به نام خالق یکتا

تمرین اول مبانی یادگیری ماشین

رگرسیون خطی

پاییز 1402

سوال اول) پیشبینی قیمت خانه

1. مقدمه

در این تمرین هدف پیاده‌سازی رگرسیون خطی با استفاده از دیتاست مربوط به قیمت خانه در بوستون که در درس کمی با آن آشنا شدید، می‌باشد. پیاده سازی باید از ابتدا (from scratch) و بدون استفاده از کتابخانه‌های آماده مربوط به الگوریتم‌های یادگیری ماشین، باشد.

2. معرفی دیتاست

مجموعه داده‌ی مسکن بوستون اطلاعات مربوط به خانه های مختلف را از طریق پارامترهای متعددی به ما می‌دهد. این دیتاست اغلب در تشخیص الگوهای یادگیری ماشین مورد استفاده قرار می‌گیرد. هدف ما پیشبینی قیمت مسکن براساس ویژگی های زیر است که در فایل csv مربوطه موجود است.

- (CRIM): نرخ جرم به ازای هر نفر در شهر
- (ZN): نسبت زمین مسکونی
- (INDUS): نسبت هکتارهای تجاری غیر خرده فروشی در شهر
- (CHAS): متغییر مجازی برخورد خانه به رودخانه چارلز
- (NOX): ترکیبات نیتروژن اکسید
- (RM): متوسط تعداد اتاق‌ها در هر مسکن
- (AGE): نسبت واحد های مسکونی که قبل از سال 1940 ساخته شده است.
- (DIS): فواصل وزنی به پنج مرکز اشتغال بوستون
- (RAD): شاخص دسترسی به بزرگراه ها
- (TAX): نرخ مالیات بر ارزش کامل
- (PTRATIO): نسبت دانش آموزش به معلم در شهر
- (B): نسبت ساکنین سیاه پوست به کل جمعیت منطقه
- (LSTAT): درصد وضعیت پایین جمعیت
- (MEDV): متغیر هدف، قیمت خانه مسکونی

3. پاک‌سازی و پیش‌پردازش

قدم اول پیش پردازش حذف داده های NULL از دیتاست میباشد. شما موظف هستید با استفاده از توابع describe(), info(), isna() سلول های NULL را یافته و در قدم بعد با کمک میانگین یا میانه، مقادیر NULL را جایگذاری کنید. (جایگذاری مقادیر NULL با استفاده از تابع fillna() انجام می‌شود.)

4. مصورسازی

با استفاده از توابعی همچون regplot(), heatmap(), scatterplot() در seaborn میتوانید رابطه هریک از ویژگی هارا با قیمت منازل مشاهده کنید.

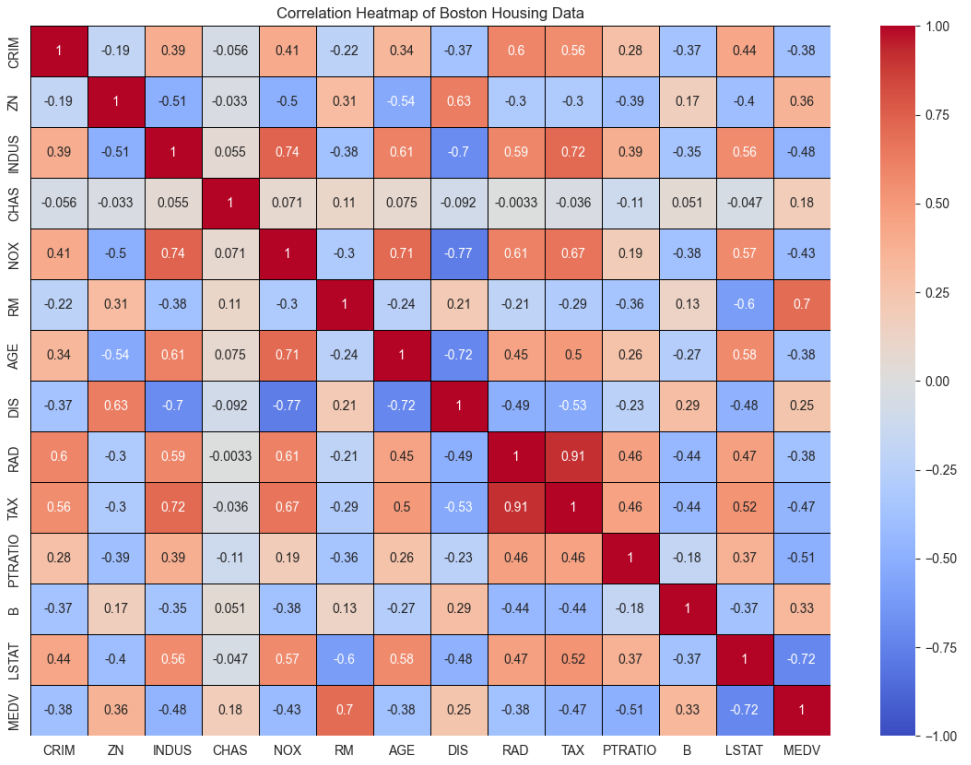

Figure 1 Heatmap

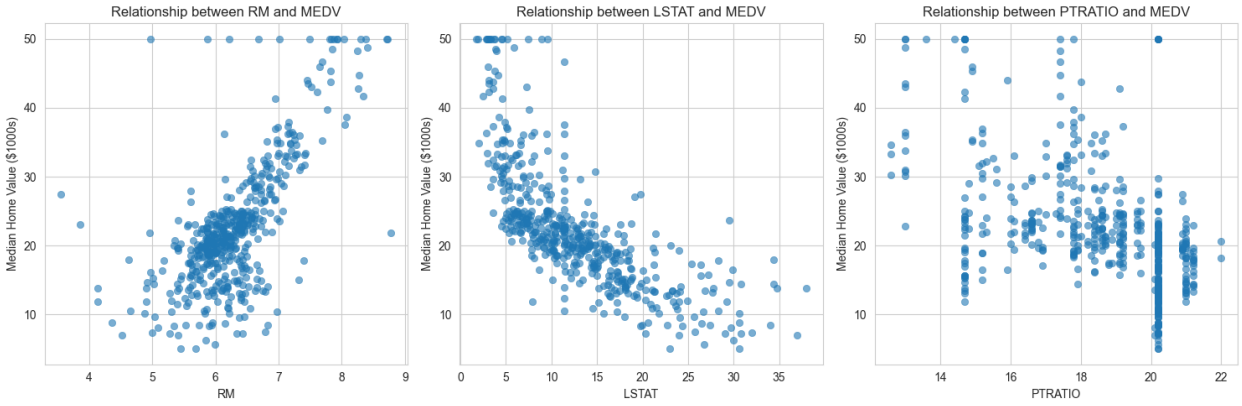

Figure 2 scatterplot

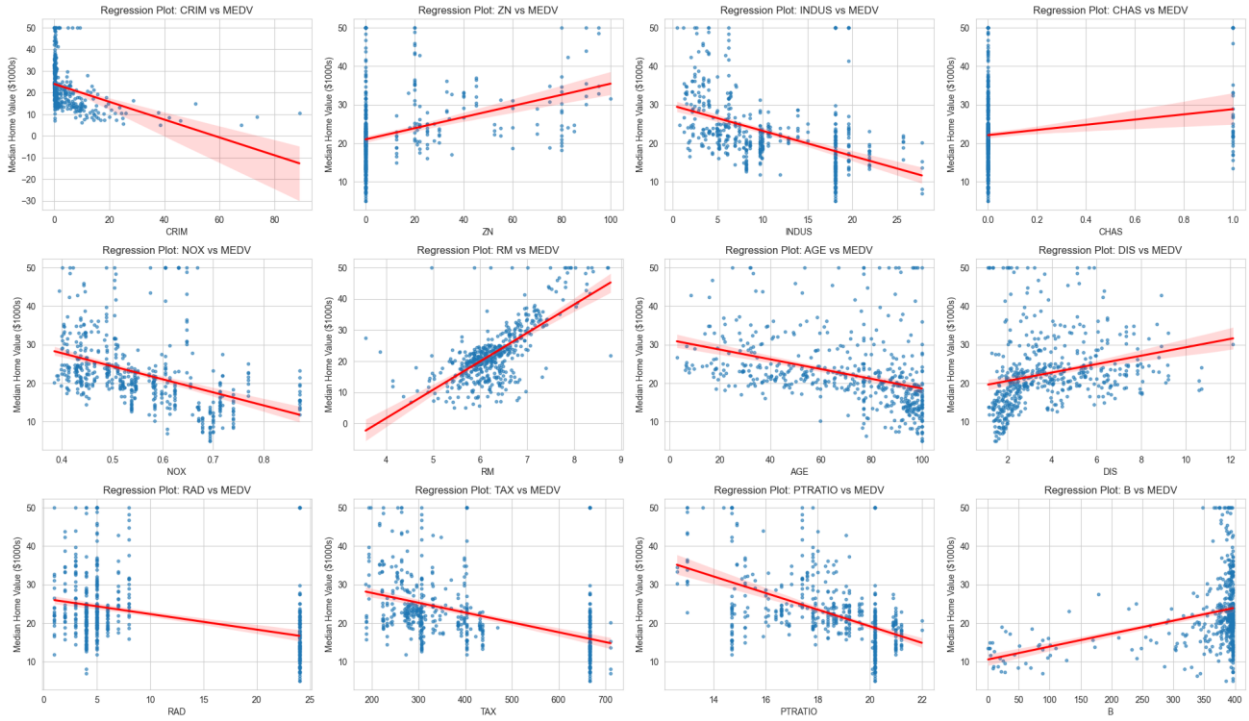

Figure 3 regplot

5. پیاده‌سازی مدل

در این پیاده‌سازی شما میبایست کلاس مربوط به رگرسیون خطی، تابع loss و بروزرسانی با استفاده از gradient descent را از ابتدا پیاده سازی کنید.

پیش از انجام عملیات یادگیری نرمالایز کردن دیتا توصیه میشود.

6. تابع هدف

تابع هدفی که باید برای برازش مدل استفاده کنید، تابع خطای جذر میانگین مربعات (RMSE) می‌باشد که نمره دهی شما نیز بر اساس اندازه همین معیار بر داده های تست در نظر گرفته خواهد شد. در گیت هاب تستی قرار داده شده که اگر خطای جذر میانگین مربعات بین پیشبینی مدل شما و جواب‌های تست، کم تر مساوی 5.2 باشد امتیاز تست را می‌گیرید و در غیر این صورت امتیازی نمی‌گیرید.

7. **آپلود**: در انتها باید نتایج مدل خود روی داده‌های تست را در فایلی به نام house_pred.csv ذخیره کنید و به همراه کد، در ریپازیتوری مربوط به تمرین در گیت‌هاب پوش کنید. دقت کنید که نام گذاری فایل باید دقیقا به شکل گفته شده باشد. همینطور فایل خروجی باید فقط و فقط یک ستون به نام MEDV داشته باشد که به تعداد سطرهای داده تست، یعنی 102 تا پیشبینی دارد.

سوال دوم) پیشبینی آب‌وهوا

1. مقدمه

در این سوال، هدف استفاده از کتابخانه Scikit-learn برای پیشبینی دما با استفاده از یک مدل رگرسیون خطی می‌باشد. دیتاست مورد استفاده در این سوال، دیتاست آب‌وهوای Weather in Szeged 2006-2016 می‌باشد.

2. معرفی دیتاست

داده‌های دیتاست، شامل اطلاعات ساعتی درباره وضعیت آب‌وهوا در منطقه ای در مجارستان بین سال‌های 2006 تا 2016 می‌باشد.

ستون‌های دیتاست عبارتند از:

- Time(زمان): این ستون معمولاً زمانی را نمایش می‌دهد که داده مربوط به آن زمان جمع‌آوری یا ثبت شده است. این ممکن است به صورت تاریخ و ساعت یا فرمت زمان دیگری مثل Unix Timestamp باشد.

- Summary(خلاصه): این ستون به صورت مختصر توضیحی از وضعیت هوا یا رویدادی که در زمان مربوط به ردیف مشخصی اتفاق افتاده است، ارائه می‌دهد. مثلاً "آفتابی" یا "بارانی" می‌تواند اطلاعاتی از نظر آب و هوا را ارائه کند.

- PrecipType(نوع باران): این ستون نوع باران یا برف که در زمان مربوط به ردیف مشخصی رخ داده است را نمایش می‌دهد. ممکن است مقادیری مانند "باران"، "برف" یا "بی‌برنگ" داشته باشد.

- Temperature(دما): دما در این ستون میزان گرمای یا سرمای هوا را به صورت عددی نشان می‌دهد، معمولاً در واحد درجه سلسیوس یا فارنهایت.

- Apparent Temperature(دمای واقعی): این مقدار نشان‌دهنده دمای واقعی حسی است که انسانها در آن شرایط هوا احساس می‌کنند. این مقدار تحت تأثیر عواملی مانند رطوبت و سرعت باد قرار می‌گیرد.

- Humidity(رطوبت): این ستون نشان‌دهنده میزان رطوبت هوا در زمان مشخصی است. معمولاً به صورت درصدی اعلام می‌شود.

- Wind Speed(سرعت باد): این مقدار سرعت حرکت باد را در زمان مربوط به ردیف مشخصی نشان می‌دهد. اعلام معمولاً به صورت متر بر ثانیه (m/s) یا مایل بر ساعت (mph) صورت می‌گیرد.

- Wind Bearing(جهت باد): این ستون جهت یا زاویه‌ای را نمایش می‌دهد که باد در زمان مشخصی از آن جهت به وجود آمده است. این زاویه معمولاً به درجه از شمال (0 درجه) محاسبه می‌شود.

- Visibility(دیدئواری): این ستون معمولاً میزان دیدئواری در زمان مشخصی را نشان می‌دهد. این مقدار به مسافتی که قابل رؤیت در شرایط هوایی مشخص است، ارجاع دارد.

- Cloud Cover(پوشش ابری): این ستون معمولاً مقداری از پوشش ابری را به صورت درصدی نمایش می‌دهد. این مقدار نشان‌دهنده تعداد و میزان ابرها در سمای مشخصی است.

- Pressure(فشار): این ستون معمولاً فشار جوی در زمان مشخصی را نمایش می‌دهد. این فشار معمولاً به واحد هکتوپاسکال (hPa)یا اینچ جیوه‌آمبر (inHg) اعلام می‌شود و نمایانگر وضعیت فشار جوی است.

هدف شما پیشبینی مقدایر Apparent Temperature با داشتن دیگر ویژگی‌هاست.

3. پیاده‌سازی مدل

در این پیاده‌سازی شما باید کلاس مربوط به رگرسیون خطی، تابع loss و بروزرسانی پارامترها با استفاده از gradient descent را از ابتدا پیاده سازی کنید.

پیش از انجام عملیات یادگیری نرمالایز کردن داده توصیه میشود.

4. تابع هدف : تابع هدفی که باید برای برازش استفاده کنید، تابع خطای میانگین مربعات (MSE) می‌باشد که نمره دهی شما نیز بر اساس اندازه همین معیار بر داده های تست در نظر گرفته خواهد شد. در گیت هاب تستی قرار داده شده که اگر خطای میانگین مربعات بین پیشبینی مدل شما و جواب‌های تست، کم تر مساوی 2.0 باشد امتیاز تست را می‌گیرید و در غیر این صورت امتیازی نمی‌گیرید.

5. آپلود: در انتها باید نتایج مدل خود را در فایلی به نام weather_pred.csv ذخیره کنید و به همراه کد، در ریپازیتوری مربوط به تمرین در گیت‌هاب پوش کنید. دقت کنید که نام گذاری فایل باید دقیقا به شکل گفته شده باشد. همینطور فایل خروجی باید فقط و فقط یک ستون به نام Apparent Temperature (C) داشته باشد که به تعداد سطرهای داده تست، یعنی 28937 تا پیشبینی دارد.

آنچه باید تحویل دهید:

- کدهای مربوط به هر دو سوال تمرین به فرمت Ipynb و به شکل نوت بوک شامل توضیحات درباره هر بخش و کار انجام شده، پیش پردازش‌ها و انتخاب‌های مختلف.
- فایل های house_pred.csvو weather_pred.csv. (دقت کنید که نام گذاری فایل ها دقیقا به همین شکل باشد.)
- تحویل تمرین فقط و فقط در گیت هاب کلاس انجام خواهد شد.
- لینک گیت هاب : https://classroom.github.com/a/_qcLooCZ
- آخرین مهلت ارسال جواب، شنبه 13 آبان ساعت 3 بامداد می‌باشد.

موفق باشید