پروژه درس یادگیری ماشین در علوم زیستی

هدف از این پروژه استفاده از آموخته های درس در یک مسئله ی واقعی بیوانفورماتیکی است. مجموعه داده ی این پروژه یک مطالعه-- ی ترنسکریپتوم (RNA-Seq) جهت بررسی ژنتیک بیماری NAFLD (Non-Alchlolic Fattly Liver Disease) می باشد. این بیماری که کبد چرب غیر الکلی نامیده می شود، بیماری بسیار شایعی در جوامع صنعتی است. در مراحل اولیه این بیماری در صورت تشخیص، بیماری برگشت پذیر می باشد ولی در مراحل نهایی آن که به فیبروز کبدی و در نهایت سیروز کبدی منتهی می شود، بیماری برگشت پذیر نبوده و معمولا برای درمان نیاز به پیوند کبد می باشد. لذا تشخیص زود هنگام التهاب کبد و نیز فیبروز کبدی جهت درمان موثر و به موقع آن ضروری می باشد. هرچند روش استاندارد برای تشخیص این بیماری استفاده از سونوگرافی کبد می باشد ولی جهت درک بهتر مکانیزم بیماری و نیز طراحی پنل های ژنی به عنوان مارکر های تشخیصی، استفاده از اطلاعات ژنتیکی مانند بیان ژن ها اهمیت بالایی دارد.

برای این منظور در این مطالعه داده های ترنسکریپتوم مربوط به بافت کبد ۱۹۲ فرد، شامل افراد سالم، افراد مبتلا به فیبروز کبدی غیر پیشرفته و افراد مبتلا به فیبروز کبدی پیشرفته توسط روش های توالی یابی نسل جدید توالی یابی شده اند. در گام اول لازم است که داده های این پروژه که لیست کد اجرای آن ها در فایل run_accession_list.txt ذخیره شده است، از دیتابیس SRA دانلود گردند. سپس داده ها با استفاده از یک پایپلاین آنالیز داده های ترنسکریپتوم مانند Hisat2+Stringtie+limma-voom پردازش شده و داده های نرمال شده بیان ژن ها بدست آیند. توجه شود که در این مطالعه به دنبال ترنسکریپت های جدید نیستیم و هدف بدست آوردن بیان ژن ها و ترنسکریپت های شناخته شده می باشد. برای این قسمت حتما از ورژن Hg38 یا GRCH38 و فایل انوتیشن مربوط به آن استفاده نمایید. چنانچه انجام این مراحل برایتان دشوار است، می توانید از فایل های داده های نرمال شده بیان ژن ها که در اختیار شما قرار می گیرد استفاده کنید. بعد از مرحله پردازش داده های خام و حصول داده های نرمال شده بیان ژن، با استفاده از آموخته های درس به سوالات زیر پاسخ دهید و کد های مربوطه را ارائه نمایید. فایل متادیتا این پروژه با نام meta_data.csvدر دسترس قرار داده شده است.

۱. با ترکیب روش های انتخاب ویژگی و کاهش ابعاد و الگوریتم های طبقه بندی ماشین هایی را طراحی نمایید که بهترین Precisionو Recall را برای طبقه بندی نمونه هایNormal ، Advance Fibrosisو Non Advanced Fibrosisانجام دهند؟ حداقل تعداد بهینه ویژگی ها چند ویژگی می باشد؟ چه تفسیر های زیستی برای این ویژگی ها می توانید ارائه نمایید؟ نتایج بدست آمده چقدر قابل تکرار می باشند؟ کدام روش طبقه بندی و با چه پارامترهایی بهترین عملکرد را داشته است؟ کدام روش انتخاب ویژگی عملکرد بهتری داشته است؟

۲. در سوال (۱)، آیا داشتن اطلاعاتی مانند جنسیت، Body mass index (BMI)، ابتلا به دیابت و نیز سن چقدر تاثیر روی عملکرد ماشین ها دارد؟ آیا این ویژگی ها جزء ویژگی های انتخاب شده هستند؟

۳. فرض کنید که معماری ماشین طبقه بندی خود را به صورت زیر تغییر دهیم. ماشین اول، طبقه بندی بین افراد سالم و افراد مبتلا به NAFLD شامل advance-fibrosis و (non-advanced-fibrosis انجام دهد. و ماشین دوم در صورتی که فرد توسط ماشین اول مبتلا به NAFLD تشخیص داده شد، طبقه بندی بین Advanced Fibrosis و-Non advance Fibrosisرا انجام دهد. برای این حالت نیز موارد خواسته شده در سوال (۱) را انجام داده و نتایج را مقایسه نمایید.

٤. (اختیاری) با استفاده از اطلاعات بیان ژن ها و نیز سایر اطلاعات موجود در فایلMeta-data ، رگرسیونی را طراحی نمایید که بتواند سن بیمار را تخمین بزند. با استفاده از روش lasso سعی کنید حداقل تعداد ویژگی را در این مدل انتخاب نمایید. ویژگی های موثر را گزارش و تفسیر نمایید. آیا جدا کردن مدل برای نمونه های زن و مرد در بهبود دقت پیش بینی موثر است؟ برای این منظور رگرسیون های جداگانه برای نمونه های مرد و زن طراحی نموده و نتایج را با حالت اولیه مقایسه نمایید.

نکات مهم:

دانشجویان می توانند در گروه های یک یا دو نفره پروژه را انجام دهند. نماینده هر گروه باید در گروه اسکایپی درس نام افراد گروه را اعلام نماید.

هرگروه باید یک گزارش کتبی که حاوی پاسخ سوالات و روش مورد استفاده باشد، تهیه کند و تا روز تحویل شفاهی پروژه به آدرس ایمیل kkavousi@yahoo.com ارسال نماید. در subject ایمیل لازم است نام، نام خانوادگی، و شماره دانشجویی اعضای گروه ذکر شود. ارائه شفاهی به صورت مجازی و از طریق google meet انجام خواهد شد.

تاریخ تحویل پروژه بین ۲۰ تا ۲۳ مرداد خواهد بود. روز و ساعت دقیق متعاقبا اطلاع رسانی خواهد شد.

موفق باشید