مدرس: **دکتر فدایی و دکتر یعقوب‌زاده** طراحان: مهسا اسکندری، بهار افشار، تینا بهزاد

مهلت تحویل: چهارشنبه ۱۲ خرداد ۱۴۰۰، ساعت ۲۳:۵۵

مقدمه

در این پروژه، هدف ما تخمین قیمت خودروها در آگهی‌های سایت دیوار است. میخواهیم به کمک کتابخانه‌ی scikit learn و مدل‌های مختلف پیاده‌سازی شده در آن، سعی کنیم برای هر مدل هایپرپارامترهای بهینه را بیابیم و در نهایت نتایج مدل‌ها را مقایسه کنیم.

معرفی مجموعه‌ی داده

مجموعه داده تعدادی آگهی درج شده برای فروش خودرو در سایت دیوار است که در فرمت csv در اختیار شما قرار گرفته است. در این داده، عنوان آگهی، توضیحات مربوط به خودرو، نوع خودرو(خودروی سبک یا سنگین)، زمان درج آگهی، تعداد تصاویر بارگذاری شده در کنار آگهی، کارکرد خودرو و سال تولید آن مشخص شده است.در نهایت یک ستون مربوط به قیمت درج‌شده در آگهی وجود دارد که همان هدف مدل خواهد بود.

فاز اول: پیش‌پردازش و مشاهده[1]ی داده

در فاز اول باید داده‌هاي خام ورودي را به مجموعه‌اي از ویژگی‌هاي قابل پردازش تبدیل کنید. به این منظور لازم است مقادیر خالی ستون ها را حذف کنید، داده‌هایی که دسته‌ای[2] هستند را با روش های مرسوم به داده‌هایی قابل استفاده برای مدلتان استفاده کنید. همچنین ستون‌های حاوی اطلاعات متنی داخل مجموعه داده را برای تحلیل های بعدی پیش‌پردازش کنید. برای این کار می‌توانید از کتابخانه‌ی هضم[3] استفاده کنید یا خودتان موارد مورد نیازتان را پیاده سازی کنید. شما باید توضیحات و تیتر هایی که موجود است را تا حد ممکن Normalize کنید. (روش های ممکن، شامل حذف کلمات پرتکرار یا همان stop words، تبدیل کلمات به ریشه آنها و ... است.) همچنین لازم است رابطه‌ی بین فیچرهایی که دارید و ستون هدف را بررسی کنید.به این منظور information gain را براي ویژگی ها محاسبه کنید و نمودار gain بر

[1] Preprocess and visualization
[2] categorical
[3] https://github.com/sobhe/hazm

حسب ویژگی ها را رسم کنید. (بـرای محاسبه میتوانید از متد mutual_info_regression در کتابخانه‌ی SciKit-Learn استفاده کنید.بخش مربوط به regression در این لینک می تواند به شما کمک کند.)

پرسش ۱) این نمودارها چه اطلاعاتی برای استفاده از فیچرها در ادامه کار میدهند؟ به طور خلاصه توضیح دهید.

هرکاری که برای آماده کردن ستون‌ها برای مدل‌هایتان انجام داده‌اید را حتما در گزارش ذکر کنید.

پرسش ۲) برای تبدیل داده‌های دسته‌ای به اطلاعات قابل استفاده برای مدلتان راه‌های متفاوتی وجود دارد.میتوانید در این لینک تعدادی از آن‌ها را مشاهده کنید. شما کدام روش را انتخاب کردید؟ چرا؟

پرسش ۳) برای پردازش داده‌های متنی، شما میتوانید تعداد مشخصی از کلمات پرتکرار متن را به عنوان فیچر در نظر بگیرید و برای هر سطر داده، مشخص کنید این کلمات چه تعداد بار تکرار شده‌اند(count vectorizer) یا از معیاری مثل tf-idf استفاده کنید که علاوه بر تعداد کلمات، تعداد بار 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 است داشتن یا نداشتن آن کلمه بار اطلاعاتی کمتری دارد. در این لینک میتوانید بیشتر در مورد این دو مدل مطالعه کنید. در یادگیری مدل‌هایتان(حداقل در یکی از مدل‌ها) استفاده از تعداد کلمات مختلف و معیارهای متفاوت را امتحان کنید و نتایج را ذکر کرده و توضیح دهید.

پرسش ۴) در برخی ستون‌ها مقادیر از دست رفته[4] وجود دارد، برای حل این مساله روش‌های متفاوتی وجود دارد، دو تا از این روش‌ها را توضیح دهید و بگویید شما از کدام یک از آن‌ها استفاده کردید. میتوانید در این لینک تعدادی از این روش‌ها را ببینید.

فاز دوم: پیش‌بینی قیمت

در مرحله‌ی اول داده‌ای که در اختیارتان قرار گرفته را به دو بخش یادگیری و تست تقسیم کنید. سپس با استفاده از کتابخانه‌ی scikit-learn با مدل‌های KNN، Linear Regression و Decision Tree مدلتان را با داده‌هایی که برای یادگیری جدا کردید train کنید و سپس به کمک مدل، ستون قیمت را برای داده‌های تست خودتان تخمین بزنید.

برای ارزیابی تخمین‌هایتان از معیار Root Mean Square Error و Mean Square Error استفاده کنید.(هر دو معیار را برای هر مورد ذکر کنید) فرمول هر کدام از این خطا ها در زیر آمده است ولی برای مطالعه‌ی بیشتر میتوانید از این لینک هم استفاده کنید. میزان خطای rmse مطلوب برای این پروژه 10,000,000 میباشد.

$$MSE = \frac{\sum_{i=1}^{n} (y_i - y_i^p)^2}{n}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

برای هر کدام از مدل‌ها جست و جو کنید که هایپرپارامترهای ذکر شده در زیر چه هستند و آن‌ها را تغییر دهید و مقدار بهینه را برای مدلتان پیدا کنید، مقدار بهینه یعنی حالتی که loss مدل کمینه باشد ولی overfitting رخ ندهد.

KNN: n_neighbors

Decision tree: max_depth, min_samples_split

[4] Missing values

در صورتی که اجرای الگوریتم KNN زمان زیادی میگیرد، میتوانید این الگوریتم را روی بخش کمتری از داده اجرا کنید. همچنین آزمودن حداقل ۳ مقدار برای هایپر پارامتر ذکر شده کافی است.

پرسش ۵) داده‌هایتان را به چه نسبتی برای یادگیری و تست تقسیم کرده‌اید؟ چرا؟ برای linear regression یکبار ۹۸ درصد داده را برای یادگیری استفاده کنید و ۲ درصد را برای تست، نتایج را با حالتی که خودتان تقسیم کردید مقایسه کنید و توضیح دهید.اگر ۴۰ درصد را برای یادگیری استفاده کنید چطور؟ نام این پدیده ها چیست؟ در مورد آن‌ها کمی توضیح دهید.

پرسش ۶) اگر max_depth را برای درختتان بسیار زیاد کنید، چه اتفاقی می‌افتد؟ اگر کم کنید چطور؟ نمودار تغییرات هر دو معیار خطا را برای داده‌های تست و یادگیری در این بازه رسم کنید.(سعی کنید حداقل ۷ مقدار را برای max depth امتحان کنید و در نمودار نمایش دهید.) نمودار را تحلیل کنید و بگویید از آن چه برداشتی دارید؟

فاز سوم: استفاده از مدل‌های تجمیعی

در این بخش برای پیش‌بینی قیمت از مدلی استفاده میکنیم که از یادگیری گروهی استفاده می کند به عبارتی از تجمیع نتایج حاصل از تعدادي مـدل، پیشبینی نـهایی را انجام می‌دهند.

به عنوان اولین مدل در این بخش از random forest استفاده کنید. 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 میشود.در مورد حداقل دو تا از هایپرپارامترهای این مدل مطالعه کنید و تاثیر تغییر این هایپرپارامترها را روی نتایجتان بسنجید.(حتما نتایج را دقیق ذکر کنید یا با رسم نمودار نشان دهید) میتوانید از این لینک استفاده کنید.(برای این کار می توانید از متد gridsearchCV کتابخانه‌ی scikit learn استفاده کنید.این لینک می تواند به شما در استفاده از این کتابخانه کمک کند.)

پرسش ۷) 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 و بگویید 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 forest). آیا نتایجی که به دست آوردید با نظرتان مطابقت دارد؟

نکات پایانی

- دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده بطور مثال نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را بطور 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.
- 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 در مورد اینکه هر مدل چگونه پیش‌بینی را انجام میدهد، خطا چطور محاسبه میشود، هر هایپر پارامتر چه چیزی را تغییر میدهد و...)
- 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.

- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت توسط ایمیل با طراحان در ارتباط باشید.
- هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید.