

Ejercicio GNNs en la práctica

Predicción de moléculas que inhiben el VIH



Introducción

En esta practica entrenaremos una red neuronal para predecir si un compuesto es capaz de inhibir la replicación del VIH. Para ello utilizaremos el dataset de MoleculeNet¹ llamada "HIV". Este dataset está compuesto de:

#graphs	#nodes	#edges	#features
41,127	~25.5	~54.9	9

Cada grafo representa una molécula, los nodos son elementos químicos y las aristas enlaces entre ellos. Los nodos tienen los siguientes atributos:

1. número atómico del átomo (por ejemplo, 6 para carbono, 8 para oxígeno)
2. quiralidad del átomo.
3. grado total del átomo.
4. carga formal del átomo (por ejemplo, -1, 0, +1).
5. número total de hidrógenos asociados al átomo.
6. número de electrones radicales asociados al átomo.
7. estado de hibridación (sp, sp², sp³).
8. Indica si el átomo es aromático (pertenece a un anillo de benceno o similar).
9. Indica si el átomo pertenece a un anillo en la estructura molecular.

¹ https://pytorch-geometric.readthedocs.io/en/stable/generated/torch_geometric.datasets.MoleculeNet.html#torch_geometric.datasets.MoleculeNet

Objetivo

Realizar los siguientes pasos

1. Cargar el dataset de MoleculeNet HIV y dividirlo en train y test.
2. Diseñar una GNN con tres capas de paso de mensajes de tipo GAT²; jumping knowledge de tipo “concat”; capa de global pooling y un MLP como clasificador.
3. Comprobar cual de los operadores simples de pooling (add, mean, max) funciona mejor para nuestra tarea.

² https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.nn.conv.GATConv.html#torch_geometric.nn.conv.GATConv