یادگیری ماشین برای بیوانفورماتیک

نیم‌سال دوم ۹۹-۹۸

دانشکده‌ی مهندسی کامپیوتر

گردآورندگان: رسا قوامی، روح‌الله پورسلیمانی

پروژه پایانی ــ فاز دوم

مقدمه

عملکرد تعدادی از داروها به گونه‌ای است که به پروتئین خاصی می‌چسبند و عملکرد آن را تغییر می‌دهند. از آنجایی که طراحی یک داروی جدید کاری بسیار زمان‌بر است و آزمایش‌های متعددی می‌طلبد، می‌توان با استفاده از مدل‌های محاسباتی پیش از شروع به انجام آزمایش‌ها از عملکرد آن اطمینان بیشتری حاصل کرد. از طرفی دیگر لازم است مطمئن شویم که داروی جدید روی سایر پروتئین‌ها تاثیری ندارد. بنابراین پیش‌بینی اتصال دارو به پروتئین از اهمیت زیادی برخوردار است. بسیاری از مدل‌ها فقط به پیش‌بینی اتصال/عدم اتصال می‌پردازند این در حالی است که میزان اتصال دارو به پروتئین یک کمیت پیوسته است. در واقع هر چقدر تمایل یک دارو برای اتصال به یک پروتئین بیشتر باشد، دز ثابتی از دارو به تعداد بیش‌تری پروتئین می‌چسبد و تاثیری که دارو می‌گذارد هم به این میزان وابسته است.

مسئله

در سال ۲۰۱۸ مدلی به اسم DeepDTA ارائه شد که میزان اتصال دارو به پروتئین را با استفاده از یک مدل عمیق پیش‌بینی می‌کرد. این مدل روی دو مجموعه داده Davis و KIBA ارزیابی شده است.

۱. مدل ارائه شده در مقاله را بررسی کنید. به نظرتان نقاط قوت و ضعف مدل ارائه شده چیست؟

۲. مدلی ارائه دهید که نقاط ضعف مذکور در بند پیش را برطرف کند. دلایلتان برای ارائه این مدل را توضیح دهید.

۳. با توجه به مدل طراحی شده توضیح دهید که داده‌ها به چه پیش‌پردازشی نیاز دارند.

۴. مقاله مرجع هر کدام از مجموعه داده‌ها را به دو بخش آموزش و آزمون تقسیم کرده است. همچنین برای تنظیم هایپرپارامترهای مدل با استفاده از ۵ fold-cross-validation مجموعه داده‌ی آموزش را به ۵ قسمت تقسیم کرده است. شما از ۴ بخش نخست آموزش به عنوان داده آموزش و از بخش آخر به عنوان داده ارزیابی استفاده کنید. مدل خود را با استفاده از pytorch پیاده‌سازی کنید. برای تعیین مدل، هایپرپارامترهای آن، نحوه و میزان آموزش از داده‌های ارزیابی استفاده کنید. مدل نهایی را روی داده‌های آزمون اجرا کرده و معیارهای میانگین مربعات خطا (MSE) و شاخص تطابق (CI) را برای خروجی‌های محاسبه کرده و با مقادیر ارائه شده در مقاله مقایسه کنید.

۵. با استفاده از حدگذاری، مسئله رگرسیون را به یک مسئله دسته‌بندی تبدیل می‌کنیم. برای حد مرزی با توجه به مقادیر پیشنهادی مقاله، برای مجموعه داده Davis از مقدار ۷ ($pK_d < ۷$ معادل اتصال است) و برای مجموعه داده KIBA از مقدار ۱۲,۱ استفاده کنید. معیارهای دقت [۱]، حساست [۲]، تشخیص‌پذیری [۳] و امتیاز F۱ را برای داده‌های آزمون گزارش کنید. با توجه به مقادیر گزارش شده، عملکرد مدل خود را توصیف کنید (برای چه چیزهایی ضعیف‌تر عمل می‌کند/ در همه موارد قوی عمل می‌کند).

۶. هیستوگرام توزیع مقادیر خروجی (میزان اتصال دارو به پروتئین) را برای تمامی جفت دارو-پروتئین‌هایی که اطلاعات میزان اتصالشان را داریم برای هر کدام از مجموعه داده‌ها به طور جداگانه ترسیم کنید. با توجه

[۱] Accuracy
[۲] Sensitivity
[۳] Specificty

به هیستوگرام به نظرتان ممکن است که مدل روی هر کدام از مجموعه داده‌ها به سمت مقادیر خاصی بایاس شود؟ برای حل این مشکل چه راه حلی را پیشنهاد می‌کنید؟ با استفاده از راه حل پیشنهادی مدل را یک بار دیگر آموزش دهید و معیارهای ارزیابی دو بند قبل را برای مجموعه داده آزمون گزارش دهید. نتایج چه فرقی با نتایج مرحله قبل داشته؟ به نظرتان چرا چنین اتفاقی افتاده؟

۷. بین دو مدلی که برای مجموعه داده‌ی Davis آموزش داده‌اید مدل بهتر را بر اساس معیارهای ارزیابی انتخاب کرده و برای سوالات بعدی از آن استفاده کنید.

۸. یکی از ایرادات بزرگی که به شبکه‌های عمیق وارد است عدم توانایی در تشریح آن‌ها است. یعنی به طور مستقیم نمی‌توانیم بگوییم چرا یک شبکه عمیق برای یک ورودی خاص، خروجی داده شده را تولید کرده است. درک نحوه عملکرد یک شبکه به اعتماد به کارکرد شبکه و اطمینان از آموزش درست آن منجر می‌شود. همچنین می‌توان با تفسیر نتایج حاصل از شبکه، به دانش بیشتری در مورد مسئله رسید. برای مثال در این مورد می‌توان گفت در صورتی که با استفاده از شبکه بتوانیم به این نکته پی ببریم که برای یک پروتئین خاص، وجود چه رشته/رشته‌های شیمیایی موجب اتصال دارو به آن می‌شود و وجود چه رشته/رشته‌هایی مانع از اتصال دارو به آن می‌شود می‌توانیم داروهای مناسب‌تری برای پروتئین طراحی کنیم. روش‌های زیادی مطرح شده‌اند که دلیل عملکرد شبکه برای یک ورودی خاص را توضیح دهند. تعدادی از این روش‌ها، قسمت‌هایی از ورودی را مشخص می‌کنند که روی نتیجه نهایی تاثیر زیادی گذاشته‌اند. برای مثال از جمله ساده‌ترین این روش‌ها می‌توان به Saliency Maps، Guided Back Propagation، Guided Grad-Cam و LRP اشاره کرد. با استفاده از منابع معرفی شده در مورد نحوه کار هر یک از این روش‌ها و پایه تئوری پشت آن‌ها (از نظر شهودی) توضیح دهید.

۹. چلنج اصلی در ارائه یک روش تفسیری برای شبکه‌های عمیق، تفسیرپذیر بودن خود تفسیر!، میزان Robust بودن آن و کاراییش در کشف درست دلایل است. برای مثال در صورتی که یک روش تفسیری به بازه‌های کم و منسجم‌تری از ورودی اشاره کند از نظر انسان قابل فهم‌تر است همچنین احتمال اینکه به نویزهای ورودی هم توجه کرده باشد در این حالت کاهش می‌یابد. ولی ممکن است یک مدل تمام بازه‌های مهم ورودی را کشف نکرده باشد و بدین ترتیب دقت کار آن پایین می‌آید. ابتدا روش‌های نام‌برده را پیاده‌سازی کرده و سپس با استفاده از آن‌ها تمامی بازه‌های ورودی که روی نتیجه خروجی تاثیر مثبت دارند را برای تمامی جفت پروتئین_داروهایی که به هم متصل می‌شوند و شبکه هم این را درست پیش‌بینی کرده پیدا کنید. برای هر روشی که نیازمند انتخاب لایه/پارامتر است، در مورد دلایل انتخابتان توضیح دهید. برای ارزیابی قابلیت درک توسط انسان در روش‌های نام‌برده، نمودار توزیع تعداد حروف انتخابی در ورودی (جمع تعداد حروف انتخابی برای دارو و پروتئین) را رسم کنید. روش‌ها را از نظر تعداد حروف انتخابی مقایسه کنید. برای مقایسه دقت روش‌های تفسیری می‌توان بازه‌های انتخابی آن‌ها را حذف کرد (جای‌گذاری با یک حرف بی‌اثر یا یک حرف رندم). اگر این بازه‌ها درست شناسایی شده باشند باید شبکه به ازای ورودی دستکاری شده جواب متفاوتی بدهد. این کار را برای تمام جفت‌هایی که به هم متصل می‌شوند و شبکه هم این را درست پیش‌بینی کرده انجام دهید. یک بار تمامی بازه‌های انتخابی در دارو را خراب کنید و بار دیگر پروتئین. دقت شبکه را روی مجموعه داده جدید بسنجید. هر روشی که مجموعه داده‌ی تولیدی از روی آن دقت پایین‌تری داشته باشد در تشخیص نواحی حساس بهتر عمل کرده است. روش‌های تفسیری را از این نظر مقایسه کنید. با استفاده از دو بررسی که انجام دادید، بهترین روش را انتخاذ کنید و برای سوالات بعدی از آن استفاده کنید.

۱۰. نواحی انتخابی برای هر پروتئین و هر دارو را در تمامی جفت‌هایی که آن پروتئین/دارو مشارکت دارند در نظر بگیرید. روشی ارائه دهید که با استفاده از آن معناداری تکرر انتخاب نواحی به ازای هر پروتئین/دارو را بین جفت‌هایی که اتصال دارند و شبکه درست پیش‌بینی کرده در مقابل جفت‌هایی که اتصال دارند ولی شبکه اشتباه پیش‌بینی کرده محاسبه کنید. دقت کنید که این کار را یک بار برای داروها و یک بار برای پروتئین‌ها به طور جداگانه انجام دهید. هدف بررسی این است که آیا نواحی یکسانی در پروتئین‌ها موجب اتصال داروها به آن‌ها می‌شود و آیا نواحی یکسانی در داروها موجب قابلیت اتصال آن‌ها به پروتئین‌ها می‌شود. نتایج حاصل را تفسیر کنید.

۱۱. با اضافه کردن ترم‌هایی به تابع هزینه یا تغییر معماری، شبکه خود را تفسیرپذیرتر کنید. یعنی آن را به گونه‌ای تغییر دهید که درک نحوه عملکردش برای انسان راحت‌تر شود یا با استفاده از روش‌های تفسیر کردن نتیجه،

نتیجه منسجم‌تر و تفسیرپذیرتری حاصل شود. شبکه را آموزش دهید. آن را از نظر معیارهای ارزیابی معرفی شده برای دسته‌بندی با شبکه قبل مقایسه کنید. همچنین بند قبل را روی شبکه جدید هم انجام دهید و با شبکه قبلی مقایسه کنید. نتایج حاصل را تفسیر کنید.

مجموعه داده‌ها

داده‌ها از https://github.com/hkmztrk/DeepDTA گیت‌هاب مقاله مرجع قابل دسترسی هستند. همانطور که گفته شد مقاله از دو مجموعه داده‌ی Davis و KIBA استفاده کرده است. مجموعه داده Davis شامل اطلاعات میزان اتصال ۳۰۰۵۶ جفت پروتئین_دارو است بین ۴۴۲ پروتئین یکتا و ۶۸ داروی یکتا و مجموعه داده KIBA شامل اطلاعات میزان اتصال ۱۱۸۲۵۴ جفت پروتئین_دارو بین ۲۲۹ پروتئین یکتا و ۲۱۱۱ داروی یکتا است. برای هر کدام از مجموعه‌ها، فایل‌های زیر وجود دارند (تنها فایل‌های مورد نیاز توضیح داده شده‌اند):

- رشته مولکول دارو به دو فرمت مختلف در دو فایل ligands-iso و ligands-can.txt قرار دارند. در این فایل رشته مولکول هر دارو بعد از شناسه آن آورده شده است. شناسه‌ها و رشته‌ها با استفاده از : و داروهای مختلف با استفاده از , از هم جدا شده‌اند. برای آشنایی بیشتر با فرمت SMILES می‌توانید از این مرجع استفاده کنید.
- رشته پروتئین‌ها در همان فرمت توضیح داده شده در بالا در فایل proteins.txt قرار دارد.
- Y شامل میزان اتصال جفت‌های پروتئین و دارو است. فایل آن یک فایل pickle است و با استفاده از کتابخانه pandas قابل باز شدن است. حاوی یک ماتریس است (یک سطر به ازای هر دارو و یک ستون به ازای هر پروتئین). اندیس هر دارو/پروتئین ترتیب آمدن شناسه آن در فایل حاوی رشته‌های دارو/پروتئین است.
- پوشه folds حاوی جفت‌هایی است که برای مجموعه آموزش/آزمون انتخاب شده‌اند. با استفاده از کد ارائه شده در این صفحه می‌توانید شناسه جفت‌ها را به پروتئین و داروی مربوطه مپ کنید. داده‌ی آزمون حاوی یک لیست از شناسه‌های جفت‌هاست و فایل داده آموزش لیستی از ۵ لیست از شناسه‌هاست که هر کدام مربوط به یک fold داده است که توسط مقاله جدا شده.

برای استفاده از داده‌ها هم می‌توانید ازکدهای گیت‌هاب معرفی شده استفاده کنید هم می‌توانید خودتان با توجه به فرمت توضیح داده شده فایل‌ها را بخوانید.

نکات مهم

۱. مستندی تهیه کنید و برای تمامی موارد خواسته شده تصمیماتی که گرفته‌اید و نتایجی که به دست آوردید در آن توضیحات کامل و جامعی ارائه دهید. لازم به ذکر است توضیحات برای تمامی تصمیمات شما و تفسیر تمامی نتایج الزامی هستند. ضمنا کامنت‌های میان کد مستند محسوب نمی‌شوند.

۲. بهترین مدل از بین مدل‌های دسته‌بندی که زده‌اید را از نظر معیارهای ارزیابی انتخاب کنید. اسکریپتی به نام Main.py تهیه کنید که با دریافت آدرس فایل ورودی و آدرس ذخیره نتایج، مدل انتخابی را روی فایل ورودی اجرا کرده و حاصل آن (اتصال/عدم اتصال) را در فایل خروجی ذخیره کند. در هر سطر فایل ورودی رشته یک پروتئین tab ترکیب شیمیایی دارو آورده شده است و لازم است در همان سطر فایل خروجی پیش‌بینی مدل خود را ذخیره کنید. نحوه اجرای اسکریپت:
python Main.py InpDir OutDir

۳. لطفا کدهای خود را به همراه فایل مستند جداگانه در پوشه‌ای زیپ کرده و در کوئرا آپلود کنید. همچنین آپلود باکسی برای مدل انتخابی شما در نظر گرفته شده. دقت کنید که لازم است مدل خود را در آن آپلود کنید.

۴. سنت حسنه مشورت برای قسمت طراحی مدل و پیاده‌سازی‌ها قابل قبول نیست. مدل‌ها و کدها شباهت سنجی می‌شوند و با خاطیان برخورد خواهد شد.

ارزیابی

- مستند ٪۲۵
- منطق پشت تصمیمات ٪۲۵
- پیاده‌سازی ٪۲۵
- نتایج ٪۲۵
- ارائه روش‌های بسیار زیبا، ایده‌های جذاب و نتایج بیش از حد انتظار در هر یک از موارد بالا تا ٪۲۰ نمره اضافی خواهد داشت (سقف جمع نمره امتیازی: ٪۵۰).