CA3

رضا چهرقانی ۸۱۰۱۰۱۴۰۱

---

1. اتوانکودر

1.1. یکی از کاربردهای اتوانکودر کاهش نویز در تصاویر می‌باشد. از آنجایی که داده‌ها اغلب دارای نویز هستند ما از نوعی از اتوانکودرها برای کاهش نویز استفاده می‌کنیم. اتوانکودر برای حذف نویزها ابتدا تصویر ورودی را encode می‌کند و هنگامی که می‌خواهد آن را decode کند سعی می‌کند تصویر را به گونه‌ای بازسازی کند که نویز نداشته باشد.

یکی دیگر از کاربردهای اتوانکودرها کاهش ابعاد داده‌ی ورودی است. خیلی از زمان ها برای سادگی کار بعد داده‌های ورودی را کاهش می‌دهیم و فقط یک سری از ویژگی‌های را که می‌خواهیم نگه می‌داریم. همان طور که در صورت پروژه گفته شده encoder داده‌های ورودی را به فضای نهان تبدیل می‌کند که این فضا دارای بعد کمتر می‌باشد و ویژگی‌های مهم و معنادار از آن استخراج شده است. برای این کار پس از آموزش اتوانکودر قسمت decoder را حذف می‌کنید و خروجی ما در واقع همان خروجی encoder می‌شود.

1.2. اتوانکودر سعی در کاهش تفاوت بین تصویر ورودی و تصویر بازسازی شده می‌کند. اما به دلیل آنکه برای ساخت تصویر بین پیکسل‌ها میانگین می‌گیرد که موجب تار شدن تصویر می‌شود. همچنین میزان تار بودن تصویر خیلی بستگی به اندازه‌ی فضای پنهان دارد. وقتی فضای پنهان کوچک باشد موجب از دست رفتن جزئیات و افزایش تاری تصویر بازسازی شده می‌شود. زیرا مدل باید تمام اطلاعات لازم را در فضای اندکی جای دهد. در فضای پنهان بزرگ پتانسیل ذخیره‌ی جزئیات بیشتر وجود دارد که موجب کاهش تاری می‌شود اما ممکن است به داده و محاسبات بیشتری نیازمند باشد.

1.3. در آزمون فرض به روش p-value اگر مقدار محاسبه شده بسیار کوچک باشد آزمون رد می‌شود. در اینجا هم این مقدار تقریبا برابر صفر می‌باشد. پس داده‌های MSE از توزیع نرمال پیروی نمی‌کنند.

2. معادله رگرسیون

2.1. نقطه پرت نقطه داده ای است که پاسخ y از روند کلی بقیه داده ها پیروی نمی کند و تفاوت قابل توجهی با سایر نقاط داده دارد. اگر یک نقطه داده دارای مقادیر x پیش‌بینی‌کننده افراطی باشد نقطه اهرمی است. به نقطه‌ای با هر دو ویژگی نقطه تاثیر گذار می‌گویند. این نقاط می توانند منجر به پیش بینی های نادرست شود. نقاط پرت می‌توانند باعث شوند که خط رگرسیون به سمت آنها منحرف شود. نقاط اهرمی می تواند باعث شود که خط رگرسیون به سمت آنها شیفت پیدا کند. نقاط تاثیرگذار می توانند تاثیر قابل توجهی بر روی خط رگرسیون داشته

باشند و باعث شوند که در جهتی حرکت کند که نمایشگر کل داده ها نیست. اگر این نقاط مشکل ساز تشخیص داده شوند حذف آن‌ها مهم است. البته ممکن است همه‌ی این نقاط مشکل ساز نباشند و دادهٔای معتبری باشند که باید در تحلیل استفاده شوند.

2.2. ضریب تعیین نسبتی از واریانس برحسب متغیر وابسته است که از متغیرهای مستقل قابل پیش‌بینی باشد (در بخش بعد از چنین فرمولی برای محاسبه‌ی آن استفاده می‌کنیم.). هنگامی که برای بررسی نیکوئی برارزش رگرسیون خطی استفاده می‌شود مقدار ضریب تعیین بین 0 و 1 قرار می‌گیرد. هنگامی که برابر صفر است یعنی مدل همان میانگین yها را پیش‌بینی می‌کند ولی وقتی یک می‌شود یعنی همه‌ی پیش‌بینی‌ها درست است. در رگرسیون خطی مقدار ضریب تعیین برابر با مجذور ضریب همبستگی بین داده‌های ورودی و مقادیر پیش‌بینی شده است. در کل ضریب تعیین ابزاری برای ارزیابی درستی پیش‌بینی یک مدل است.

2.3. در رگرسیون بر پایه هشت داده اصلی ضریب تعیین به عدد یک نزدیک است و این یعنی نشانه‌ی درستی بالای پیش‌بینی است. همانطور هم که از نمودار پیداست نقاط نزدیک خط رگرسیون می‌باشند. در رگرسیون دومی با اضافه کردن نقطه دور افتاده ضریب تعیین به شدت کاهش می‌یابد. دلیل آن از نمودار پیداست. زیرا نقطه‌ای که اضافه کرده‌ایم به شدت از خط رگرسیون دور است. در سومی هم به همین شکل است زیرا نقطه اهرمی از خط رگرسیون فاصله دارد. اما در چهارمی وقتی که نقطه‌ی دور افتاده-اهرمی را اضافه می‌کنیم این نقطه بسیار به خط رگرسیون نزدیک است و مدل پیش بینی کننده‌ی ما را منحرف نمی‌کند و ضریب تعیین در این حالت نزدیک به یک خواهد بود. پس در نتیجه دو نقطه‌ی دور افتاده و اهرمی مشکل‌ساز هستند و برای افزایش ضریب تعیین و در نتیجه افزایش دقت پیش بینی باید آن ها را حذف کرد اما نقطه‌ی دور افتاده-اهرمی مشکل‌ساز نیست و حتی می‌توان به عنوان داده‌ی معتبر نیز در نظر گرفت.

2.4. چند نوع مدل رگرسیونی داریم که هر کدام به نحوی تاثیر این نقاط دور افتاده یا اهرمی را کاهش می‌دهد. یکی از آن ها رگرسیون قوی است. که روش‌هایی برای محدود کردن تأثیر نقض مفروضات صحیح توسط فرآیند تولید داده‌های اصولی بر تخمین‌های رگرسیون دارد. برای مثال تابع Huber loss جایگزینی قوی برای روش Least Squares است که هنگام محاسبه سهم داده‌های پرت را کاهش می‌دهد و در نتیجه تاثیر آن ها را بر تخمین رگرسیونی کاهش می‌دهد. یکی دیگر از انواع رگرسیون، رگرسیون وزن‌دار است که به هر داده وزن مشخصی در محاسبات می‌دهد. در این نوع رگرسیون اگر به داده‌ها وزن مناسبی داده‌شود می‌توان خطا را به حداقل رساند و به حداکثر دقت رسید. یکی دیگر از کارهایی که می‌توان کرد تشخیص داده‌های مشکل‌ساز و حذف آن ها است. همچنین اگر متغیر وابسته خیلی منحرف شده بود با در نظر گرفتن این تغییر می‌توان تاثیر نقاط پرت را کاهش داد.

3. قضیه حد مرکزی و نمونه‌گیری

3.1. به نظرم می توانیم از داده‌های سایر بازیکنان میانگین بگیریم و حاصل را برای بازیکنانی که اطلاعات آنها را نداریم قرار دهیم.

3.2. مقدار min برابر با کوچکترین داده و max برابر با بزرگترین داده است. حال اگر اعداد را از کوچک به بزرگ مرتب کنید عدد وسط برابر با میانه Q2 می‌شود (اگر تعداد داده‌ها زوج بود میانه برابر میانگین دو عدد وسط می‌شود.). حال اگر اعداد بین min و میانه را در نظر بگیریم میانه‌ی این اعداد برابر است با Q1 و به طور مشابه میانه‌ی اعداد بین میانه و max برابر Q3 می‌باشد. البته تابع sns.boxplot که از آن در رسم نمودار جعبه‌ای استفاده کردیم بر اساس IQR تعدادی از داده‌ها را به عنوان داده‌ی پرت در نظر می‌گیرد.

3.3. ب. به ازای هر نقطه‌ی (x, y) در نمودار Q-Q مقادیر x برابر توزیع اول (معمولا توزیع نظری) و مقادیر y برابر توزیع دوم می‌شود. البته اندازه‌ی نمونه دو توزیع باید برابر باشد و باید آن ها مرتب شده باشند.

ج. همان طور که از نمودار پیداست نقاط نمودار Q-Q با تقریب خوبی روی خط y=x قرار می‌گیرند. به همین دلیل می‌توان نتیجه گرفت که توزیع وزن بازیکنان تقریبا دارای توزیع نرمال است.

د. همان طور که از مقادیر پیداست مقدار p_value تقریبا برابر نیم است که مقدار زیادی محسوب می‌شود پس آزمون فرض پذیرفته می‌شود. پس طبق این می‌توان نتیجه گرفت وزن بازیکنان از توزیع نرمال پیروی می‌کند.

ه. با توجه به نمودار Q-Q با افزایش n توزیع وزن بازیکنان انطباق بیشتری با توزیع نرمال پیدا می‌کند. یعنی توزیع بازیکنان به سمت توزیع نرمال حرکت می‌کند. پس می‌توان نتیجه گرفت توزیع وزن بازیکنان، نرمال است. همچنین می‌توان نتیجه گرفت وقتی هر چه اندازه‌ی نمونه بزرگتر باشد نتیجه‌ی ما دقیق‌تر خواهد بود و می‌توانیم پارامترهای توزیع آن ها را به صورت دقیق‌تری بدست بیاوریم.

3.4. همان طور که از مقادیر p_value پیداست با افزایش n مقدار آن به سمت صفر میل می‌کند. زیرا با افزایش نمونه‌ها معلوم می‌شود که آن ها متعلق به توزیع نرمال نیستند (زیرا متعلق به توزیع پواسون هستند.). در n=5 مقدار p_value بیشتر از 0.05 است (مقدار significance level برابر 0.05 است.). پس نمی‌توانیم فرض صفر را رد کنیم. اما در n=50, 5000 مقدار p_value کمتر از 0.05 است پس می‌توانیم آن ها را رد کنیم و بگوییم توزیع آن ها نرمال نیست. قضیه حد مرکزی می‌گوید وقتی تعداد زیادی متغیر تصادفی با هم جمع می‌شوند توزیع متغیر تصادفی حاصل نرمال خواهد بود. اما ما در اینجا هیچ جمعی انجام نداده‌ایم.