trained to predict the properties of chemical compounds based on their molecular structure. In this context, each chemical compound is represented as a graph, where the nodes represent atoms and the edges represent chemical bonds between the atoms.

The molecular structure of a chemical compound can have a significant impact on its physical and chemical properties, such as solubility, reactivity, and toxicity. By training a machine learning model to classify chemical compounds based on their molecular structure, it is possible to predict their properties and identify compounds with desirable characteristics for specific applications, such as drug discovery or materials science.

Graph neural networks (GNNs) have proven to be a powerful tool for chemical compound classification due to their ability to model the complex relationships between atoms in a molecule. GNNs can learn to extract relevant features from the graph structure, such as the presence of certain substructures or the length and type of chemical bonds, and use these features to predict the properties of the compound.

3.1 Dataset

The dataset for this task is BBBP. 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. This dataset contains 2000 graphs or molecules with 127 vertex features, 12 edge features, and 200 molecular features, each of which is explained below.

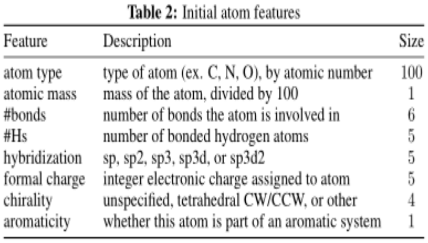

Table 2: Initial atom features

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

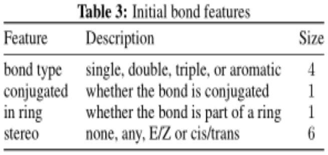

Table 3: Initial bond features

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

In this GNN we use of cross entropy as loss function and ROC-AUC as metric and node features and edge features . finally resaults are in below:

	GNN	Message	Aggregation	Dropout	Batch-n	Test score
Model 1	GCN (2 layer)	default GCN	default GCN	False	False	0.61
Model 2	GNN2	u-add-v	mean	False	False	0.58
Model 3	GNN2	u-mul-v	sum	False	False	0.65
Model 4	GNN2(4 layer)	u-mul-v	sum	False	False	0.74
Model 5	GNN2(4 layer)	u-mul-v	sum	False	True	0.82
Model 6	GNN2(4 layer)	u-mul-v	sum	True	True	0.60
Model 7	GNN2	copy-e	sum	False	False	0.63

Table 1: result of classification method on dataset

As you can see my best result is Model 5.

4 Graph Regression

Graph regression is a type of machine learning technique that involves predicting a continuous output variable from a graph-structured input. In graph regression, the input data is represented as a graph, where the nodes represent entities and the edges represent the relationships between those entities. The goal of graph regression is to learn a mapping between the input graph and a continuous output variable, such as a numeric value or a vector.

Graph regression can be used in a variety of applications, such as drug discovery, materials science, and protein structure prediction. In drug discovery, for example, graph regression can be used to predict the binding affinity of a candidate drug molecule to a target protein, given a graph representation of the molecule and the protein.

Graph regression models can be constructed using a variety of machine learning techniques, such as graph neural networks (GNNs), kernel methods, and decision trees. GNNs are a popular choice for graph regression because they are designed to operate on graph-structured data and can learn hierarchical representations of the input graph. GNNs can be trained using various loss functions, such as mean squared error (MSE) or mean absolute error (MAE), and can be optimized using standard backpropagation techniques.

4.1 Dataset

The dataset for this task is ESOL. ESOL is a small dataset consisting of water solubility data for some compounds. Like the previous dataset this dataset contains 2000 graphs or molecules with 127 vertex features, 12 edge features, and 200 molecular features.

In this GNN we use of MSE as loss function and RMSE as metric and node features and edge features . finally resaults are in below:

	GNN	Message	Aggregation	Dropout	Batch-n	Test score
Model 1	GCN (2 layer)	default GCN	default GCN	False	False	2.61
Model 2	GNN2	u-add-v	mean	False	False	2.24
Model 3	GNN2	u-mul-v	sum	False	False	3.18
Model 4	GNN2(4 layer)	u-mul-v	sum	False	False	2.00
Model 5	GNN2(4 layer)	u-mul-v	sum	False	True	3.15
Model 6	GNN2(4 layer)	u-mul-v	sum	True	True	3.15
Model 7	GNN2	copy-e	sum	False	False	2.45

Table 2: result of regression method on dataset

As you can see my best result is Model 4.