دانشکده علوم

استاد درس:دکتر طاهری
بهار ۱۴۰۲
گزارش پروژه پایانی
درس گراف کاوی

زهرا حق شناس

1.معرفی

هدف این گزارش کار ارائه جزئیات پروژه با هدف پیش‌بینی خواص مولکولی با استفاده از مدل شبکه عصبی نمودار (GNN) است. این پروژه از مجموعه داده BBBP (Blood-Brain Barrier Penetration)استفاده می کند که نفوذپذیری ترکیبات به سد خونی مغزی را ثبت می کند. این گزارش شرح مجموعه داده ها، مروری بر کد استفاده شده و خروجی به دست آمده از اجرای کد روی مجموعه داده را پوشش می دهد.

2. توضیحات مجموعه داده

مجموعه داده های BBBP از داده های مولکولی مربوط به نفوذپذیری سد خونی-مغزی ترکیبات تشکیل شده است. مجموعه داده ها شامل 2039 ترکیب است که هر کدام با مجموعه ای از توصیفگرهای مولکولی نشان داده می شوند. هدف این است که پیش‌بینی شود آیا یک ترکیب معین نسبت به سد خونی مغزی نفوذپذیر است یا خیر، که نشان‌دهنده توانایی آن برای عبور از این سد فیزیولوژیکی است. مجموعه داده به عنوان یک کار طبقه بندی طبقه بندی می شود، که در آن متغیر هدف باینری است (قابل نفوذ یا غیرقابل نفوذ) .مجموعه داده ها از مطالعه مارتینز و همکاران استخراج شده است. (2012) در مورد مدل سازی و پیش بینی نفوذپذیری سد خونی مغزی. نویسندگان از یک رویکرد بیزی برای مدل‌سازی نفوذ سد خونی مغزی سیلیکونی استفاده کردند

3.مروری بر کد

کد ارائه شده برای این پروژه به زبان پایتون نوشته شده است و از کتابخانه های DGL (Deep Graph Libraryو PyTorch استفاده می کند. از چندین بخش تشکیل شده است:

نصب و راه اندازی: کد با نصب کتابخانه های مورد نیاز و تنظیم DGL به PyTorch آغاز می شود.

مجموعه داده های سفارشی :PyTorch کد یک کلاس مجموعه داده سفارشی، DGLDatasetClassرا برای بارگیری و سازماندهی مجموعه داده های BBBP برای وظایف طبقه بندی تعریف می کند.

Train، Validationو :Test Set کد مجموعه‌های قطار، اعتبارسنجی و آزمایش را با ایجاد نمونه‌هایی از DGLDatasetClass برای تقسیم‌بندی داده‌های مربوطه، مقداردهی اولیه می‌کند.که 1631 داده مربوط به train_set، 203 داده مربوط به val_setو 205 داده مربوط به test_setاست.

Data Loader:کد یک تابع بارگذار داده، لودر را برای ایجاد بارگذارهای داده برای قطار، اعتبارسنجی و مجموعه های آزمایشی تعریف می کند و پردازش دسته ای کارآمد را در طول آموزش و ارزیابی امکان پذیر می کند.

تعریف مدل :GNN کد یک مدل GNN به نام GNN را تعریف می کند که لایه های کانولوشن گراف و گذر رو به جلو را برای نمودارهای مولکولی پیاده سازی می کند.

Loss Function:کد یک تابع از دست دادن سفارشی، loss_funcرا ارائه می‌کند، که با در نظر گرفتن هرگونه پوشش داده، تلفات بین خروجی مدل و برچسب‌های حقیقت زمین را محاسبه می‌کند.

آموزش و ارزیابی: کد شامل توابعی برای آموزش و ارزیابی مدل است. تابع آموزش مدل را با استفاده از مجموعه قطار آموزش می دهد و عملکرد آن را در مجموعه اعتبارسنجی تأیید می کند. تابع ارزیابی عملکرد مدل را در مجموعه تست محاسبه می کند.

4.خروجی کد

پس از اجرای کد روی مجموعه دادهBBBP ، نتایج زیر به دست آمد:

میانگین امتیاز معتبر: 0.819

امتیاز آزمون: 0.623

زمان اجرا: 55.949 ثانیه

«متوسط امتیاز معتبر» میانگین عملکرد مدل را در مجموعه اعتبارسنجی نشان می‌دهد و نشان می‌دهد که چقدر به داده‌های دیده نشده تعمیم می‌یابد. "امتیاز تست" عملکرد مدل را در مجموعه آزمایش نشان می دهد و تخمینی از توانایی آن در پیش بینی نفوذپذیری سد خونی-مغزی ترکیبات ارائه می دهد. زمان اجرا نشان دهنده مدت زمان اجرای کد و به دست آوردن نتایج است.

5. نتیجه گیری

در نتیجه، این گزارش کار پروژه‌ای را با تمرکز بر پیش‌بینی نفوذپذیری سد خونی-مغزی ترکیبات با استفاده از مدل شبکه عصبی نمودار توصیف می‌کند. این کد از کتابخانه های DGL و PyTorch برای پردازش مجموعه داده هایBBBP ، آموزش مدل و ارزیابی عملکرد آن استفاده می کند. بر اساس معیارهای خروجی ارائه شده، مدل به میانگین امتیاز اعتبار 0/819 و نمره آزمون 0/623 دست یافت. زمان اجرای کد 55.949 ثانیه بود.

این نتایج نشان می‌دهد که مدل سطح معقولی از عملکرد را در پیش‌بینی نفوذپذیری سد خونی مغزی نشان می‌دهد. با این حال، تجزیه و تحلیل بیشتر و مقایسه با سایر مدل ها یا خطوط پایه برای تعیین اثربخشی کلی رویکرد ضروری است.

گزارش کار: پیش بینی فعالیت مهاری ترکیباتHIV

1.معرفی

هدف اصلی این گزارش کار ارائه یک تجزیه و تحلیل جامع از یک پروژه متمرکز بر پیش‌بینی فعالیت مهاری ترکیبات HIV است. این پروژه از مجموعه داده های HIV از برنامه درمان دارویی AIDS Antivirus Screen (DTP)استفاده می کند. مجموعه داده ها حاوی اطلاعاتی در مورد بیش از 41000 ترکیب است و هدف آن پیش بینی توانایی آنها در مهار تکثیر HIV است. این پروژه از یک رویکرد طبقه بندی استفاده می کند، که در آن ترکیبات به کلاس های فعال بازدارنده تکثیر (HIV) یا غیرفعال طبقه بندی می شوند. معیار ارزیابی برای این کار، امتیاز ROC-AUC است.

2.توضیحات مجموعه داده

مجموعه داده های HIV بخشی از برنامه درمان دارویی AIDS Antivirus Screen (DTP) است. این شامل داده های مولکولی مربوط به بیش از 41000 ترکیب و فعالیت مربوط به مهار HIV مربوط به آنها است. برای هر ترکیب، مجموعه‌ای از توصیفگرها و ویژگی‌های مولکولی ارائه می‌شود که ویژگی‌های مختلف مرتبط با مهار HIV را نشان می‌دهد. متغیر هدف باینری است، با ترکیبات به عنوان فعال (1) یا غیر فعال (0) بر اساس توانایی آنها در مهار تکثیر HIV برچسب گذاری شده است.

-نام مجموعه داده: مجموعه دادهHIV

-دسته: بیوفیزیک

-تعداد وظایف: 1

-تعداد ترکیبات: 41127

-نوع وظیفه: طبقه بندی
-معیار ارزیابیROC-AUC :

3.نمای کلی کد

کد ارائه شده برای این پروژه در پایتون پیاده سازی شده و از کتابخانه هاییماندPyTorch ، DGL (Deep Graph Library)و سایر بسته های مربوطه استفاده می کند. کد مراحل کلیدی زیر را انجام می دهد:

-بارگذاری و پیش پردازش داده ها: کد مجموعه داده های HIV را بارگیری می کند و آن را پیش پردازش می کند تا توصیفگرها و ویژگی های مولکولی به همراه برچسب های مربوط به آنها آماده شود. داده ها به مجموعه های قطار، اعتبار سنجی و آزمایش تقسیم می شوند. که 32901 داده مربوط به train_set، 4112 داده مربوط به val_setو 4114 داده مربوط به test_setاست.

-نمایش گراف مولکولی: برای رسیدگی به ساختارهای مولکولی، کد برای هر ترکیب، نمایش گراف را می سازد. نمودار با گره‌هایی که اتم‌ها را نشان می‌دهند و یال‌هایی که پیوندهای شیمیایی را نشان می‌دهند ایجاد می‌شود. این ساختار نمودار به مدل اجازه می دهد تا روابط فضایی و برهمکنش های بین اتم ها را ثبت کند.

-مدل شبکه عصبی گراف: کد یک مدل شبکه عصبی گراف (GNN) را تعریف می کند که قادر به پردازش نمودارهای مولکولی است GNN .از لایه‌های کانولوشن گراف تشکیل شده است که اطلاعات اتم‌های همسایه را جمع‌آوری می‌کند و نمایش گره‌ها را به‌طور مکرر به‌روزرسانی می‌کند. مدل یاد می گیرد که ویژگی های مولکولی معنی دار را برای کار پیش بینی استخراج کند.

-آموزش: کد مدل GNN را با استفاده از مجموعه قطار آماده شده آموزش می دهد. پارامترهای

مدل را با استفاده از روش‌های بهینه‌سازی مبتنی بر گرادیان و انتشار پس‌انداز بهینه می‌کند. در طول آموزش، مدل به حداقل رساندن از دست دادن طبقه بندی بین پیش بینی های خود و برچسب های واقعی است.

-اعتبارسنجی: پس از هر دوره آموزشی، کد عملکرد مدل را در مجموعه اعتبارسنجی ارزیابی می کند تا بر توانایی تعمیم آن نظارت کند. امتیاز ROC-AUC به عنوان معیار ارزیابی محاسبه می شود.

-تست: پس از تکمیل آموزش مدل، کد عملکرد آن را در مجموعه تست ارزیابی می کند تا تخمین نهایی قابلیت های پیش بینی آن را ارائه دهد. امتیاز ROC-AUC در مجموعه تست ثبت می شود.

4.خروجی کد

با اجرای کد روی مجموعه داده هایHIV ، خروجی زیر به دست می آید:

میانگین امتیاز ROC-AUC در مجموعه اعتبارسنجی: 0.762
امتیاز ROC-AUC در مجموعه تست: 0.702
زمان اجرا: 783.420 ثانیه
"متوسط امتیاز ROC-AUC در مجموعه اعتبار سنجی" نشان دهنده عملکرد مدل در داده های اعتبارسنجی در طول آموزش است. این نشان می‌دهد که چقدر این مدل به داده‌های دیده نشده تعمیم می‌یابد و به عنوان معیاری برای توانایی آن در پیش‌بینی فعالیت مهار HIV عمل می‌کند.

"امتیاز ROC-AUC در مجموعه تست" ارزیابی نهایی عملکرد مدل بر روی داده های آزمایشی دیده نشده را نشان می دهد. این تخمینی از قابلیت های پیش بینی مدل در شناسایی ترکیبات با فعالیت مهار HIV ارائه می دهد.

"زمان اجرا" مدت زمان مورد نیاز برای آموزش مدل و ارزیابی عملکرد آن در مجموعه

تست را نشان می دهد. این نشان دهنده کارایی اجرای کد است.

5.نتیجه گیری

در نتیجه، این گزارش کاری پروژه‌ای را با تمرکز بر پیش‌بینی فعالیت مهار HIV ترکیبات با استفاده از مدل شبکه عصبی گراف شرح داد. این پروژه از مجموعه داده های HIV از برنامه درمان دارویی AIDS Screen ضد ویروسی [8] استفاده کرد و یک رویکرد طبقه بندی را به کار گرفت. کد بارگذاری داده ها، پیش پردازش، ساخت گراف، آموزش مدل GNNو مراحل ارزیابی را پیاده سازی کرد. خروجی کد شامل امتیاز ROC-AUC در مجموعه اعتبار سنجی و مجموعه تست به همراه زمان اجرا بود.

بر اساس خروجی ارائه شده، مدل به امتیاز ROC-AUC 0.702 در مجموعه آزمایشی دست یافت که توانایی آن را در پیش بینی فعالیت مهار HIV ترکیبات نشان می دهد. زمان اجرای کد 783.420 ثانیه بود که نشان دهنده کارایی راه حل پیاده سازی شده است