تمرین سری ۱ واحد درسی یادگیری ماشین

جناب آقای دکتر فراهانی

دستیاران آموزشی : نوید کاشی ، علی شریفی

گردآورنده : سیدمازیار مزاری

اسفند ماه 1399

مقدمه

هدف کلی تمرین زیر پیش بینی متراژ خانه(پذیرایی) خانه های کشور آلمان میباشد

برای رسیدن به این مهم ابتدا داده های خانه ها را ابتدا را پاکسازی میکنم و سپس با استفاده از یک مدل رگرسیون خطی به پیش بینی داده ها میپردازیم، در طول این تمرین سوالاتی از جانب خود فرد یا صاحب دیتا مطرح میشود که به بررسی آنها می‌پردازیم

پیش‌پردازش داده ها

ابتدا نگاهی کلی به دیتاست موجود میکنیم، این دیتاست از ۴۹ داده ستون یا ویژگی تشکیل شده است که یکی از آنها متغیر هدف میباشد(living space) می باشد، سایر متغیر ها متغیر مستقل می باشند.

این داده ها شامل ۱۹ متغیر کیفی می‌باشند که به بررسی آنها میپردازیم.

داده های ما دارای مقادیر گمشده می‌باشند که باید برای پردازش حذف یا جایگزین با مقداری شوند که در پخش پاکسازی مقادیر گمشده بررسی میشوند، ابتدا متغیر هدف را بررسی میکنیم و اگر مقدار گمشده در نمونه های ما باشند باید آن نمونه ها را از دیتاست حذف کنیم که خوشبختانه همچین داده هایی موجود نمیباشند.

مقادیر گمشده

برای حذف یا جایگزین کردن مقادیر گمشده باید ابتدا دیتای خود را بررسی کنیم، همه ی ستون هایی که بیش از نصف آنها مقادیر گمشده دارند را حذف میکنیم چراکه نتیجه ی درستی نمیتوان از مدل گرفت همچنین اگر بخواهیم همه ی مقادیر گمشده این ستون هارا حذف کنیم فقط با ۳۳ نمونه سر و کار خواهیم زد که اصلا مناسب نیست، پس از حذف این ستون ها مقادیر گمشده ی عددی را با مقدار میانگین آن ستون پر میکنیم و برای متغیر های کیفی از داده ی پرتکرار(مد) برای جایگزینی داده های کیفی استفاده میکنیم.

حال جداول ما عاری از هرگونه مقدار گمشده میباشند، لازم به ذکر است ستون هایی مانند ملاک خانه و یا توضیحی که برای خانه آمده است کمکی به مدل ما نمیکند و تاثیری در نتیجه ندارد پس آن سری دیتا هارا نیز میتوانیم از دیتاست حذف نماییم.

داده های جغرافیایی

بـرای بررسی داده های جغرافیایی که به صورت کیفی میباشند دو راهکار را پیشنهاد کردیم که بررسی تاثیر آنها بر روی مدل میپردازیم.

اولین راهگار میتواند جایگزینی داده های کیفی با طول عرض جغرافیایی باشد، اینکار بدون افزایش بعد داده ها و اضافه کردن ستون اضافی میتوان داده های کیفی را به کمی تبدیل کرد تا مدل بتواند با آنها کار کنند، اما این شیوه مشکلی دارد و آن نیز این است که مدل به این داده ها بصورت عددی نگاه میکند و نه موقعیت، یعنی برای مثال هرچه طول جغرافیایی بزرگ یا کوچک شود تاثیری بر متراژ خانه ها ندارد و دارای تناوب است.

رویکرد بعدی جایگزینی داده ها با متغیر های ساختگی(dummy variable) می‌باشد، معایب این کار این است که تعداد ستون ها بالا میباشد پس ما فقط از ۱۶ منطقه استفاده میکنیم و باقی داده های جغرافیایی را حذف میکنیم.

نرمال سازی و مقیاس بندی داده ها

نرمال سازی و اسکیل کردن داده ها به ما کمک میکنند تا داده ها مقیاس تقریبا برابری با یکدیگر داشته باشند، برای اینکار از مقیاس بندی استاندارد یا استاندارد اسکیلینگ استفاده کردیم، مقدار هر متغیر در هر ستون را از میانگین آنها کم کرده و سپس تقسیم بر انحراف معیار میکنیم، این کار باعث میشود که داده های نزدیک به میانگین به صفر میل کنند و میانگین صفر میشوند، و مقدار هر داده تعداد واحدهایی که از میانگین دور است را به ما نشان می‌دهد.

حذف داده ها پرت

برای حذف داده های پرت، داده هایی را که بیشتر از ۳ انحراف معیار از میانگین داده ها دور است را حذف میکنیم که ابعاد دیتاست را خیلی کم نمیکند و تنها ۶ درصد از داده ها از بین میرود اما کمک زیادی به مدل میکند و دقت را افزایش می‌دهد.

کوریلیشن

ماتریس کوریلیشن را رسم می‌کنیم. این ماتریس رابطه ی هر متغیر را با متغیر هدف نشان می‌دهد. علت پایین بودن رنج اعداد بدلیل این می‌باشد که متغیر های مستقل مقیاس بندی شده اند و دارای داده ی پرت نمی‌باشند، اما متغیر وابسته یا هدف مقیاس بندی نشده است.
بیشترین کوریلیشن هم با هدف متغیر living space range دارد که منطقی می‌باشد.

PCA

ابعدا دیتاست پس از اضافه کردن متغیر های ساختگی خیلی زیاد شد که یکی از روش های کاهش بعد برای این عملیات می‌باشد. ما برای انجام این کار ۹۰ درصد داده ها را انتخاب کردیم تا حفظ شوند، که ابعاد این داده ها به ۲۸ ستون یا کامپوننت کاهش پیدا می‌کنند.

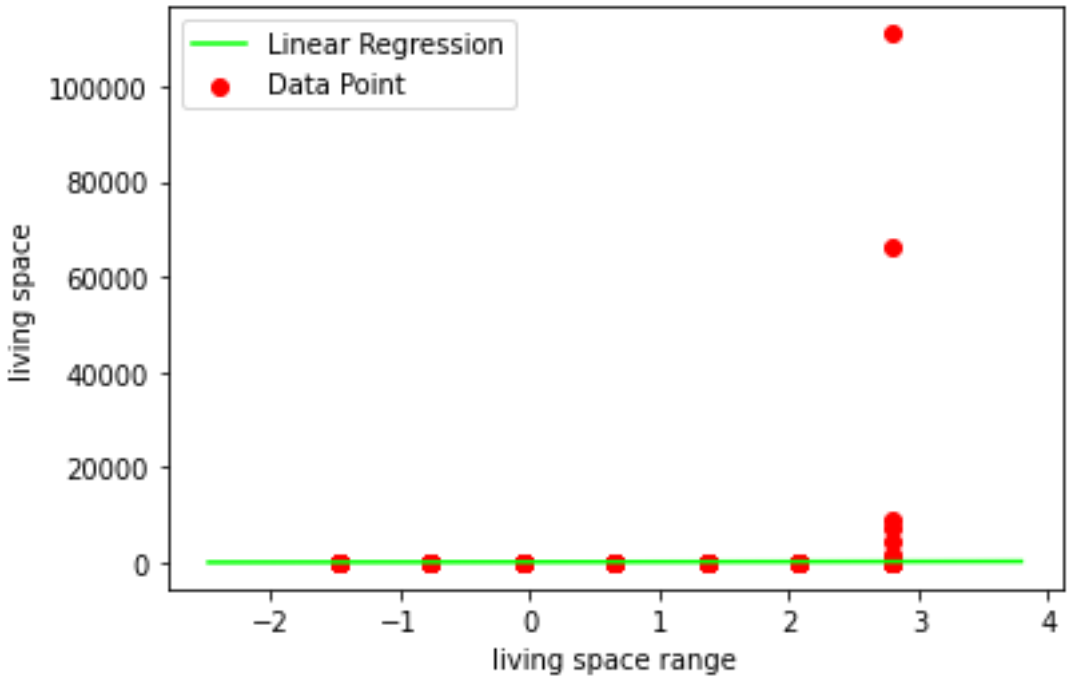
جدا سازی داده ها

داده ها را به دو بخش داده ها آموزشی(train) و تست تقسیم میکنیم که ۸۰ درصد آموزشی هست و ۲۰ درصد داده های تست، از داده های آموزشی برای آموزش مدل و از داده های تست جهت ارزیابی مدلی که توسط داده های تست آموزش داده شده استفاده می‌شود.

مدل رگرسیون

ابتدا مدل را تنها با یک متغیر که بیشترین کوریلیشن را با هدف دارد آموزش میدهیم، ارزیابی این مدل توسط خطای mse بررسی میشود، توزیح مدل به شرح زیر می‌باشد.

```
b1 = 32.8702789596806 b0 = 74.4196739327577
```




```
2374.027647218314 = test mse erro
```

حال برای مدل رگرسیون با بیش از یک متغیر از روش گرادیان کاهشی استفاده میکنیم، دلیل استفاده از گرادیان کاهش ابعاد بسیار مدل می‌باشد، چرا که اگر به روش جبر خطی و همه ی مدل را در یک عملیات حساب کنیم بار محاسباتی بسیار زیاد می‌شود.معیار سنجش برای این مدل هم mse می‌باشد.

	test	pred
215173	108.00	122.208036
215174	52.80	52.366281
215175	58.00	50.925901
215176	83.05	96.743490
215177	105.00	111.792276

```
scratch_linear_mse = 2376.486758884171
```

در آخر نیز از مدل کتابخانه ی sklearnاستفاده میکنیم که به شرح زیر است:

```
mean_squared_error = 2386.0118124073633
```

ارزیابی مدل

خطای حاصل از مدل آماده ی پکیج sklearn از رگرسیونی که به پیاده سازی کردیم کمی بیشتر است و خطای مدل دستی کمتر می باشد.

اگر ۱۵ درصد بالا یا پایین مقدار متغیر هدف را در نظر بگیریم و اگر مقادیر پیش‌بینی شده در این بازه قرار گرفت به آن مقدار درست و در غیر این صورت مقدار نادرست به آن میدهیم، این مقدار را دقت می‌نامیم.

دقت مدل پکیج sklearn ۷۰ درصد می‌باشد.