پروژه پایان‌ترم درس هوش محاسباتی

پیش‌بینی تعاملات دارو-هدف با استفاده از روش‌های ترکیبی یادگیری ماشین و یادگیری عمیق

بهار ۱۴۰۴

پیش‌بینی تعاملات دارو-هدف یکی از چالش‌های مهم در کشف دارو و پزشکی شخص‌محور است که با شناسایی ارتباطات مولکولی بین داروها و پروتئین‌های هدف، مسیرهای جدیدی برای درمان بیماری‌ها و بازپروژه‌گذاری داروهای موجود فراهم می‌کند. در این پروژه، از روش‌های ترکیبی یادگیری ماشین و یادگیری عمیق برای پیش‌بینی این تعاملات با استفاده از دیتاست شبکه‌ای Drug-Target Interaction استفاده می‌شود. با استخراج ویژگی‌های ساختاری داروها و توالی پروتئین‌ها، مدل‌های مختلفی آموزش داده می‌شوند تا دقت پیش‌بینی را افزایش دهند. هدف نهایی، ارائه یک چارچوب هوشمند برای شناسایی تعاملات ناشناخته دارو-هدف با کاربرد در طراحی داروهای جدید و کاهش هزینه‌های تحقیقاتی است. این پروژه می‌تواند به عنوان پایه‌ای برای تحقیقات آینده در حوزه هوش مصنوعی، پزشکی و بیوانفورماتیک مورد استفاده قرار گیرد.

مرحله ۱: پیش‌پردازش داده‌ها

در این بخش، نیاز است ابتدا دیتاست SNAP Drug-Target Interaction را بارگذاری نمایید. این دیتاست، شامل ۵۰۱۷ دارو، ۲۳۲۴ پروتئین و ۱۵۱۳۸ تعامل است. در صورت تمایل، می‌توانید با جستجو دیتاست‌های دیگری را نیز استفاده نمایید. در گزارش خود، حتما دیتاست‌هایی را که استفاده نموده اید را به‌طور کامل توضیح دهید. در ادامه مراحل زیر را برای پیش‌پردازش داده‌ها انجام دهید.

۱-۱. پاکسازی داده‌ها

در این مرحله، نیاز است داده‌های موجود در دیتاست خود را پاکسازی نمایید. بدین‌منظور، عملیات‌هایی همچون حذف داده‌های تکراری و یا مدیریت مقادیم گم‌شده و از این قبیل می‌بایست انجام شود.

۱-۲. استخراج ویژگی‌ها برای داروها

در این مرحله، نیاز است ویژگی‌هایی برای داروها در نظر گرفته و استخراج گردد. در این مرحله، می‌بایست تحقیق نمایید که از چه ویژگی‌هایی استفاده می‌گردد و در گزارش خود با ذکر منبع توضیح دهید. همچنین می‌توانید بر اساس مطالعات خود ویژگی‌هایی را نیز پیشنهاد دهید. بعنوان مثال، ویژگی‌های داروها می‌تواند ویژگی‌های شیمیایی آن یا اثرانگشت دارو باشد.

۱-۳. استخراج ویژگی‌ها برای پروتئین‌ها

در این مرحله نیز، می‌بایست همچون مرحله‌ی ماقبل، ویژگی‌هایی برای پروتئین‌ها پیدا کنید و این ویژگی‌ها میبایست پشتوانه علمی داشته باشد. به‌عنوان مثال، توالی آمینواسیدها می‌تواند ویژگی خوبی باشد.

۱-۴. تعدیل داده‌ها

در صورت نیاز و عدم تعادل در داده‌ها، می‌بایست از تکنیک‌های تعادل (Balancing) استفاده نمایید. روش خود را در گزارش کامل توضیح دهید.

۱-۵. کاهش ابعاد

در آخرین بخش مرحله‌ی پیش‌پردازش داده‌ها، در صورت نیاز، با استفاده از هر روش دلخواه، می‌بایست ابعاد ویژگی‌ها را کاهش دهید. علت روش انتخابی خود را در گزارش به‌صورت تمام و کمال بنویسید.

مرحله ۲: مدلسازی

پس از پیش‌پردازش داده‌ها، نیاز است که مدل‌سازی انجام شود. در ابتدا، با مطالعه‌ی چند مقاله‌ی ۵ سال اخیر در این حیطه، چند مدل را انتخاب نمایید؛ این مدل‌ها می‌توانند مدل‌های یادگیری ماشین همانند Random Forest، SVM، MLP و یا مدل‌های یادگیری عمیق همانند GNNها و GCNها باشند. از هر نوع مدل، حداقل ۲ مدل انتخاب شود. در گزارش خود، علت انتخاب مدل را نیز تمام و کمال شرح دهید.

پس از انتخاب مدل، نیاز است داده‌ها را به داده‌های آموزش و آزمون تقسیم نموده و مدل‌ها را آموزش داده و تحت آزمایش قرار دهید. آموزش مدل‌ها را می‌توانید با استفاده از هر روش دلخواه همچون تقسیم به داده‌های آموزش و آزمون، روش K Fold یا هر روش دیگری انجام دهید. توجه فرمایید که این مورد را در گزارش خود به‌صورت تمام و کمال توضیح دهید.

سوال ۳: ارزیابی مدل

پس از آموزش مدل، نیاز است با استفاده از داده‌های آزمون یا با روش Cross-Validation، مدل خود را ارزیابی نمایید. میبایست مدل‌های خود را با تمامی معیارهای ارزیابی بیازمایید و بهترینشان را انتخاب نمایید. سپس می‌بایست دقت مدل خود را با سایر مقالات که مطالعه نموده اید مقایسه نمایید. برای گرفتن نمره‌ی این بخش، می‌بایست دقت مدل شما حداکثر ۵ درصد از کمترین دقت مدل موجود در مقالاتی که خوانده اید کمتر باشد. به ازای هر درصد دقت کمتر، ۲ درصد نمره این سوال کسر خواهد شد. همچنین دقت کنید که دقت مدل خود را می‌بایست با دقت مدل‌های مقالات ۲۰۲۰ به بعد مقایسه نمایید.

سوال ۴: خروجی‌های پروژه

- مدل‌های آموزش‌دیده
- مقایسه عملکرد روش‌های محتلف و مقالات پیشین
- کشف تعاملات جدید (آنهایی که در دیتاست وجود ندارند)

- گزارش‌نویسی بر اساس متودولوژی، نتایج و کارهای انجام‌شده