مدرس: **دکتر فدایی و دکتر یعقوب‌زاده** طراحان: آتیه آرمین، بهار افشار، نسترن جمالی‌پور

مهلت تحویل: سه شنبه ۲۳ آذر ۱۴۰۰ ساعت ۲۳:۵۵

مقدمه

در این پروژه هدف آشنایی با روش‌های یادگیری ماشین به وسیله کتابخانه Sickit-Learn می‌باشد. این پروژه در چهار فاز تعریف شده است. در فاز صفر به بررسی و تجزیه تحلیل داده می‌پردازید. در فاز اول با روش‌های پیش‌پردازش آشنا می‌شوید، سپس به استخراج ویژگی از ستون‌های متنی می‌پردازید. در فاز دوم با استفاده از چند مدل معروف کتابخانه Sickit-Learn به پیشبینی هدف و بهینه‌سازی این مدل‌ها می‌پردازید. و در نهایت در فاز سوم با روش‌های یادگیری تجمعی آشنا می‌شوید و نتایج حاصل از این نوع مدل‌ها را با نتایج فاز قبل مقایسه می‌کنید.

معرفی مجموعه داده

مجموعه داده‌ای که در اختیار شما قرار گرفته، مجموعه‌ای حاوی تعدادی فیلم و مجموعه تلویزیونی موجود در سایت نتفلیکس و آمازون می‌باشد. این مجموعه داده حاوی ویژگی‌های عنوان، بازیگران، کشور، سال ساخت، ژانر، خلاصه فیلم و نوع فیلم (فیلم سینمایی یا مجموعه تلویزیونی) می‌باشد. در طول این پروژه به پیشبینی ستون نوع فیلم خواهیم پرداخت.

روش حل مسئله:

فاز صفر: بررسی داده

اولین گام در هر پروژه‌ی یادگیری ماشینی، مشاهده، شناخت و بررسی داده‌ها و ارتباط میان آنها است. به این منظور قدم‌های زیر را انجام دهید.

۱. با استفاده از متدهای describe و info کتابخانه pandas، ساختار کلی داده‌ها را بررسی کنید.

۲. درصد داده‌های از دست رفته هر ویژگی را پیدا کنید و نمایش دهید.

فاز یک: پیش‌پردازش

پیش‌پردازش داده‌ها مهم‌ترین گام در فرایند پروژه‌های یادگیری ماشین می‌باشد. در این فاز داده‌های خام ورودی باید به مجموعه‌ای از داده‌های قابل پردازش تبدیل شود. به این منظور قدم‌های زیر را انجام دهید.

۱. برای رفع مشکل داده‌های گمشده، روش‌های زیادی وجود دارد. به بررسی دو مورد از این روش‌ها بپردازید سپس یکی را برای به‌کارگیری روی داده‌ها انتخاب کنید و دلیل انتخاب خود را ذکر کنید.

۲. برای ویژگی‌های عددی Normalization یا Standardization به چه منظور استفاده می‌شود؟ شما کدام روش را انتخاب می‌کنید؟ چرا؟

۳. برای اینکه مدل ما بتواند با داده‌های دسته‌ای کار کند، روش‌های بسیاری وجود دارد. دو روش را توضیح دهید و بیان کنید از کدام روش استفاده کردید. چرا؟

۴. در مجموعه دادگان داده شده، ستون listed_in که نشان‌دهنده ژانر های یک فیلم است به صورت سری است. برای استفاده از این ستون از چه روش هایی می‌توان استفاده کرد؟ روش انتخابی خود را توضیح دهید.

استخراج ویژگی از متن

در این مجموعه ستون‌های متنی عنوان، خلاصه فیلم و بازیگران وجود دارد. برای پردازش این داده‌ها، شما میتوانید تعداد مشخصی از کلمات پرتکرار متن را به عنوان فیچر در نظر بگیرید و برای هر سطر داده، مشخص کنید این کلمات چند بار تکرار شده‌اند (count vectorizer) استفاده کنید. البته برای این کار معیار دیگری به نام tf-idf هم هست که علاوه بر تعداد کلمات، تعداد تکرار یک کلمه را در هم در نظر می‌گیرد، به عنوان مثال اگر کلمه‌ای زیاد در عبارت‌ها تکرار شود، ارزش آن را کم می‌کند چرا که وقتی در همه‌ی عبارت‌ها وجود داشته باشد، داشتن یا نداشتن آن کلمه بار اطلاعاتی کمتری خواهد داشت. در این لینک میتوانید بیشتر در مورد این دو مدل مطالعه کنید. در یادگیری حداقل یکی از مدل‌هایتان استفاده از count vectorizer را امتحان کنید. نتایج را ذکر کرده و توضیح دهید. بررسی کنید که استخراج ویژگی بیشتر چه تاثیری بر یادگیری مدل می‌گذارد.

بررسی روابط بین ویژگی‌ها

اکنون که فاز پیش‌پردازش داده‌ها به اتمام رسیده می‌خواهیم روابط بین ویژگی‌ها را به صورت دقیق‌تر بررسی کنیم تا ویژگی‌های بهتر را شناسایی کنیم. به این منظور information gain بین ویژگی‌ها را محاسبه کرده و نمودار gain را بر اساس ویژگی‌ها رسم کنید. نمودار خود را تحلیل کنید. این نمودار چه اطلاعاتی در مورد استفاده از ویژگی‌ها در ادامه کار به شما می‌دهد؟

فاز دوم: پیشبینی هدف و بهینه‌سازی مدل‌ها

در این فاز از پروژه به کمک کتابخانه‌ی Scikit-Learn مدل بر پایه‌ی Decision Tree پیاده‌سازی می‌کنید. سپس با استفاده از هایپر پارامترها به بهینه‌سازی این مدل خواهید پرداخت.

قبل از هر چیز لازم است تا داده‌ی خود را به دو بخش یادگیری و تست تقسیم کنید. سپس به تخمین ویژگی هدف بپردازید. برای ارزیابی تخمین‌هایتان از معیار accuracy استفاده کنید. نتیجه مطلوب برای این پروژه دستیابی به نتیجه ۹۰٪ برای معیار ذکر شده در حداقل یکی از مدل ها می‌باشد. برای بررسی بهتر نتایج هر مدل confusion-matrix مربوط به آن را رسم کنید و در مورد نتایج توضیح دهید.

برای هرکدام از مدل خود تحقیق کنید که هایپرپارامترهای max_depth و min_samples_split چه هستند. مقادیر آنها را برای مدل خود طوری تغییر دهید تا مدل بهینه شود (بهینه‌سازی مدل‌ها به این منظور است که خطا کمینه شود اما overfitting رخ ندهد). برای یافتن مقدار بهینه می‌توانید از تابع GridSearchCV کتابخانه‌ی Scikit-Learn استفاده کنید. برای اطلاعات بیشتر به این لینک مراجعه کنید.

در نهایت به پرسش‌های زیر پاسخ دهید.

۱. داده‌هایتان را به چه نسبتی برای یادگیری و تست تقسیم کرده‌اید؟ چرا؟ یکبار ۹۸ درصد داده را برای یادگیری و ۲ درصد را برای تست استفاده کنید. نتایج را با حالتی که خودتان تقسیم کردید مقایسه کنید و توضیح دهید. اگر ۴۰ درصد را برای یادگیری استفاده کنید چه اتفاقی می‌افتد؟ نام این پدیده‌ها چیست؟ در مورد آن‌ها کمی توضیح دهید.

۲. اگر max_depth را برای درختتان بسیار زیاد یا بسیار کم کنید چه اتفاقی می‌افتد؟ نمودار تغییرات هر دو معیار خطا را برای داده‌های تست و یادگیری در این بازه رسم و تحلیل کنید (سعی کنید حداقل ۷ مقدار را برای max depth امتحان کنید و در نمودار نمایش دهید.).

فاز سوم: پیشبینی با استفاده از یادگیری گروهی

یادگیری گروهی به این معناست که پیشبینی نهایی را با تجمیع نتایج حاصل از چند مدل انجام دهیم. در این فاز به پیاده‌سازی و تحلیل نتایج مدل Random Forrest می‌پردازیم.

در این مدل، تعدادی Decision Tree ساخته می‌شود که هرکدام جداگانه و با فیچرهای متفاوت آموزش می‌بینند. سپس برای تخمین نهایی بین نتایج درخت‌ها نوعی رای‌گیری انجام می‌شود. در مورد حداقل دو عدد از هایپرپارامترهای این مدل مطالعه کنید و تاثیر تغییر این هایپرپارامترها را روی نتایجتان را با رسم نمودار و ذکر دقیق نتایج بسنجید.

نتایج این مدل را با مدل Decision Tree مقایسه کنید. در مورد bias و variance و ارتباط بین آن ها در این لینک مطالعه کنید. به نظر شما از نظر هر کدام از دو مورد bias و variance یک مدل تنها (Decision Tree) بهتر عمل می‌کند یا یک مدل تجمیعی (Random Forrest)؟ آیا نتایجی که به دست آوردید با نظرتان مطابقت دارد؟

نکات پایانی

۱. دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده مانند نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را به طور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. اگر در جایی ذکر شده مقایسه‌ای انجام دهید، حتما نتایج را دقیق ذکر کنید و سپس آن‌ها را تحلیل و مقایسه کنید.

۲. در همه‌ی بخش‌ها مجازید از متدهای کتابخانه‌ی Scikit-Learn استفاده کنید ولی باید اطلاعات لازم در مورد هر کاری که انجام می‌دهید را داشته باشید، در هنگام تحویل ممکن است در مورد هرکدام از شما سوال پرسیده شود. (به عنوان مثال هر مدل چگونه پیش‌بینی را انجام می‌دهد؟، خطا چطور محاسبه می‌شود؟، هر هایپر پارامتر چه چیزی را تغییر می‌دهد؟ و...)

۳. نتایج و گزارش خود را در یک فایل فشرده با عنوان AI_CA4_<#SID>.zip تحویل دهید. محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل های مورد نیاز برای اجرای آن باشد. توضیح و نمایش خروجی های خواسته شده بخشی از نمره این تمرین را تشکیل می‌دهد. از نمایش درست خروجی های مورد نیاز در فایل html مطمئن شوید.

۴. در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس یا گروه تلگرام مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت از طریق ایمیل با طراحان در ارتباط باشید.

۵. هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید.

موفق باشید!