

تمرین سری ۱ واحد درسی داده کاوی

جناب آقای دکتر فراهانی
دستیاران آموزشی : نوید کاشی ، علی شریفی

۱ اسفند ۱۳۹۹

۱ مقدمه

هدف کلی تمرین جاری آشنایی بیشتر با مدل رگرسیون و پیش بینی و چالش های مرتبط با پیش بینی داده است . همانند تمامی پروژه های مرتبط با داده در ابتدا لازم است که بررسی هایی پیرامون داده برای درک بهتر از داده ها صورت بگیرد . در طول فرایند کار با داده ممکن است سوالاتی از جانب خود فرد و یا از جانب صاحب دیتا مطرح شود که این سوالات و فرضیات باید بررسی شوند و تحلیل گردد.

در تمرین جاری ابتدا لازم است داده ها خوانده شود که در این بخش میتوان از پکیج Pandas بهره برد سپس لازم است بر روی داده ها فرایندهایی صورت بگیرد که این فرایندها را به صورت کلی میتوان به صورت زیر مطرح کرد.

۱. حذف داده های پرت

گاهها در داده ها ممکن است داده های تحت عنوان داده های پرت وجود داشته باشند که ممکن است به دلایل مختلفی این داده ها ثبت شده باشند به طور مثال مشکل در سنسورهای ثبت داده و یا اشکال در ورود داده ها و این داده ها معمولا باعث افزایش پیچیدگی مدل میشوند و سبب میشود مدل عمومی بودن را از دست بدهد لذا بهتر است که این داده های پرت از مجموعه داده ها حذف گردد . برای این کار ممکن است استراتژی های مختلفی اتخاذ گردد ، یکی از استراتژی های ساده در این بخش میتواند حذف داده هایی باشد که در فاصله ای بیش از سه برابر انحراف معیار از میانگین وجود دارد . برای این کار میتوانید از پکیج Pandas و متد apply و یا با فیلترکردن دیتافریم صورت بگیرد .

۲. مدیریت Null

در داده ها ممکن است فیچرهایی موجود باشد که آنها ثبت نشده و به اصطلاح Null هستند

که ممکن است به دلیل خطا در هنگام ثبت داده ها و یا عدم وجود اطلاعات از آن فیچر ثبت گردید باشد . راه های مختلفی برای مدیریت این نوع داده ها وجود دارد که عبارتند از :

- حذف این نمونه هایی که یک یا چند فیچر آنها وجود ندارد . این استراتژی میتواند در زمانی که تمامی فیچرهای مهم باشد و حتی عدم وجود یک فیچر اخلاقی در مدل و تحلیل ایجاد کند به کار برود البته باید توجه کنیم در صورتی که تعداد زیادی از نمونه های دارای یک یا چند فیچر خالی باشند بخش زیادی از داده را از دست میدهند که در این حالت بهتر است به سراغ استراتژی های دیگری بروید .
- برای این کار میتوانید از پکیج Pandas و متد dropna() بهره ببرید .
- ممکن است به جای حذف نمونه ها از دیتاست شما فیچرهایی را حذف کنیم که تعداد زیادی از نمونه ها ، فاقد این فیچر میباشند . این کار میتواند باعث شود که تعداد نمونه ها کاهش نیابد ولی ابعاد داده کاهش یابد .
- برای این کار میتوانید از پکیج Pandas و دستور del بهره ببرید .
- گاهی خود Null بودن یک فیچر میتواند برای ما یک اطلاعات ارزشمندی در نظر گرفته شود و Null بودن را به عنوان یک Category در فیچرهای Categorical و در فیچرهای عددی به عنوان یک مقدار در نظر گرفت .
- ممکن است فیچرهای Null را با استفاده از استراتژی هایی پر کنید . این استراتژی ها عبارتند از :

- پر کردن داده های categorical با پر تکرار ترین category.

- پر کردن داده های عددی با استفاده از میانگین مقادیر غیر Null آن فیچر.
- استفاده از یک مدل برای پر کردن این جاهای خالی . در این روش ما از یک مدل همانند رگرسیون برای مقادیر عددی یا رگرسیون لجستیک برای مقادیر categorical و یا شبکه عصبی بهره میبریم . با استفاده از سایر فیچرهایی که null نمیباشند ما مدل را آموزش میدهم یعنی سایر فیچرها به عنوان ورودی مدل و فیچری که میخواهیم آن را پر کنیم به عنوان تارگت در نظر گرفته میشود . سپس در بخش تست ما به پیش بینی یا تخمین مقادیر خالی میپردازیم .

۳. انجام پیش پردازش هایی بر روی داده

این بخش بسیار وابسته به داده ها و تسک های خواسته شده میباشد . در این بخش ممکن است بنا بر صورت مسئله نیاز به انجام فرایندهایی باشد از قبیل تبدیل داده های Categorical اسمی به مقادیر عددی . این کار به این دلیل صورت میگیرد که در مدل ما نیاز به کار با اعداد داریم و string ها برای ما معنی نخواهند داشت .(برای این بخش میتوانید از پکیج sklearn و یا از پکیج Pandas بهره ببرید . ممکن است در بخش دیگری ، نیاز به انجام اسکیل کردن داده ها با استفاده از روش

های مختلف همانند min-max scaling، ... باشد . (برای این بخش میتوانید از پکیج sklearn و یا از پکیج Pandas بهره ببرید .)
اگر در داده ها ما به صورت اسمی باشند گاهی نیاز است فرایندهایی بر روی این string ها انجام گردد همانند این که تمامی حروف lower case شوند و یا غیره .

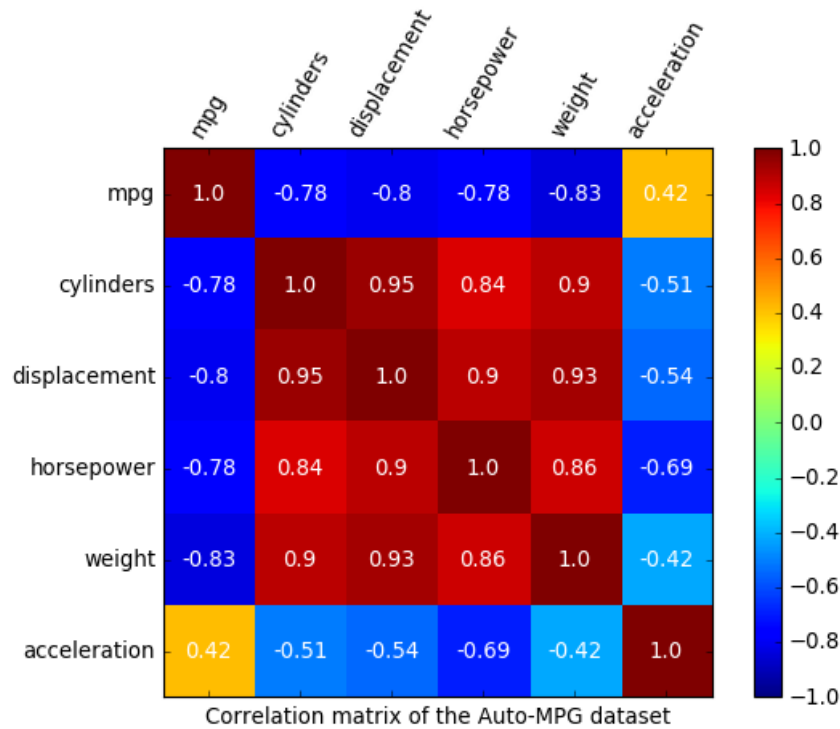
۲ شرح دیتاست

دیتاست تمرین جاری از سایت Kaggle در آدرس زیر
<https://www.kaggle.com/corrieaar/apartment-rental-offers-in-germany>
قابل دسترس است. دیتاست پیرامون متراژ خانه ها در کشور آلمان براساس فیچرها میباشد .
دیتاست شامل ۴۸ فیچر و ۲۶۸,۸۵۰ نمونه میباشد که بخشی از این فیچرها Categorical و بخشی از این داده ها عددی میباشد .
در این دیتاست هدف تخمین متراژ خانه براساس فیچرهای داده شده است یعنی شما ۴۸ فیچر را دارید یعنی از شما میخواهیم که با مدل رگرسیون خود مقادیر livingSpace را تخمین بزنید .
ممکن است در دیتاست فیچرهایی وجود داشته باشند که بی معنی باشند همانند scoutId .
* در صورتی که دیتا های categorical شما زیاد هستند در هنگام استفاده از ایده one hot encoding میتوان با دسته بندی مجدد آنها تعداد category ها را کاهش دهید به طور مثال اگر در دیتاست شما ۳۲۵ شهر وجود داشته باشد میتوان با اضافه کردن اطلاعات جغرافیایی ۳۲۵ شهر را به ۴ دسته بزرگ شهرهای شمالی ، جنوبی ، غربی ، شرقی تقسیم بندی کرد .
* ایده های خلاقانه در جهت کار بر روی داده ها مشمول نمره امتیازی خواهد شد.
* برای یافتن ایده های جدید و خلاقانه میتوانید دیتاست های مشابه همانند دیتاست Boston House Price را مشاهده کنید .

۳ تمرین

در تمرین جاری ابتدا مراحل لازم و مورد نیاز برای دیتاست را که برخی از آنها در بخش مقدمه ذکر شده است را بر روی داده ها انجام دهید. سپس لازم است به ۶ سوال زیر به دقت پاسخ داده شود.

۱. کدام فیچرها بیشترین کورولیشن خطی را با هدف دارا میباشند؟ (برای این کار میتوان نمودار correlation را رسم کنید و سپس به بررسی آنها بپردازید.)
به طور مثال در شکل ۱ اگر horsepower همان تارگت ما باشد فیچر displacement بیشترین رابطه مستقیم و فیچر mpg بیشترین رابطه معکوس را دارد.



شکل ۱: نمونه ای از یک correlation matrix

۲. ۵ فرض بر روی داده ها مطرح کنید و سپس آزمون فرض مناسب آن را مطرح کنید و فرض را تایید یا رد کنید.
به عنوان مثال فرض کنید دیتاست در خصوص قیمت ماشین که شامل فیچرهایی زیر باشد

- شرکت سازنده خودرو (۵ شرکت)

- دنده اتوماتیک بودن یا نبودن خودرو

یکی از فرض می‌تواند به این صورت باشد که آیا شرکت سازنده در قیمت تاثیر دارد یا نه که می‌توان با استفاده از Anova و تحلیل نتایج آن (p-value) به این سوال جواب داد. البته توجه داریم که شرط استفاده از این آزمون فرض نرمال بودن داده‌ها می‌باشد که این مطلب حتما باید در پاسخ ذکر شود.

۳. بدون استفاده از پکیج‌های بدون استفاده از پکیج `scikit-learn` یا پکیج `statsmodels` به پیاده‌سازی یک رگرسیون بپردازید. برای راحت‌تر شدن کار شما استفاده از `numpy` مجاز می‌باشد.

۴. مدل ساخته شده خود را برای حالت‌های مختلفی زیر آزمایش کنید و نتایج را تحلیل کنید و با سایر حالات مقایسه کنید.

- در مدل خود یکبار از همه فیچرها استفاده کنید
- در مدل خود فقط از فیچرهایی استفاده کنید که کورولیشن خطی آنها بیشتر از سایرین است این مقدار را خود شما مشخص می‌کنید به طور مثال فیچرهایی را در نظر بگیرید که کورولیشن خطی آنها با تارگت بیش از 0.7 و کمتر از 0.7- است.
- با استفاده از الگوریتم PCA سعی کنید ابعاد دیتاست را کاهش دهید. (برای سادگی می‌توانید در PCA برای pov مقادیر مختلف 0.7, 0.9, 0.95, 0.99 در نظر بگیرید.)
- Error خود را در مدل تغییر دهید. یعنی یکبار مدل پیاده‌سازی خود بدون پکیج `scikit-learn` و یا پکیج `statsmodels` را با `Square Error` بسازید و یکبار با `Absolute Error` بسازید و مورد (ب) سوال ۶ را برای آنها مقایسه کنید. ارورهای معروف برای مسائل رگرسیون به صورت زیر می‌باشند. تعداد نمونه‌ها N و مقدار واقعی نمونه t ام و y^t مقدار پیش‌بینی شده برای نمونه t ام در نظر می‌گیریم.

– Square Error

$$\frac{1}{2} \sum_{t=1}^N (r^t - y^t)^2 \quad (1)$$

– Absolute Error

$$\frac{1}{2} \sum_{t=1}^N |r^t - y^t| \quad (2)$$

epsilon-sensitive Error –

$$\sum_{t=1}^N 1 (|r^t - y^t| > \epsilon) (|r^t - y^t| - \epsilon) \quad (3)$$

تابع 1 اگر آرگومان ورودیش درست باشد مقدار ۱ برمیگرداند در غیر این صورت مقدار ۰ باز میگرداند .

۵. در اخر مراحل سوال ۴ را مدل های ساخته شده با پکیج `scikit-learn` و یا پکیج `statsmodels` تکرار کنید و نتایج را با مدل خود مقایسه کنید . این مقایسه میتواند از جهت سرعت اجرا و معیار های مطرح شده در سوال (۶) باشد .

۶. برای هر مدل به دو صورت میزان کارایی مدل خود را بیان کنید .

(آ) مقادیر واقعی و مقادیر پیش بینی شده را داریم . برای تک تک نمونه ها خطای `Absolute Error` را حساب کنید سپس از این مقادیر میانگین بگیرید و برای هر مدل این مقدار را گزارش کنید .

(ب) یک بازه ۵ درصدی و یا ۱۰ درصدی پیرامون مقادیر واقعی در نظر بگیرید سپس اگر مقدار پیش بینی شده در این بازه قرار گرفت به آن نمونه لیبل درست و در غیر این صورت لیبل نادرست در نظر بگیرید . سپس دقت را با استفاده از این لیبل ها بیان کنید .

$$\text{accuracy} = \frac{\text{no instances with true labels}}{\text{no total instances}} \quad (4)$$

به طور مثال فرض کنید قیمت واقعی نمونه ۱۰۰ باشد حال بازه ۱۰ درصدی آن میشود یک بازه با کران بالای ۱۱۰ و کران پایین ۹۰ . اگر مقدار پیش بینی شده ما در بازه ۹۰ تا ۱۱۰ قرار گیرد به آن نمونه لیبل درست و در غیر این صورت لیبل نادرست اعمال میکنیم .

۴ نحوه تحویل

نحوه تحویل نتایج و کدها بعد تر اطلاع رسانی میگردد . مهلت تمرین ۱۰ روز میباشد بازه تاخیر برای این پروژه ۳ روز میباشد که تا پایان روز ۱۱ ام ۵ درصد از نمره ، تا پایان روز ۱۲ ام ۱۰ درصد نمره ، تا پایان روز ۱۳ ام ، ۱۵ درصد نمره و پس از روز ۱۴ ام ۱۰۰ درصد نمره شما کسر خواهد شد .