

آنالیز داده با حذف داده های خارج از محدوده

درس کیهان شناسی

زمستان ۱۴۰۱

هنگامی که با تعداد داده های زیادی سر و کار دارید، ممکن است با داده هایی مواجه شوید که با مقادیر مورد انتظار شما تفاوت بسیاری دارند بدین صورت که بسیار بزرگ تر یا بسیار کوچک تر هستند. اینگونه داده ها که آن ها را تحت عنوان داده های خارج از محدوده می شناسیم (outliers)، می توانند نتیجه مسائلی مانند خطای انسانی، تجهیزات ناکافی و یا در مواردی نمونه برداری ضعیف باشند. صرف نظر از نحوه ورود آنها به داده ها، می توانند تأثیر زیادی بر تجزیه و تحلیل آماری نتایج داشته باشند. برای حل این مشکل روش های گوناگونی پیشنهاد شده است اما پیش از هر چیز باید یاد بگیریم که چگونه و با چه معیاری می توان این داده ها را تشخیص داد. به طور کلی راه هایی را که برای تشخیص این داده ها وجود دارند می توان به سه دسته تقسیم کرد:

۱. انجام فرآیندهای visualization بر روی داده ها:

برای این کار می توانید از برنامه های مختلف مانند پایتون، ممتیکا و ... کمک بگیرید. نمودارهایی که در این مورد بیشترین کاربرد را دارند نمودارهای هیستوگرام، پراکندگی، و یا نمودارهای جعبه ای هستند. با توجه به میزان تراکم داده ها، می توان تصمیم گرفت که داده ها در چه بازه ای معتبر هستند. این روش به دقت روش های آماری نیست اما در موارد ابتدایی و در جها دست یافتن به یک دید کلی نسبت به داده ها، می تواند مفید باشد.

۲. استفاده از IQR (Interquartile range):

این روش معیار است برای این که پراکندگی آماری داده ها را با استفاده از آن بررسی کنیم. در این روش، داده ها به سه دسته تقسیم می شوند که دسته اول شامل $1/4$ از داده هاست که کم ترین مقدار را دارند (Q_1)، دسته سوم شامل $1/4$ از داده هاست که بیشترین مقدار را دارند (Q_3)، و دسته دوم شامل داده هاییست که بین ربع اول و ربع سوم واقع شده اند و بازه مطلوب ما را می سازد که همان IQR می باشد. در واقع این روش بر اساس پیدا کردن میانه اعداد در بازه های مختلف می باشد. بنابراین می توان گفت:

$$IQR = Q_3 - Q_1$$

این روش بیشتر برای توزیع های نرمال به کار می روند. اگر به عنوان مثال توزیعی $log - normal$ داشتیم، چگونه باید این روش را اعمال کنیم؟

۳. استفاده از میانگین و انحراف از معیار:

همانطور که میدانید، انحراف از معیار کمیتی است که پراکندگی داده ها حول مقدار میانگین را توصیف می کند و به ما می گوید که مقدار استاندارد این انحراف تا چه حدی می تواند باشد. بنابراین انتظار خواهیم داشت که داده های مطلوب برای ما داده هایی باشند که از این حد تجاوز نکنند. اگر میانگین را با نماد avg و انحراف از معیار را با std نمایش دهیم، بازه مطلوب برای داده های x_i بدین صورت خواهد بود:

$$x_i \in (avg - std, avg + std)$$

۴. Z-score:

در این مدل، تعداد نمونه هایی بررسی می شود که انحراف از معیار این نمونه ها، باعث شده است داده ها از مقدار میانگینشان فاصله بگیرند. نکته ای که در این مدل وجود دارد این است که برای توزیع های نرمال به خوبی عمل می کند و معمولاً در مواردی که تعداد داده ها کم هستند کاربرد بیشتری دارد. به شکل زیر تعریف می شود:

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

که در رابطه فوق \bar{x} نشان دهنده میانگین داده ها و σ انحراف از معیار می باشد. طبق تجربه، معمولاً داده هایی با $|z_i| \leq 3$ داده پرت محسوب می شوند. اما باید توجه داشت که این معیار از شرایطی به شرایط دیگر ممکن است تغییر کند.

۵. روش باقی مانده ($residuals$):

در مدل رگرسیون خطی، باقی مانده های معمولی ($ordinary residuals$) به صورت تفاوت میان داده های مشاهده شده و داده های مورد انتظار تعریف می شوند. داده هایی با داشتن باقی مانده معمولی بزرگ، معمولاً داده پرت محسوب می شوند چراکه در مدل رگرسیون خطی قرار نمی گیرند و با خط فیت شده فاصله زیادی خواهند داشت. نکته ای که وجود دارد این است که آیا ما مطمئن هستیم که واحد اندازه گیری تمام داده ها یکی است؟ به عنوان مثال اگر در داده مربوط به فواصل کهکشان ها مدل فیت شده بر اساس فاصله کهکشان ها برحسب کیلومتر باشد، داده ای که با مگا پارسک بیان شده باشد تفاوت زیادی با خط فیت شده خواهد داشت (در صورتی که داده پرت نیست). برای رفع این مشکل، می توان از روش $Studentized residuals$ استفاده کرد. در این روش، با



تقسیم باقی مانده ها بر انحراف از معیار داده، واحد اندازه گیری را از بین می برند و کمیت را بی بعد می کنند. آیا محدودیت دیگری برای این مدل به ذهنتان می رسد؟

حال که با روش های فوق داده های خارج از محدوده را مشخص کردیم، می توانیم تصمیم بگیریم که با این داده ها از چه طریقی برخورد کنیم. به طور کلی دو روش برای این کار وجود دارد:

۱. یکی از روش های متداول این است که داده های پرت را حذف کنید. استفاده از این روش اگرچه روش آسانی است اما باید توجه داشت که در مواردی ممکن است دقت کار را کاهش دهد. این مورد زمانی که تعداد زیادی از داده ها خارج از محدوده محسوب می شوند مناسب نیست زیرا باعث می شود که بخش زیاد از داده ی در دسترس حذف شود. اما در موارد دیگر می تواند گزینه خوبی باشد. بنابراین استفاده از این روش، وابسته به مراحل پیشین در تشخیص داده های پرت می باشد.

۲. روش دیگری که می توان از آن بهره برد این است که حداقل یا حداکثر مقادیر مجاز را به داده های پرت اختصاص دهیم. با ذکر مثالی این موضوع را روشن تر خواهیم کرد: اگر از روش IQR برای تشخیص داده های خارج از محدوده استفاده کرده باشیم و قصد حذف داده های خارج از محدوده IQR را نداشته باشیم، می توان بیشترین مقدار مجاز واقع در این ناحیه را پیدا کرد و به تمام اعداد واقع در ناحیه Q_3 این مقدار را نسبت داد. از طرفی دیگر می توان کم ترین مقدار مجاز در این ناحیه را نیز پیدا کرد و این مقدار را به تمام اعداد واقع در ناحیه Q_1 نسبت داد. به این ترتیب داده ای را حذف نکرده ایم و تنها تمام داده های موجود را به بازه مجاز منتقل کرده ایم.

نکته ۱: باید به این موضوع دقت شود که حذف کردن یا نگه داشتن داده های پرت، ملزم دقت به موارد زیادی است. به نظر شما در هر یک از موارد زیر بهتر است اینگونه داده ها نگه داشته شوند یا کنار گذاشته شوند؟

- داده ای مشخصا به خاطر اندازه گیری اشتباه وارد داده ها شده باشد.
- داده پرت تنها بر روی  اثر بگذارد و روی  های اولیه ما اثری نداشته باشد.
- این داده هم بر روی نتایج و هم بر فرض های اولیه اثر داشته باشد.
- به خوبی از محدوده داده های مورد نظرمون واقع باشیم.
- تعداد داده هایمان خیلی زیاد باشد.
- تعداد داده های خارج از محدوده زیاد باشد.

■ نتایجی که داده ها بیان میکنند در حوزه های مهمی باشد.

نکته ۲: به تفاوت های میان نویز و داده پرت فکر کنید. کدام یک برای ما سودمند است و کدام یک نیست؟ سعی کنید که این موضوع را در موارد فیزیکی بررسی کنید.