رگرسیون

تمرین درس داده کاوی بخش دوم – استاد مهردوست

کیان رضایی

رگرسیون چیست؟

- در مدل‌های آماری، تحلیل رگرسیون یا تحلیل ارتباط یک فرایند آماری برای تخمین روابط بین متغیرها است. این روش شامل تکنیک‌های زیادی برای مدل‌سازی و تحلیل متغیرهای خاص و منحصر بفرد، با تمرکز بر رابطه بین متغیر وابسته و یک یا چند متغیر مستقل، است. تحلیل رگرسیون کمک می‌کند در فهم اینکه چگونه مقدار متغیر وابسته با تغییر هرکدام از متغیرهای مستقل و با ثابت بودن دیگر متغیرهای مستقل تغییر می‌کند.
- بیشترین کاربرد تحلیل رگرسیون تخمین امید ریاضی شرطی متغیر وابسته از متغیرهای مستقل معین است که معادل مقدار متوسط متغیر وابسته است وقتی که متغیرهای مستقل ثابت هستند. کمترین کاربرد آن تمرکز روی چندک یا پارامتر مکانی توزیع شرطی متغیر وابسته از متغیر مستقل معین است. در همه موارد هدف تخمین یک تابع از متغیرهای مستقل است که تابع رگرسیون نامیده شده‌است.

رگرسیون چیست؟

- در تحلیل رگرسیون تعیین پراکندگی متغیر وابسته اطراف تابع رگرسیون مورد توجه است که می‌تواند توسط یک توزیع احتمال توضیح داده شود. تحلیل رگرسیون به صورت گسترده برای پیش‌بینی استفاده شده‌است. تحلیل رگرسیون همچنین برای شناخت ارتباط میان متغیر مستقل و وابسته و شکل این روابط استفاده شده‌است. در شرایط خاصی این تحلیل برای استنتاج روابط عالی بین متغیرهای مستقل و وابسته می‌تواند استفاده شود. هرچند این می‌تواند موجب روابط اشتباه یا باطل شود بنابراین احتیاط قابل توصیه است.

رگرسیون چیست؟

- تکنیک‌های زیادی برای انجام تحلیل رگرسیون توسعه داده شده‌است. روش‌های آشنا همچون رگرسیون خطی و حداقل مربعات که پارامتری هستند، در واقع در آن تابع رگرسیون تحت یک تعداد محدودی از پارامترهای ناشناخته از داده‌ها تخمین زده شده‌است. رگرسیون غیر پارامتری به روش‌هایی اشاره می‌کند که به توابع رگرسیون اجازه می‌دهد تا در یک مجموعه مشخص از توابع با احتمال پارامترهای نامحدود قرار گیرند.

- تحلیل رگرسیونی یا تحلیل وایازشی فن و تکنیکی آماری برای بررسی و مدل‌سازی ارتباط بین متغیرها است. رگرسیون تقریباً در هر زمینه‌ای از جمله مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی، بیولوژی و علوم اجتماعی برای برآورد و پیش‌بینی مورد نیاز است.

مثال

- رابطه بین قد و وزن انسانها را در نظر بگیرید. همه می دانیم که این رابطه یک رابطه مستقیم ریاضی و صد درصدی نیست که لزوما هر که قد بلندتری داشته باشد وزن بیشتری داشته باشد، اما می توان گفت که با احتمال قابل قبولی افراد با قد بلندتر، وزن بیشتری نیز دارند. در اینجا پیش بینی وزن از روی قد و بیان ارتباط بین این متغیر با روش آماری رگرسیون خطی صورت می پذیرد که این رابطه را به صورت کمی به ما نشان می دهد.

رگرسیون را با معادله رگرسیون بیان می کنند. در مثال فوق معادله رگرسیون خطی می تواند به صورت زیر باشد:

متغیر وزن = متغیر قد * a + b

ترسیم این خط پس از محاسبه ضرایب a و b ما را به خط رگرسیون می رساند.

مثال

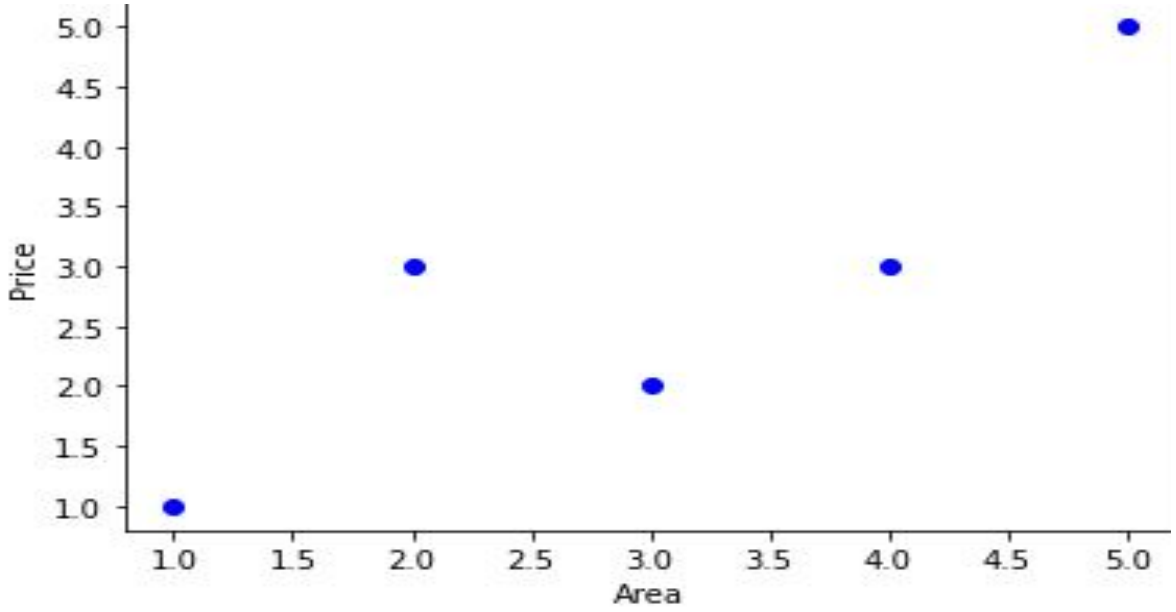
- فرض می کنیم که جدول زیر نشان دهنده متراژ و قیمت چند خانه است.

Area (x)	Price (y)
1	1
2	3
4	3
3	2
5	5

ستون Area در این جدول نشان دهنده متراژ و ستون Price نشان دهنده قیمت آن خانه است. برای مثال قیمت خانه ۱ متری ۱ تومان، خانه ۲ متری ۳ تومان است.

ادامه...

- اگر مقادیر جدول بالا را بر روی نمودار رسم کنیم، متوجه وجود یک رابطه خطی بین متراژ و قیمت می شویم، به طوری که در نتیجه افزایش متراژ، قیمت نیز افزایش می یابد.



ادامه...

- در چنین مواردی که یک رابطه خطی بین متغیرها وجود دارد، می توانیم برای پیش بینی مقادیر جدید، از رگرسیون خطی استفاده کنیم.

انواع رگرسیون...

- انواع بسیار زیادی از مدل های رگرسیونی وجود دارند. ما در اینجا به ۱۴ نوع از آنها اشاره میکنیم اما در بخش پیاده سازی به مدل های پرکاربرد بسنده میکنیم.

تکنیک های رگرسیونی به شرح زیرند :

۱. رگرسیون خطی (Linear Regression)
۲. رگرسیون لجستیک (Logistic Regression)
۳. رگرسیون چندکی (Quantile Regression)
۴. رگرسیون ستیغی (Ridge Regression)
۵. رگرسیون لاسو (Lasso Regression)
۶. رگرسیون شبکه الاستیک (Elastic Net Regression)

انواع رگرسیون - ادامه...

۷.رگرسیون مولفه های اصلی (Principle Component Regression)
۸.رگرسیون کمترین مربعات جزئی (Partial Least Square Regression)
۹.رگرسیون بردار پشتیبان (Support Vector Regression)
۱۰.رگرسیون ترتیبی (Ordinal Regression)
۱۱.رگرسیون پواسون (Poisson Regression)
۱۲.رگرسیون دو جمله ای منفی (Negative Binomial Regression)
۱۳.رگرسیون شبه پواسون (Quasi Poisson Regression)
۱۴.رگرسیون کاکس (Cox Regression)

انواع رگرسیون

- آگاهی از این روش‌ها به یک دانشمند داده کمک می‌کند که بهترین روش و الگو را برای تحلیل داده‌های خود به کار ببرد و در نتیجه مدل‌های ساخته شده از بیشترین کارایی و دقت برخوردار شوند.
- هر یک از روش های رگرسیونی، پیش فرض های مخصوص خود را دارد که برحسب ویژگی و مشخصات متغیرهای توصیفی (Explanatory Variables) و متغیر پاسخ (Response Variable) تعیین میشوند. توجه داشته باشید که گاهی به متغیر های توصیفی، متغیر های مستقل و به متغیر های پاسخ متغیر وابسته میگویند.

انواع رگرسیون – رگرسیون خطی

- در رگرسیون خطی (Linear Regression)، به طور کلی دو نوع متغیر وجود دارد:
- متغیرهای مستقل، که اغلب با x نشان داده می شوند.
- متغیرهای وابسته، که در نتیجه متغیرهای مستقل محاسبه می شوند و به طور معمول با y نشان داده می شوند.
- زمانی که فقط یک متغیر مستقل وجود داشته باشد، مدل رگرسیونی خطی را ساده (Simple Regression) می‌نامند و اگر بیش از یک متغیر مستقل (توصیفی) وجود داشته باشد، رگرسیون را چندگانه (Multiple Regression) می‌گویند.

فرمول کلی رگرسیون خطی ساده

- $y = \beta_0 + \beta_1 X + \varepsilon$
- با استفاده از این فرمول ما می توانیم به ازای هر ورودی x ، خروجی y را محاسبه کنیم.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

فرمول کلی رگرسیون خطی ساده

- مشخص است که ابتدا باید B۱ را حساب کنیم و سپس B۰ را به دست بیاوریم.

- برای محاسبه صورت کسر فرمول B۱ ، برای هر ردیف، مقدار ستون x را از میانگین xها کم می کنیم، مقدار ستون y را از میانگین yها کم می کنیم، حاصل را در هم ضرب می کنیم.

- برای تمامی ردیف ها این فرآیند را انجام می دهیم و در پایان تمامی این مقادیر را با هم جمع می کنیم.

انواع رگرسیون – رگرسیون لجستیک

- در رگرسیون لجستیک (Logistic Regression) ، متغیر وابسته، به صورت دو دویی است. به این معنی که مقادیر آن به دو طبقه صفر و یک دسته‌بندی شده‌اند. البته زمانی که از رگرسیون چند جمله‌ای لجستیک (Multinomial Logistic Regression) استفاده می‌کنید، ممکن است تعداد سطوح متغیر طبقه‌ای بیشتر از دو باشد. در این حال مدل رگرسیون لجستیک به شکل زیر نوشته می‌شود.

$$P(Y=1) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}$$

انواع رگرسیون – رگرسیون لجستیک

- واضح است که در این مدل رگرسیونی، خطاها، دارای توزیع نرمال نیستند و متغیر وابسته دارای توزیع دو یا چند جمله‌ای است در نتیجه نمی‌توان از مدل رگرسیون ساده یا خطی استفاده کرد.
- معمولا از این شیوه یا مدل رگرسیونی، برای طبقه‌بندی کردن مشاهدات جدید برحسب مقادیر قبلی استفاده می‌کنند و به نوع «یادگیری نظارت شده» محسوب می‌شود. به این ترتیب اگر مقدار $P(Y = 1)$ یک مقدار آستانه (مثلا ۰/۵) بیشتر باشد، آن مشاهده را در گروه ۱ طبقه بندی میکنیم.

انواع رگرسیون – رگرسیون شبکه الاستیک

- «رگرسیون شبکه الاستیک» (Elastic Net Regression)، با ترکیب رگرسیون لاسو و رگرسیون ستیغی، بر معایب آن‌ها غلبه کرده و جایگزین مطمئن برای آن‌ها است. به این ترتیب اگر با مدلی مواجه هستید که متغیرهای توصیفی آن با یکدیگر همبستگی دارند، بهتر است از رگرسیون شبکه الاستیک استفاده کنید.
- به این ترتیب یک قاعده‌سازی مرتبه ۱ و ۲ روی مدل همزمان اعمال می‌شود. در نتیجه تابع هدف در رگرسیون شبکه الاستیک به صورت زیر نوشته خواهد شد.

$$\min\left(\sum \epsilon^2 + \lambda_1 \sum \beta_i + \lambda_2 \sum |\beta_i|\right)$$

انواع رگرسیون – رگرسیون کمترین مربعات جزئی

- زمانی که بین متغیرهای توصیفی، وابستگی شدید وجود داشته باشد، به جای رگرسیون مولفه‌های اصلی بهتر است از رگرسیون کمترین مربعات جزئی Partial Least Square Regression استفاده شود. همچنین زمانی که تعداد متغیرهای توصیفی زیاد هستند و می‌خواهیم موثرترین متغیرها در مدل حضور داشته باشند، از رگرسیون کمترین مربعات جزئی (PLS) استفاده می‌کنیم. در ادامه رگرسیون مولفه های اصلی را نیز شرح میدهیم.
- هم در روش رگرسیون مولفه‌های اصلی و هم کمترین مربعات جزئی، متغیر جدیدی به عنوان متغیر پیش‌گو ساخته می‌شود که به آن مولفه گفته می‌شود. این متغیر جدید، ترکیب خطی از متغیرهای توصیفی است. ولی تفاوت در این است که در تحلیل رگرسیون PCR، مولفه‌ها براساس توصیف واریانس کل متغیرهای توصیفی تولید می‌شوند بدون آنکه به مقایر متغیر پاسخ توجه شود. در حالیکه در PLS با در نظر گرفتن متغیر پاسخ و متغیرهای پیشگو، مولفه‌ها تولید می‌شوند و در نهایت مدلی ایجاد می‌شود که با کمترین عوامل، بهترین برازش را دارد.

انواع رگرسیون – رگرسیون مولفه های اصلی

- زمانی که هم‌خطی یا هم‌خطی چندگانه در مدل رگرسیونی وجود داشته باشد، بهتر است از مدل رگرسیون مولفه‌های اصلی (Principle Component Regression) که به اختصار با PCR نشان داده می‌شود، استفاده کنیم.

- رگرسیون مولفه‌های اصلی در دو گام اجرا می‌شود.

۱. استخراج مولفه‌های اصلی براساس متغیرهای توصیفی

۲. اجرای رگرسیون براساس مولفه‌های ایجاد شده به عنوان متغیرهای مستقل با متغیر پاسخ

انواع رگرسیون – رگرسیون مولفه های اصلی

- به این ترتیب، مشکل هم‌خطی یا هم‌خطی چندگانه از مدل رگرسیونی خارج شده و از طرفی با توجه به استفاده از مولفه‌های کمتر از تعداد متغیرهای توصیفی، ابعاد یا تعداد متغیرهای به کار رفته در مدل رگرسیونی نیز کاهش می‌یابد.
- بخش اول در محاسبات مربوط به رگرسیون مولفه های اصلی، تعیین "بار های عاملی" است که به کمک آن مولفه ها ایجاد میشوند. هر مولفه مثل U_i به صورت زیر تشکیل میشود.

$$U_i = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

انواع رگرسیون – رگرسیون مولفه های اصلی

- البته دقت داشته باشید به شرطی که بارها B_i در شرط زیر صدق کنند.

$$\sum_{i=1}^{p} \beta_i^2 = 1$$

اولین مولفه اصلی دارای بیشترین سهم از واریانس متغیر پاسخ را در خود جای داده است. به همین ترتیب، مولفه‌های بعدی، سهم کمتری در بیان واریانس کل متغیر پاسخ خواهند داشت.

موضوع دیگری که در مورد مولفه‌های اصلی وجود دارد، ناهمبسته بودن آن‌ها است. به این معنی که ضریب همبستگی بین مولفه‌ها تقریبا صفر است. در نتیجه مشکل هم‌خطی یا هم‌خطی چندگانه در مدل ایجاد شده، از بین خواهد رفت.

انواع رگرسیون – رگرسیون مولفه های اصلی

- همچنین از آنجایی که مقدار p را می‌توان کمتر یا مساوی با k انتخاب کرد، کاهش بعد مسئله نیز از مزایای استفاده از PCR محسوب می‌شود. در نتیجه می‌توان به جای استفاده از مدل با ۱۰ متغیر توصیفی، فقط با ۲ یا ۳ مولفه، مدل رگرسیونی را ایجاد کرد بطوریکه کمترین میزان اطلاعات در مورد متغیر وابسته، در مدل از بین رفته یا نادیده گرفته شده باشد.
- باید این موضوع را در نظر بگیرم که استفاده از PCR، روشی برای تعیین ویژگی‌های موثر در مدل رگرسیونی نیست بلکه با بهره‌گیری از آن، مولفه‌های جدیدی ایجاد می‌شود که بیشترین توصیف یا سهم تغییرات برای متغیر وابسته را در خود دارند. در نتیجه نمی‌توان گفت که کدام متغیر توصیفی، بیشترین نقش را در مدل رگرسیونی PCR دارد.

منابع

- Linear regression – Wikipedia
- Regression Definition (investopedia.com)
- Regression Analysis - Formulas, Explanation, Examples and Definitions (corporatefinanceinstitute.com)
- Introduction to Linear Regression Analysis (Wiley Series in Probability and Statistics): Montgomery, Douglas C., Peck, Elizabeth A., Vining, G. Geoffrey: 9781119578727: Amazon.com: Books