به نام خداوند رنگین کمان

پروژه درس گراف کاوی

استاد: دکتر زهرا طاهری

زهرا تبیانیان

فهرست مطالب

مقدمه

در این گزارش به بررسی مدل‌های متنوع شبکه عصبی گرافی جهت انجام کلاس‌بندی و رگرسیون گرافی روی مجموعه داده‌های مولکولی پرداختیم. شبکه‌های عصبی گرافی (GNN) که به عنوان یک کاندید قدرتمند برای مدل‌سازی داده‌های با ساختار گراف توسعه یافته‌اند، از داغ‌ترین موضوعات در سال‌های اخیر هستند. یک مولکول را می توان به طور طبیعی به عنوان یک گراف نشان داد که در آن اتم ها (راس‌ها) با پیوندهای شیمیایی (یال‌ها) به هم متصل می شوند.

کلاس بندی گراف‌ها

کلاس بندی گراف مسئله تعیین دسته یا برچسب گراف است. اگر مجموعه داده ای متشکل از تعداد زیادی گراف ورودی داشته باشیم، مسأله این است که هر یک از گراف‌ها را به دسته یا برچسب هدف درست خود کلاس بندی کنیم.

می‌خواهیم روی مجموعه داده‌ی BBBP کلاس بندی گرافی انجام دهیم. این مجموعه داده از یک مطالعه اخیر در مورد مدل‌سازی و پیش‌بینی نفوذپذیری مانع می‌آیند. مسأله به این صورت است: پیش بینی اینکه یک ترکیب شیمیایی نسبت به سد خونی-مغزی نفوذ پذیر هست یا خیر. بنا بر این مسئله از نوع کلاس بندی دو کلاسه(binary classification) است.

قابل توجه است که مجموعه داده های مورد استفاده(هم BBBP و هم مجموعه داده ای که در رگرسیون به کار بردیم)، دارای ویژگی های اولیه ی راسی(اتم)، یالی(پیوند بین اتم ها) و گرافی هستند.تمامی این ویژگی ها توسط پکیج RDKit محاسبه شده اند. سایز ویژگی های اولیه گرافی ۲۰۰ است. این ویژگی‌ها به خروجی GNN متصل می‌شوند تا از طریق یک شبکه عصبی feed forward با لایه های پنهان که با تابع فعال‌سازی ReLU برای تولید پیش‌بینی‌های ویژگی‌ها استفاده می‌شوند، بگذرند. ویژگی های راسی و یالی در جداول زیر لیست شده اند.

جدول ۱: ویژگی های راسی(اتم ها)

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

جدول ۲: ویژگی های یالی(پیوند ها)

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

برای حل این مسئله کلاس بندی و همچنین رگرسیون، پنج شبکه عصبی گرافی متنوع پیاده سازی کردیم:

- model_1: یک GCN ساده با دو لایه GraphConv
- model_2: یک GCN با چهار لایه GraphConv به همراه لایه های batch normalization و همچنین اعمال dropout
- model_3: یک GraphSAGE با دولایه SAGEConvپیاده سازی شده
- model_4: یک GrapgSAGE با چهار لایه SAGEConv به همراه لایه های batch normalization و اعمال dropout
- model_5: یک شبکه عصبی Custom با چهار لایه و لایه های batch normalization و اعمال dropout.

نتایج هر مدل بر اساس متریک ROC-AUC در جدول زیر آمده است. شایان ذکر است که تعداد واحد های پنهان را ۱۰۰، تعداد task ها را با توجه به مجموعه داده مذکور ۱، تعداد ایپاک ها را ۱۰۰ و patience را ۱۰ در نظر گرفتیم.

جدول ۳: نتایج کلاس بندی

	Message func.	Aggregation func.	number of conv. layers	BN	Dropout	Test Score
model_1	default GCN	default GCN	2	False	False	0.628
model_2	default GCN	default GCN	4	True	True	0.737
model_3	copy_u	mean	2	False	False	0.59
model_4	copy_u	mean	4	True	True	0.811
model_5	u_mul_v	sum	4	True	True	0.628

طبق نتایج به دست آمده مدل ۴ از همه بهتر عمل کرده است.

رگرسیون گراف‌ها

رگرسیون گراف‌ها شبیه به کلاس بندی آن‌هاست و تفاوت در تابع loss و متریک عملکرد است. همچنین کلاس دیتاست نیز در کلاس بندی و رگرسیون با هم تفاوت دارند چرا که در رگرسیون حتما باید scaling و نرمالایز کردن داده ها انجام شود.

برای رگرسیون از مجموعه داده FreeSolv استفاده کردیم. این مجموعه داده از پایگاه داده Free Solvation انتخاب شده است، که حاوی انرژی آزاد هیدراتاسیون مولکول های کوچک در آب از آزمایش ها و محاسبات انرژی آزاد شیمیایی است. در واقع هدف پیش بینی میزان انرژی آزاد هیدراتاسیون یک مولکول کوچک در آب است.

ویژگی‌های راسی و یالی اولیه به همان صورتی است که در بخش کلاس بندی به آن اشاره شد.

برای حل این مسئله رگرسیون از همان شبکه هایی که در بخش کلاس بندی داشتیم استفاده کردیم. با این تفاوت جزیی که شبکه custom آخر کمی متفاوت است. در جدول زیر جزییات بیشتر و نتیجه هر مدل بر حسب متریک RMSE آمده است.

جدول ۴: نتایج رگرسیون

	Message func.	Aggregation func.	number of conv. layers	BN	Dropout	Test Score
model_1	default GCN	default GCN	2	False	False	3.632
model_2	default GCN	default GCN	4	True	True	2.693
model_3	copy_u	mean	2	False	False	2.462
model_4	copy_u	mean	4	True	True	1.864
model_5	u_add_v	mean	4	True	True	2.727

در اینجا نیز طبق نتایج به دست آمده مدل ۴ عملکرد بهتری نسبت به دیگر مدل‌ها داشته است.