دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده علوم کامپیوتر

تمرین سری سوم

مهندسی ویژگی‌ها شامل کاهش ویژگی، انتخاب ویژگی و استخراج ویژگی

نگارش
ثمین مهدی پور
۹۸۳۹۰۳۹

استاد راهنما
دکتر قطعی

استادمشاور
دکتر یوسفی مهر

مهر ۱۴۰۲

چکیده

در این گزارش مجموعه داده ای با هدف بررسی تاثیر روش های مهندسی ویژگی شامل کاهش ویژگی، انتخاب ویژگی، استخراج ویژگی پس از اعمال پیش پردازش و پاکسازی داده ها مورد تحلیل قرار گرفته است. در گام اول داده ها مورد بررسی قرار گرفته و داده های گم شده و خارج از محدوده مدیریت شده اند. در گام بعدی پس از اعمال سه الگوریتم رگرسیون خطی، درخت تصمیم و جنگل تصادفی اخرین روش را انتخاب کرده و چگونگی تاثیرگذاری مهندسی ویژگی هارا بر روی آن با معیار های انتخاب شده بررسی کرده ایم.

واژه‌های کلیدی:

انتخاب ویژگی، استخراج ویژگی، پاکسازی داده‌ها، کاهش ویژگی، مهندسی ویژگی

فهرست مطالب

صفحه

فهرست اشکال

صفحه

فصل اول

مقدمه

مقدمه

۱-۱-پیش پردازش

در مواجه با یک مجموعه داده قبل از هرکاری نیاز است آن را مورد تحلیل قرار داده و پس از بررسی و اعمال تست های مختلف داده هارا پالایش کنیم که به این فرایند پیش پردازش گفته میشود. با انجام این روش ها امکان کاوش هدفمند بر روی داده ها افزایش می‌یابد. در ادامه به بررسی گام های طی شده در این پروژه میپردازیم:

- در اولین گام داده هارا مورد بررسی آماری قرار دادیم. به این معنا که برای هر ویژگی تعداد، حداقل، حداکثر، میانگین و موارد مشابه را بررسی و تصویر سازی کردیم. با توجه به درک بدست امده از این بخش نیاز به مدیریت چند مورد بود

۱- مدیریت داده های گم شده: داده هایی که خالی بودند و برای آنها مقداری مشخص نشده بود، با توجه به ویژگی مورد بررسی برای این داده ها تصمیم گیری شد و مقادیر لازم جایگزین شدند.

۲- مدیریت داده های تکراری: پس از بررسی به عمل آمده داده تکراری ای در این مجموعه داده یافت نشد.

۳- مدیریت داده های خارج از محدوده: داده های خارج از محدوده که عموما باعث انحراف از هدف مدل میشود باید مدیریت میشدند که این کار با شناسایی و سپس حذف آنها انجام شد.

در نهایت میتوان اعمال روش های بالا را به نوعی پاکسازی داده ها درنظر گرفت که امکان کار با داده را در گام های بعدی برای ما تسهیل میکند

نهایتا داده های ویژگی های عددی نرمالسازی شدند که برای پردازش قابل فهم و درک تر باشند.

- در گام بعدی داده ها را به دو بخش آموزش و آزمون به نسبت ۸۰ به ۲۰ تقسیم کردیم. با توجه به مجموعه داده مدنظر و انتخاب ویژگی هدف سه الگوریتم رگرسیون خطی، درخت تصمیم و جنگل تصادفی انتخاب و معیار های خطای میانگین مربعات[۱] و ضریب تعیین[۲] را برای بررسی چگونگی عملکرد آنها انتخاب کردیم. این معیارها بصورت زیر تعریف میشوند:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

۱-۲- کاهش ویژگی

در مسائل پردازش داده و یادگیری ماشین، اغلب با مجموعه‌های داده حاوی تعداد زیادی ویژگی (یا ویژگی‌های) روبرو هستیم که ممکن است برخی از آنها اطلاعات تکراری یا غیرضروری داشته باشند. کاهش ویژگی به معنای کاهش تعداد ویژگی‌ها با حفظ حداکثر اطلاعات ممکن است. این عمل به منظور افزایش کارآمدی و کارایی مدل‌های یادگیری ماشین صورت می‌گیرد. در این پروژه ما از PCA , SelectFromModel استفاده کرده و تاثیر آن را روی خطای ناشی از linear regression بررسی کردیم.

[۱] Mean Square Error

[۲] R Squared

در ادامه بصورت خلاصه چگونگی عملکرد این دو شیوه را توضیح خواهیم داد:

PCA -۱-۲-۱

PCA یا Principal Component Analysis یک روش محاسباتی برای کاهش ابعاد داده‌ها است. هدف اصلی این روش، کاهش تعداد ویژگی‌ها یا ویژگی‌های مستقل موجود در داده به یک تعداد کمتر ویژگی است که به عنوان اجزای اصلی یا Principal Components شناخته می‌شوند.

مراحل کلی PCA به شرح زیر است:

۱. محاسبه ماتریس همبستگی (Covariance Matrix): ابتدا، ماتریس همبستگی بین تمام ویژگی‌های داده محاسبه می‌شود. این ماتریس نشان‌دهنده ارتباطات و وابستگی‌های مختلف میان ویژگی‌هاست.

۲. محاسبه ویژه‌بردارها و ارزهای ویژه (Eigenvectors and Eigenvalues): با محاسبه ویژه‌بردارها و ارزهای ویژه ماتریس همبستگی، اجزای اصلی یا Principal Components مشخص می‌شوند. این اجزا به نحوی انتخاب می‌شوند که ارتباط ماکزیمم و وابستگی کمترین ممکن باشد.

۳. انتخاب اجزا و ایجاد ماتریس تبدیل (Selecting Components and Creating Transformation Matrix): بر اساس ارزهای ویژه، اجزا اصلی انتخاب می‌شوند. ماتریس تبدیلی همچنین با استفاده از این اجزا ساخته می‌شود.

۴. تبدیل داده به اجزا اصلی (Transforming Data to Principal Components): داده‌های اولیه با استفاده از ماتریس تبدیل به اجزا اصلی تبدیل می‌شوند. این مرحله باعث کاهش ابعاد داده می‌شود.

PCA به عنوان یک تکنیک کلی مورد استفاده قرار می‌گیرد و در مواردی که تعداد ویژگی‌ها زیاد است یا در صورت وجود همبستگی بین ویژگی‌ها، کارآیی بالایی دارد.

۱-۲-۲- SelectFromModel

یک کلاس در scikit-learn است که برای انتخاب ویژگی‌های مهم از یک مدل استفاده می‌شود. این روش بر اساس اهمیت وزنی ویژگی‌ها در یک مدل یادگیری ماشین انجام می‌شود. در صورتی که اهمیت یک ویژگی بیشتر از حد آستانه‌ای (threshold) باشد، آن ویژگی انتخاب می‌شود.

۱-۳- انتخاب ویژگی

در این رویکرد، تلاش برای انتخاب زیرمجموعه‌ای از ویژگی‌ها به منظور استفاده از آنها در مدل یادگیری ماشین انجام می‌شود. این انتخاب معمولاً بر اساس معیارهای مختلفی نظیر اهمیت ویژگی‌ها، اطلاعات متقاطع و یا سایر معیارهای مشابه صورت می‌پذیرد. در این پروژه ما از PCA, Variance Thresholding استفاده کردیم. در ادامه بصورت مختصر این دو روش را توضیح میدهیم:

۱-۳-۱- PCA

تکنیک کاهش ابعاد PCA (تجزیه و تحلیل مؤلفه‌های اصلی) یک روش معمول در انتخاب ویژگی‌ها در مسائل یادگیری ماشین است. این روش به صورت خلاصه به کاهش تعداد ویژگی‌های یک مجموعه داده می‌پردازد در حالی که تلاش می‌کند اطلاعات مهم و اصلی مجموعه داده را حفظ کند .

مراحل کلی اجرای PCA به شرح زیر است:

۱ .استانداردسازی داده: ابتدا داده‌ها را استانداردسازی می‌کنیم تا همگی در یک مقیاس باشند.

۲ .ماتریس کوواریانس: ماتریس کوواریانس داده‌ها را محاسبه می‌کنیم.

۳ .مقادیر و بردارهای ویژه: مقادیر و بردارهای ویژه ماتریس کوواریانس را استخراج می‌کنیم.

۴.انتخاب تعداد مؤلفه‌ها: تعداد مؤلفه‌های موردنظر را انتخاب می‌کنیم که معمولاً بر اساس مقادیر ویژه بزرگتر مشخص می‌شود.

۵ .تبدیل داده به فضای اصلی: داده‌ها را با استفاده از مؤلفه‌های اصلی تبدیل می‌کنیم.

۶.انتخاب ویژگی‌ها: مؤلفه‌های اصلی که مرتبط‌ترین اطلاعات را حاصل می‌کنند به عنوان ویژگی‌های اصلی مدل انتخاب می‌شوند.

۱-۳-۲- Variance Thresholding

Variance Thresholdingیکی از روش‌های مرسوم در انتخاب ویژگی در مسائل یادگیری ماشین است که بر اساس واریانس (پراکندگی) ویژگی‌ها عمل می‌کند. هدف اصلی این روش، حذف ویژگی‌هایی است که واریانس کمی دارند و اطلاعات مفید زیادی را به مدل ارائه نمی‌دهند. واریانس نشان‌دهنده اندازه پراکندگی داده‌ها است؛ به عبارت دیگر، میزان تغییرات مقادیر یک ویژگی.

مراحل اجرای Variance Thresholding به شرح زیر است:

۱ .محاسبه واریانس: برای هر ویژگی، مقدار واریانس محاسبه می‌شود.

۲ .تعیین یک آستانه :(Threshold) یک آستانه (حداقل مقدار واریانس مجاز) تعیین می‌شود. ویژگی‌هایی که واریانس آن‌ها از این آستانه کمتر است، حذف می‌شوند.

۳.حذف ویژگی‌های با واریانس کم: تمام ویژگی‌هایی که واریانس آن‌ها از آستانه کمتر است، حذف می‌شوند.

این روش به خصوص در مواقعی مفید است که ویژگی‌هایی با تغییرات کم در داده‌ها وجود داشته باشند و احتمال آنکه این ویژگی‌ها اطلاعات مهمی به مدل ارائه دهند کم باشد. از این روش می‌توان در پیش‌پردازش داده‌ها قبل از اعمال مدل‌های یادگیری ماشین برای افزایش سرعت آموزش مدل‌ها و کاهش پیچیدگی آن‌ها استفاده کرد.

۱-۴- استخراج ویژگی

در این رویکرد، سعی بر این است که اطلاعات مهم و مفید از مجموعه ویژگی‌های اصلی استخراج شوند و به صورت ویژگی‌های جدید یا تبدیل‌شده نمایش داده شوند. این فرایند ممکن است با استفاده از تکنیک‌های مختلفی نظیر تحلیل مؤلفه اصلی (PCA) یا تبدیل موجک انجام شود. در ادامه روش های استفاده شده را بصورت مختصر معرفی میکنیم:

۱-٤-۱- PCA

استخراج ویژگی با استفاده از تجزیه و تحلیل مؤلفه‌های اصلی یا PCA یکی از روش‌های متداول در پردازش سیگنال و تحلیل داده‌ها است. درPCA ، هدف اصلی تبدیل مجموعه‌ای از داده‌های اولیه به مجموعه‌ای از ویژگی‌های جدید به نحوی است که بیشتر اطلاعات مهم داده‌ها در ویژگی‌های جدید حفظ شود و اطلاعات اضافی یا تکراری حذف گردد.

۱-٤-۲- تبدیل موجک

تبدیل موجک یک روش مؤثر برای استخراج ویژگی از سیگنال‌ها و داده‌ها است. این روش از ایده‌های تئوری موجک (Wavelet Theory) برای تجزیه و تحلیل سیگنال‌ها استفاده می‌کند. در اینجا به صورت خلاصه، مراحل تبدیل موجک و کاربردهای آن در استخراج ویژگی را توضیح می‌دهم:

تبدیل موجک یک بعدی:

برای یک سیگنال یک بعدی، تبدیل موجک اطلاعات آن را در دامنه زمان-فرکانس تجزیه و تحلیل می‌کند. تبدیل موجک یک بعدی به دسته‌هایی از سیگنال‌های موجک تجزیه می‌شود، که هر کدام ویژگی‌های خاص خود را از سیگنال اصلی استخراج می‌کنند.

تبدیل موجک دو بعدی:

در مورد تصاویر و داده‌های دو بعدی، تبدیل موجک به دسته‌هایی از تصاویر موجک منجر می‌شود. این تصاویر موجک نشان‌دهنده ویژگی‌های مختلف تصویر در فرکانس‌ها و مقادیر مختلف زمان یا مکان هستند.

کاربردهای استخراج ویژگی:

- کاهش ابعاد: تبدیل موجک می‌تواند منجر به کاهش ابعاد داده‌ها شود، به خصوص در صورتی که اطلاعات مهم در برخی از فرکانس‌ها یا مقادیر زمانی نقش مهمی داشته باشند.

- تشخیص لبه: تبدیل موجک برای تشخیص لبه‌ها و ساختارهای جزئی در سیگنال‌ها و تصاویر بسیار مفید است.

- حذف نویز: قابلیت تبدیل موجک در تفکیک اطلاعات مهم از نویزها، به ویژه در داده‌های غیرمنظم و آشفته، موجب می‌شود که این روش به عنوان یک ابزار قدرتمند در پیش‌پردازش داده‌ها استفاده شود.

- تحلیل فرکانسی: تبدیل موجک به خوبی می‌تواند تغییرات فرکانسی در داده‌ها را مدیریت و تجزیه و تحلیل کند.

- تطبیق‌پذیری: به دلیل قابلیت تطبیق‌پذیری به ساختارهای مختلف، تبدیل موجک در بسیاری از حوزه‌ها از جمله پردازش تصویر، سیگنال‌های زمانی و داده‌های چند بعدی مورد استفاده قرار می‌گیرد.

فصل دوم

پیاده‌سازی

۲-۱-معرفی مجموعه داده

این مجموعه داده Airbnb برای سال ۲۰۱۹ در شهر نیویورک، نیازمندی‌ها و فعالیت‌های مربوط به رزروها را شامل می‌شود. این داده به ما اطلاعات کاملی از میزبان‌ها، دسترسی جغرافیایی به مکان‌ها، و معیارهایی برای پیش‌بینی و استنباط فراهم می‌کند. از این مجموعه داده می‌توان سوالاتی را بررسی کرد، از جمله مفهوم میزبان‌ها و مناطق مختلف، استفاده‌های ممکن از پیش‌بینی‌ها (مانند: موقعیت‌ها، قیمت‌ها، نقدها و غیره)، شناخت میزبان‌های پرکار و دلایل فعالیت بیشتر آن‌ها، و تفاوت‌های قابل مشاهده در ترافیک بین مناطق و دلایل احتمالی آن. این مجموعه داده یک منبع ارزشمند برای کسانی است که به دنبال درک بهتر از فعالیت‌های Airbnb در نیویورک هستند.

در بررسی اولیه مشاهده شد که این مجموعه داده حاوی ۱۶ ستون و ۴۸۸۹۵ سطر است که در ادامه پیش پردازش روی آنها انجام شد. ویژگی هدف در این مجموعه قیمت در نظر گرفته شد پس وظیفه مدل ما تخمین زدن خانه Airbnb بوسیله روش های رگرسیون بود.

۲-۲- پیش پردازش و پاکسازی داده‌ها

۲-۲-۱- مدیریت داده های گم شده

ابتدا بررسی کردیم که مقدارهای گم شده در مجموعه داده به چه صورت است که مشخص شد ویژگی های name, host_name که مقادیر غیرعددی داشته به ترتیب ۱۶ و ۲۱ داده گم شده دارند که به نسبت داده کلی قابل صرف نظر بود پس این سطرها حذف شدند.

ویژگی های last_review , reviews_per_month اما ۱۰۰۵۲ داده گم شده داشتند که قابل صرف نظر نبود. پس از تبدیل ویژگی last_review به قالب تاریخی زمانی معمول و بررسی همبستگی آن با دیگر ویژگی ها مشاهده کردیم که این ویژگی اصلا در ماتریس همبستگی محاسبه شده حضور ندارد پس با صرف نظر کردن از احتمال رابطه داشتن این ویژگی با دیگر ویژگی ها این ستون را حذف کردیم.

اما برای ستون reviews_per_month به علت وجود همبستگی با دیگر ویژگی ها مقدار میانگین جایگزین مقادیر گم شده شد.

۲-۲-۲- مدیریت داده های تکراری

پس از بررسی داده ها مشخص شد که در این مجموعه داده، داده تکراری ای موجود نمیباشد پس نیاز به اقدامی نبود.

۲-۲-۳- مدیریت داده های خارج از محدوده

در این بخش با استفاده از باکس پلات ها و شیوه IQR داده های خارج از محدوده عددی را شناسایی کردیم. تعداد داده ها ۴۸۸۹۵ سطر بود و تعداد داده های خارج از محدوده در چند ویژگی بصورت زیر بود:

- Price : 2971
- Minimum_nights : 6605
- reviews_per_month : 4099
- calculated_host_listings_count : 7080

که در مقایسه با ۴۸۸۹۵ داده قابل صرف نظر بودند پس این داده ها را حذف کردیم. چگونگی پخش شدن داده قبل و پس از حذف کردن داده های خارج از محدوده در کد نمایش داده شده و مورد مقایسه قرار گرفته اند (توسط باکس پلات و نمودار های هیستوگرام)

۲-۲-۳- نرمال سازی

در این مجموعه داده بنظر میرسید که اعمال نرمالسازی برای ستون هایی مانند latitude، longitude، price، minimum_nights، number_of_reviews، reviews_per_month، calculated_host_listings_count و availability_365 مفید باشد پس با استفاده از MinMaxScaler نرمالسازی روی این ستون ها انجام و چگونگی توزیع آنها قبل و پس از نرمالسازی توسط نمودارهای هیستوگرام در کد نمایش و مقایسه شد.

۲-۳- رگرسیون

پس از انتخاب ستون قیمت به عنوان ویژگی هدف که قصد داشتیم آن را تخمین بزنیم سه روش رگرسیون (رگرسیون خطی، درخت تصمیم و جنگل تصادفی) را پس از تقسیم کردن داده ها به نسبت ۲۰ به ۸۰ برای داده آزمون و آموزش انتخاب کردیم.

خروجی بصورت زیر بود:

	Linear Regression	Decision Tree	Random Forest
Mean Square Error	0.047	0.031	0.02
R Squared	-0.167	0.234	0.51

۲-٤-کاهش ویژگی

با استفاده از دو شیوه SelectFromModel , PCA که پیش تر توضیح داده بودیم جلو رفتیم، در ابتدا قصد داشتیم تاثیر آن را بر روی جنگل تصادفی بررسی کنیم و این کار را بر روی PCA کردیم ولی اعمال روش دوم روی آن بسیار زمان بر بود بنابراین هردو روش مجددا روی رگرسیون خطی نیز مورد بررسی قرار گرفتند:

PCA

	Linear Regression	Random Forest
Mean Square Error	0.023	0.021
R Squared	0.423	0.47

SelectFromModel

	Linear Regression	Random Forest
Mean Square Error	0.040	-
R Squared	3.28 * 10^-5	-

همانطور که مشاهده میشود شیوه اول تاثیر منفی روی خروجی جنگل تصادفی داشته ولی شدیدا خطای رگرسیون خطی را کاهش داده و آن را نزدیک خروجی جنگل تصادفی آورده با اینکه میدانیم جنگل

تصادفی بسیار زمانبر و پیچیده تر است. پس PCA بخوبی عمل کرده است. در مـورد شـیوه دوم بهبـود چشم گیری مشاهده نمیشود.

۲-۵- انتخاب ویژگی

با استفاده از دو شیوه Variance Thresholding , PCA که پیش تر توضیح داده بودیم جلو رفتیم، هردو روش روی رگرسیون خطی و جنگل تصادفی مورد بررسی قرار گرفتند:

PCA

	Linear Regression	Random Forest
Mean Square Error	0.032	0.028
R Squared	0.151	0.31

Variance Thresholding

	Linear Regression	Random Forest
Mean Square Error	0.027	0.019
R Squared	0.308	0.51

مشاهده میشود که PCA اثر منفی روی جنگل تصادفی ولی مثبت روی رگرسیون خطی داشته اما تاثیر شیوه دوم به مراتب بیشتر است و تاثیر شدیدا مثبتی روی کاهش خطای هردو الگوریتم داشته است.

۲-۶-استخراج ویژگی

با استفاده از دو شیوه PCAو تبدیل موجک که پیش تر توضیح داده بودیم جلو رفتیم، هردو روش روی رگرسیون خطی و جنگل تصادفی مورد بررسی قرار گرفتند:

PCA

	Linear Regression	Random Forest
Mean Square Error	0.034	0.028
R Squared	0.151	0.31

Wavelet transform

	Linear Regression	Random Forest
Mean Square Error	0.035	0.034
R Squared	0.112	0.154

مشاهده میشود که PCA عملکرد به مراتب بهتری از تبدیل موجک داشته و هردو اثری منفی بر جنگل تصادفی ولی مثبت بر رگرسیون خطی داشته اند.

فصل سوم

جمع‌بندی

پس از بررسی ها میتوان گفت:

- در شیوه کاهش ویژگی بهترین عملکرد را PCA روی رگرسیون خطی داشته است
- در شیوه انتخاب ویژگی بهترین عملکرد را Variance Thresholding روی هردو داشته است.
- در استخراج ویژگی عملکرد PCA بهتر از تبدیل موجک بوده است.
- در نهایت بهترین عملکرد روی هردو الگوریتم رگرسیون را Variance Thresholding در انتخاب ویژگی داشته است.
- بنظر میرسد الگوریتم پیچیده و زمان بری مانند جنگل تصادفی از روش های مهندسی ویژگی بهره چندانی نمیبرد و حتی ممکن است که از نظر زمانی و هزینه ای شرایط را پیچیده تر کند ولی اعمال این ویژگی ها روی الگوریتم ساده ای مثل رگرسیون خطی مارا در زمان و با پیچیدگی محاسباتی بسیار کمتر به خروجی ای نزدیک جنگل تصادفی میرساند.

منابع و مراجع

[1] https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

پیوست

کد به همراه گزارش و مجموعه داده قرار داده شده و لینک زیر کد روی کولب است. ترجیحا از فایل و دیتاست برای بررسی استفاده کنید.

- https://colab.research.google.com/drive/1BFjFl8WnZPr_3ssJIIqXHdrCMUd9LlYi?usp=sharing
- Dataset: https://docs.google.com/spreadsheets/d/1cdYqmVJLADsjan0lModr_IFvSOlFE2F31lIeQoI_1As/edit?usp=drive_link
- Code:

  https://drive.google.com/file/d/1ZThN_iO6DoiY4hRNyhP2qgkmfdNAlORt/view?usp=drive_link
- Github:

  https://github.com/Precioux/Data-Mining/tree/master/Data%20Cleaning%20and%20Feature%20Engineering

Abstract

In this report, a dataset has been analyzed with the aim of investigating the effect of feature engineering methods, including feature reduction, feature selection, feature extraction after applying pre-processing and data cleaning. In the first step, the data was examined and the missing and out-of-range data were managed. In the next step, after applying three linear regression algorithms, decision tree and random forest, we have chosen the last method and we have checked how the feature engineering affects it with the selected criteria.

Key Words: Feature selection, Feature extraction, Data cleaning, Feature reduction, Feature engineering

Amirkabir University of Technology
(Tehran Polytechnic)

Math and Computer Science

Data Mining Course - Project 3

Feature engineering includes feature reduction, feature selection and feature extraction

By
Samin Mahdipour

Supervisor
Dr. Ghatee

Advisor
Dr. Yousofi Mehr

October 2023