به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

آمار و احتمال مهندسی

گزارش تمرین کامپیوتری سوم

نام و نام خانوادگی:

نیلوفر مرتضوی

شماره دانشجویی:

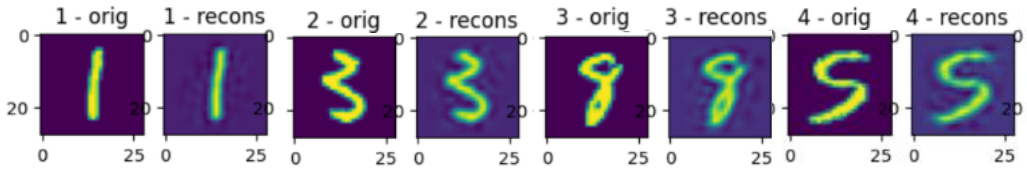
220701096

Question 1. Mean Squared Error

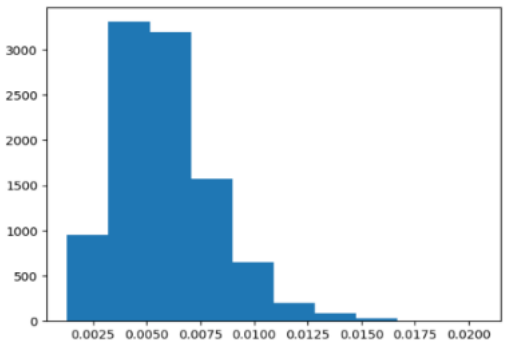
1- یکی از کاربردهای autoencoder ها کاهش ابعاد لایه های ورودی است. هم چنین می توان از آن در از بین بردن نویز تصاویر و بالا بردن وضوح نیز اسفاده کرد.

2- خطای اتوانکودر ها حاصل تفاوت بین داده های ورودی و داده بازسازی شده توسط آنها است. فضای نهان فضایی است که به داده های ورودی برای encode شدن اختصاص داده می شود. اگر این فضای نهان، به اندازه کافی بزرگ نباشد، درواقع اتوانکودر نمی تواند ورودی را با تمام ویژگی هایش به صورت کامل ذخیره کند. از طرفی دیگر اگر این فضای اختصاص داده شده بیش از حد نیاز بزرگ باشد، ممکن است encode کردن داده ها با خطا روبرو شود. در هر کدام از این حالات، بازسازی داده های ما به مشکل میخورد. پس اتوانکودر، میتواند تا حدودی ما را در جریان مناسب بودن یا نبودن ابعاد انتخاب شده برای فضای نهان قرار دهد.

3- ج) خروجی های تصادفی این قسمت در یک مرحله اجرای برنامه:



د) مطابق فرمول MSE برای این بخش، خروجی تابع به شکل زیر است:



ه) ابتدا MSE هر کدام از تصاویر را به دست می آوریم و سپس به صورت تصادفی، 1000 نمونه از مقادیر را انتخاب کرده و میانگین و واریانس این مقادیر را به تابع مربوط به آزمون می دهیم.

P-value مقدار کمتر از 0.05 دارد و بنابراین داده های MSE از توزیع نرمال مرتبط با میانگین واریانس حساب شده پیروی نمی کنند.

Question 2. Regression & Least Squares

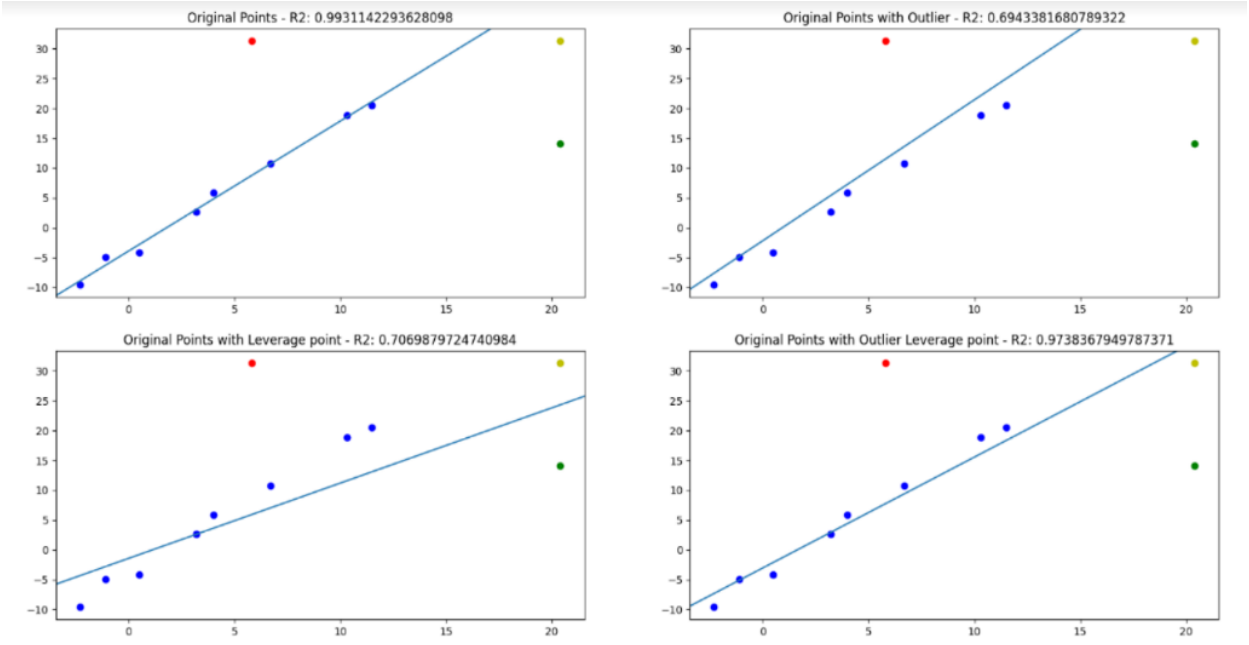
1- نقاط پرت دسته ای از داده ها هستند که با بقیهٔ داده ها تفاوت زیادی دارند و روند حرکت بقیه داده ها را دنبال نمی کنند. داده های پرت میتوانند تاثیر زیادی روی شیب خط حاصل از رگرسیون بگذارند و باعث انحراف در فرآیند آنالیز و تخمین داده ها شوند.

نقاط اهرمی یا همان leverage high، دسته ای از نقاط هستند که مقدارشان نسبت بقیه داده ها خیلی زیاد و یا خیلی کم باشد. نقاط اهرمی، همانند نقاط پرت، می توانند خط نتیجهٔ حاصل از رگریسون را دچار انحراف کنند و در نتیجه منجر به نتایج دارای خطا بشوند.

2- ضریب تعیین، معیاری است برای اینکه بفهمیم نتایج پیشبینی شده توسط یک مدل آماری، تا چه اندازه نسبت به تغییرات و گوناگونی پوشش داده شده در دادهٔ خروجی، واقعی و درست هستند. در بحث رگرسیون خطی، ضریب تعیین را می توان به سادگی برابر با مربع ضریب همبستگی بین مقادیر مشاهده شده و مقادیر پیشبینی کننده دانست که با r`2 نشان داده میشود.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

3- در نمودارهای خروجی زیر، نقاط آبی همان 8 نقطهٔ اصلی، نقطهٔ قرمز نقطهٔ پرت، نقطهٔ سبز نقطهٔ اهرمی و نقطهٔ زرد نقطهٔ با هر دو ویژگی است.



در نمودار اول، رگرسیون عملکرد خوبی داشته و ضریب تعیین ما هم بسیار به عدد 1 نزدیک است.

در نمودار دوم، نقطهٔ پرت را در داده های خود وارد کردیم که باعث انحراف نتیجه رگرسیون و در نتیجه کاهش ضریب تعیین شده است.

در نمودار سوم نقطهٔ اهرمی را در داده هاب خود وارد کردیم و در نتیجه، خط حاصل از رگرسیون به سمت نقطهٔ اهرمی منحرف شده است و در نتیجه،شاهد کاهش ضریب تعیین هستیم. مشاهده می شود که نقطهٔ پرت سبب انحراف بیشتری در نتیجه رگرسیون شده است.

در نمودار آخر، نقطهٔ با دو ویژگی پرت و اهرمی بودن را به داده های خود وارد کرده ایم و از آنجایی که این نقطه تا حدودی در راستای روند تغییرات داده های اصلی بوده، همانطور که مشاهده می شود، باعث انحراف خیلی کمی شده و ضریب تعیین نسبت به حالت اول که فقط دادههای اصلی حضور داشتند، دارای اختلاف خیلی خیلی کمی است و در این حالت هم ضریب تعیین به عدد یک بسیار نزدیک است.

4- می توانیم برای حذف داده های نامتعارف قبل از شروع عملیات این داده ها را پیش پرداز کنیم تا داده های مرتب تر و بدون انحراف داشته باشیم. راه دیگر استفاده از مدل هایی است که حساسیت کمی به این نوع داده ها دارند و یا میزان حساسیت شان از مدل مبتی بر کمترین مربع خطا کمتر است. معمول ترین مدل برای اینکه حساسیت کمتری نسبت به نقاط دورافتاده در داده داشته باشیم، M-estimation یا همان برآورد M است که معمول ترین جایگزین برای مدل مبتنی بر کمترین مربع خطا به شمار می رود.

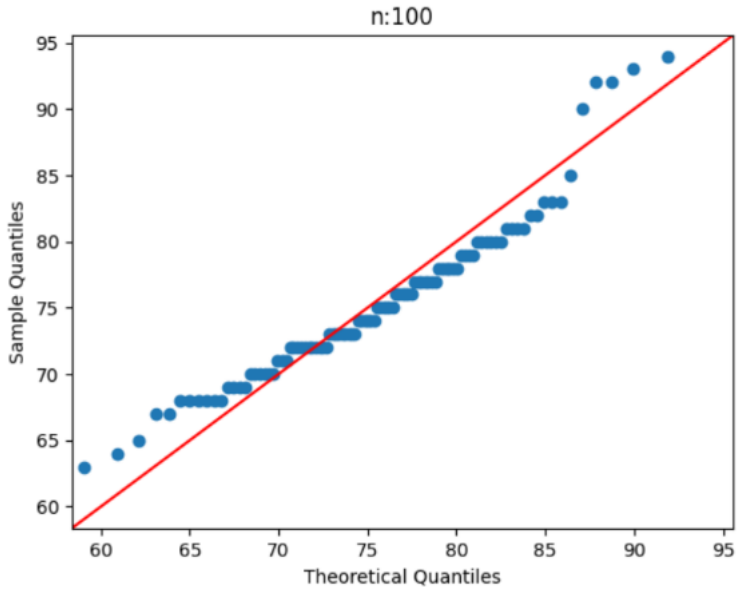
Question 3. Central Limit Theorem & Sampling

1- می توان مقادیر Icons و A/N را با میانگین مقادیر ستونشان جایگذاری کرد. برای همین برای ستون های pace و dribbling، ابتدا مقادیر A/N را با استفاده از تابع numeric_to به NaN تبدیل میکنیم و سپس از مقادیر آن ستون میانگین گرفته ایم. سپس مقادیر NaN را با مقدار میانگین جایگذاری کرده ایم. با توجه به خروجی کد، مقادیر به درستی با میانگین جایگذاری شده اند.

2- با توجه به این مقادیر، کمترین سن بازیکنان 17 و بیشترین88 سال است. همچنین چارک اول سن بازیکنان برابر با 23 سال و چارک دوم یا همان میانه سن بازیکنان برابر با 26 سال است. چارک سوم سن بازیکنان نیز برابر با 30 سال می باشد. این اعداد نشاندهندۀ این مهم است که تراکم سن بازیکنان در حدود رنج 23 الی 30 بوده است و دادۀ 88برای سن یک مقدار نامتعارف به شمار میرود. همچنین نیمی از سن بازیکنان کمتر از 26 و نیمی دیگر از بازیکنان سنی بیشتر از 26 سال دارند.

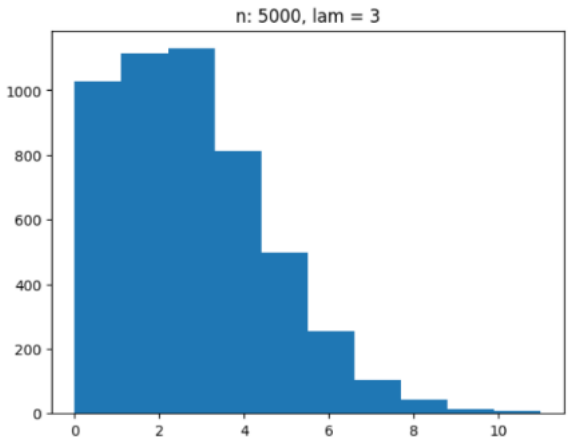
3- ب) استفادۀ این نمودار برای توزیع های داده ها است. اگر توزیع داده های مقایسه شده در این نمودار یکسان باشد، مقادیر این نمودار روی خط x = y قرار میگیرند و هر چه فرم نمودار به این خط نزدیکتر باشد، نشان می دهد که دو داده تطابق بیشتری دارند.

ج) نقاط نمودار تا حد قابل قبولی با خط x = y تطابق دارند. با توجه به قضیۀ حد مرکزی، انتظار داریم با افزایش مقدار n، این میزان انطباق با توزیع نرمال بیشتر شود.



ه) با افزایش مقدار n، تطابق نمودار با خط x = y بیشتر است و در نتیجه نمونه حاصل با توزیع نرمال تطابق بیشتری دارد.

4- الف)



ب) با مشاهده نمودار ها می توان فهمید که با افزایش مقدار n، مقدار value-p کاهش یافته است و نمودار ما از حالت خط x = y بیشتر فاصله گرفته است. این مورد ممکن است تا حدودی با انتظار ما از قضیه حد مرکزی تفاوت داشته باشد. باید توجه کنیم که که اعداد موجود در نمونه همگی دارای توزیع پواسون هستند و طبیعتاً با افزایش تعداد دادهای که از این توزیع برداشته میشود، بیشتر توزیع پواسون اولیه ترسیم می شود که متفاوت از توزیع نرمال خواهد بود.