گزارش پرو**ژه3**

برنا فروهری 810101480

سوال اول:

بخش1:

کاربرد هاي autoencoderها:

+ کاهش ابعاد ورودي: ما می توانیم از اتوانکودرها براي کاهش ابعاد لایه هاي داده ورودي داده شده استفاده کنیم.

++ از بین بردن نویز تصاویر: می توان با ترین کردن انکودرها، از آنها براي حذف یا کاهش نویز تصاویر استفاده کرد. با ترین کردن انکودرها براي مینیمم کردن خطاي بعد از بازسازي تصویر نسبت به خروجی ابتدایی، میتوان به این نتیجه رسید.

بخش 2:

خطاي اتوانکودر در واقع سنجشی براي عملکرد آن است و حاصل از وجود اختلاف میان داده ورودي و خروجی بازسازي شده توسط اتوانکودر است. فضاي نهان، فضایی با بعد کمتر است که داده ورودي ما به آنجا encode می شود. ابعاد این فضا، یکی از مهم ترین عوامل در عملکرد autoencoder است به طوري که ارتباط نزدیکی میان بعد فضاي نهان و میزان خطاي اتوانکودر وجود دارد. اگر ابعاد فضاي نهان، نسبت به داده ورودي خیلی کوچک انتخاب شود، اتوانکودر ممکن است نتواند تمامی ویژگی هاي مهم داده ورودي را در حین encode کردن پوشش دهد اما اگر ابعاد این فضا نسبتاً بزرگ انتخاب شود، تبدیل داده ورودي به ابعاد بزرگ تر ممکن است باعث وقوع خطا شود و در نتیجه داده بازسازي شده خروجی را دچار مشکل می کند. بنابراین، میان فضاي پنهان و ابعاد آن با میزان خطاي بازسازي اتوانکودر، ارتباط نزدیکی وجود دارد و همانطور که دیده شد، ابعاد فضاي نهان می تواند در مقدار خطاي اتوانکودر تاثیر بگذارد و همچنین میزان خطاي اتوانکودر، میتواند تا حدودي ما را از مناسب بودن یا نبودن ابعاد انتخاب شده براي فضاي نهان مطلع سازد. هدف در استفاده از اتوانکودر، یافتن تعادلی میان ابعاد و ویژگیهاي فضاي نهان و میزان دقت اتوانکودر در بازسازي است .

بخش 3 :

آ:

کد در zip قرار داده شده است( (Q1

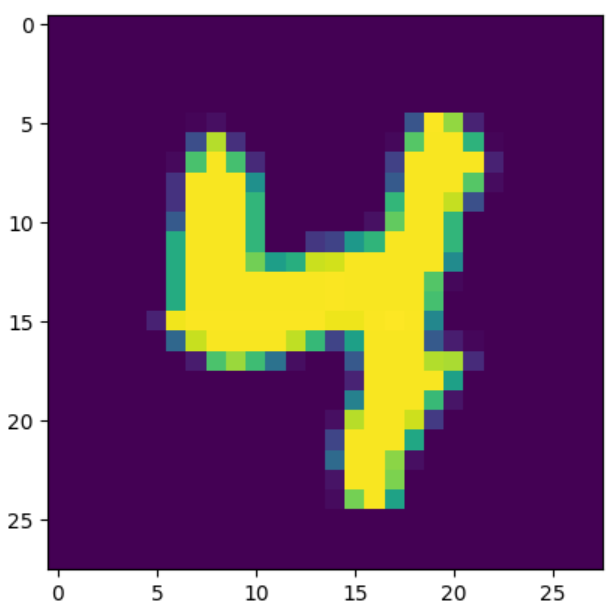
ب:

کد در zip قرار داده شده است( (Q1

ج:

با توجه به اینکه عکس ها با اندیس random و تصادفی انتخاب می شوند بنابراین پس از هر بار اجرا ممکن است خروجی ها متفاوت باشند.

تصاویر اصلی و بازسازی شده:

A green and blue image

Description automatically generated

A yellow and green number on a purple background

Description automatically generatedA green and yellow number on a blue background

Description automatically generated

A yellow and purple square with numbers

Description automatically generatedA green and blue square with numbers

Description automatically generated

A yellow and green circle

Description automatically generatedA green and blue circle with numbers

Description automatically generated

د:

با توجه به فرمول MSE آرایه اختلاف و مربع آن را تشکیل داده و بعد از آن جمع و تقسیم را محاسبه می کنیم.

نمودار hist خواسته شده:

A graph of a number of blue bars

Description automatically generated with medium confidence

ه:

ابتدا MSE هر کدام از عکسها را به دست می آوریم و سپس به صورت تصادفی، 1000 نمونه از این مقادیر را انتخاب می کنیم .

میانگین و واریانس را حساب کرده و به تابع مربوط به آزمون می دهیم.

value-p محاسبه شده کمتر از 0.05 می باشد و بنابراین داده هاي MSE هماهنگ با توزیع نرمال با میانگین و واریانس به دست آمده نیست .چون که نمونه تصادفی می باشد، ممکن است مقدار value-p در هر بار اجرا مقدار کمی تفاوت داشته باشد که در خروجی کد در Q1 قابل مشاهده می باشد.

سوال دوم:

بخش1:

داده هایی را که با داده های دیگر تفاوت زیادي دارند و روند حرکتی آنها با بقیه دادهها متفاوت است را نقاط پرت گویند.

علت های وجود این گونه داده ها در مجموعه داده :

+ به خاطر تنوع ناگهانی در خود داده ها

++خطاي تجربی در هنگام جمع آوري داده ها

در مجموعه داده سوال در مقدار 31 روند کلی دچار تغییری مشهود شده (در بین نقاط با مولفه x مشابه) بنابراین یک نقطه پرت محسوب می شود. داده هاي پرت ممکن است تاثیر زیادي روي شیب خط حاصل از رگرسیون بگذارند و بنابراین باعث انحراف در آنالیز و تخمین داده ها شوند. دسته اي از نقاط که مقدارشان نسبت به بقیه دادهها خیلی زیاد و یا خیلی کم باشد یا به عبارتی نقاطی که نسبت به داده هاي کنارشان دچار اختلاف زیاد شده اند را نقاط اهرمی گویند.

مثلا در میان داده هاي داده شده، مقدار مقدار مولفۀ x، در بین نقاط با مولفه y مشابه، یک تغییر ناگهانی داشته و درنتیجه یک نقطه اهرمی است.

نقاط اهرمی، مانند نقاط پرت، ممکن است خط نتیجۀ حاصل از رگریسون را دچار انحراف کنند و در نتیجه باعث انحراف در آنالیز و تخمین داده ها شوند.

بخش 2:

ضریب تعیین معیاري است که با آن متوجه اینکه نتایج تولید یا پیشبینی شده توسط یک مدل آماري، تا چه اندازه نسبت به تغییرات و گوناگونی پوشش داده شده در داده خروجی، واقعی و درست هستند.

به صورت دقیق تر، ضریب تعیین، برابر با میزان تغییراتی در متغیر وابسته است که توسط متغیرهاي مستقل قابل پیشبینی است.

در بحث رگرسیون خطی، ضریب تعیین برابر با مربع ضریب همبستگی بین مقادیر مشاهده شده و مقادیر پیشبینی کننده است که با 2rنشان داده می شود.

مقدار ضریب تعیین معمولا عددي بین 0 و 1 است.

ضریب تعیین = 0: مقادیر پیشبینی شده توسط مدل آماري ما برتري نسبت به زمانی که براي پیشبینی از میانگین متغیر استفاده کنیم، ندارد.

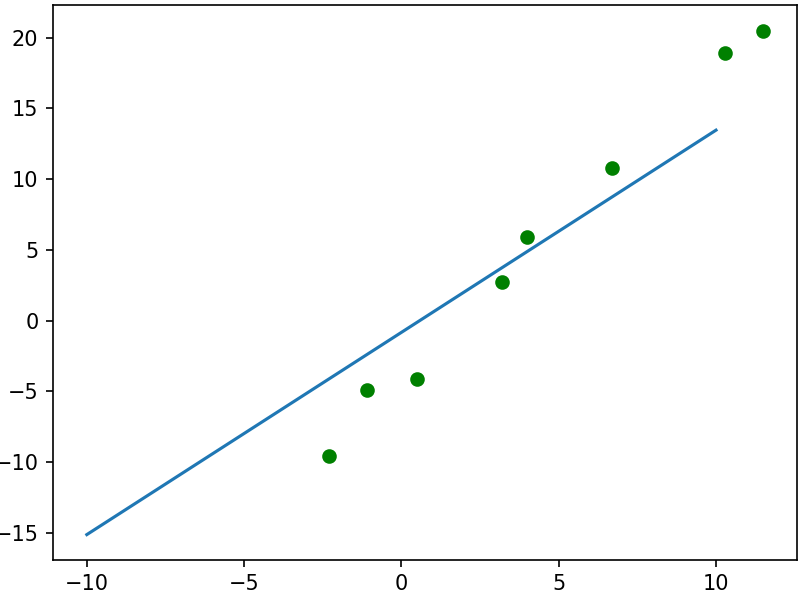
ضریب تعیین = 1: پیش بینی مدل ما، کامل و بدون نقص بوده است.

بنابراین اگر مقدار ضریب همبستگی به 1 نزدیک شود، پیش بینی ما بهتر بوده است.

در این سوال، انتظار داریم هر جا ضریب تعیین ما به 1 نزدیکتر باشد: + رگرسیون بهتر عمل کرده باشد

++ رگرسیون خطی تطلبق بیشتري با نقاط داشته باشد(با توجه به شکل خط حاصل از رگرسیون خطی و مختصات نقاط)

بخش 3:



A line graph with green dots

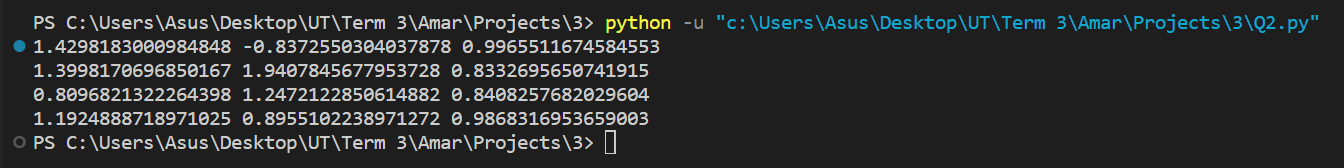
Description automatically generated

A line with green dots and a blue line

Description automatically generated

A line graph with green dots

Description automatically generated



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

نمودار دوم: نقطۀ پرت را شامل داده هاي خود کردیم و باعث انحراف نتیجه رگرسیون و کاهش ضریب تعیین شده است.

نمودار سوم: نقطۀ اهرمی را شامل داده هاي خود کردیم. خط حاصل از رگرسیون به سمت نقطه اهرمی منحرف و ضریب تعیین کاهش.

\*\*\*نقطه پرت سبب انحراف بیشتر و کاهش بیشار عملکرد رگرسیون نسبت به نقطه اهرمی \*\*\*

نمودار چهارم: نقطه با هردو ویژگی پرت و اهرمی بودن را شامل داده هاي خود کردیم و چون نقطه تا حدودي در روند داده های اصلی است، باعث انحراف خیلی کم و اختلاف کم با ضریب تعیین حالت اول (یه 1 میل می کند)

بخش 4:

حساسیت زیاد به داده هاي غیرمتعارف از خصوصیت رگرسیون خطی مبتنی بر کمترین مربع خطا است

راه حل 1:

ابتدا داده هاي نامتعارف را با پیش پردازش حذف کنیم و سپس رگرسیون را انجام دهیم.

راه حل 2:

استفاده از مدل هایی که حساسیت کمی به داده های نامتعارف دارند. معمول ترین مدل، برآورد M است .

برآورد M: تخمین برآورد درست نمایی بیشینه .

در این مدل برآورد، پیش بینی و به دست آوردن رابطه خطی با استفاده از بهینه سازي و مشتقگیري از یک تابع آماري است.

سوال سوم:

بخش 1:

بهترین راه جابگزینی مقادیر نامعلوم همان جایگذاري مقادیر با میانگین مقادیر ستونشان است.

بنابراین برای ستون های مربوطه، ابتدا مقادیر نامعلوم را به NaN (با کمک تابع to\_numeric ) تبدیل کرده و سپس از مقادیر آن ستون میانگین می گیریم. NaN در میانگین خود به خود حساب نمی شود)

سپس مقادیر NaN را با مقدار میانگین جایگذاري می کنیم. (با کمک تابع fillna )

بخش 2:

کمترین سن بازیکنان 17 سال

بیشترین سن بازیکنان 88 سال

چارك اول سن بازیکنان 23 سال

چارك دوم (میانه) سن بازیکنان 26 سال

چارك سـوم سن بازیکنان 30 سال

اعداد بالا و نمودار جعبه ای بیانگر این است که سن بازیکنان در بازه 23 الی 30 متراکم بوده است و داده 88 یک مقدار نامتعارف به شمار می رود. همینطور نصف سن بازیکنان بیشتر از 26 و نصف دیگر کمتر از 26 می باشند

A graph with a line graph and a line graph

Description automatically generated

بخش 3:

آ:

کد و مقادیربه دست آمده در Q3 قابل مشاهده می باشد و چون نمونه به صورت تصادفی با استفاده از تابع choice.random انتخاب شده، میانگین و واریانس و انحراف معیار هر بار ممکن است متفاوت باشند.

ب:

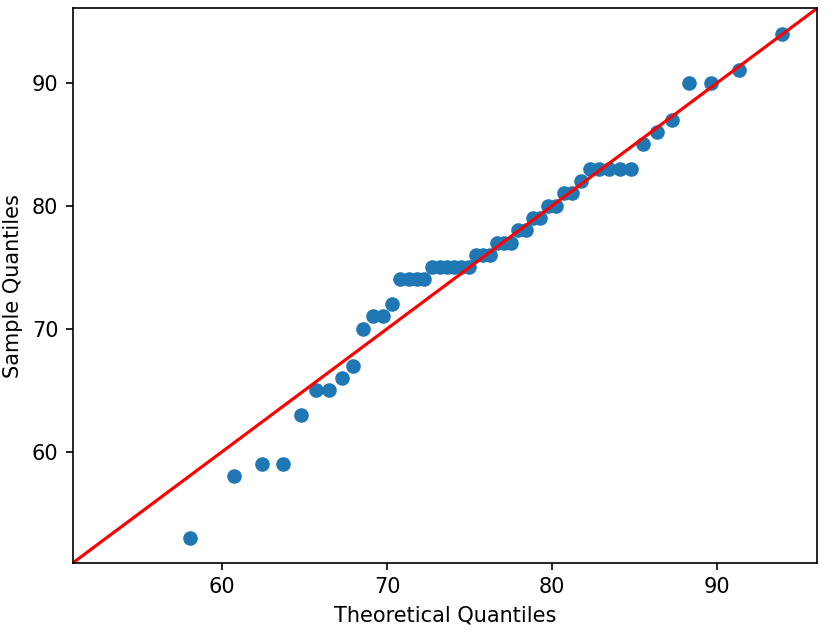
استفاده این نمودار براي بررسی یکسان بودن توزیع های دو نمونه که آیا توزیع ها یکسان هستند یا خیر.

روش کار:

این نمودار مقادیر چندك مجموعه اول را در برابر مقادیر چندک براي مجموعه دوم نشان می دهد.

اگر توزیع داده ها یکسان باشد مقادیر این نمودار روي خط y = x قرار می گیرند و هرچه نمودار به این خط نزدیکتر باشد، دو توزیع تطابق بیشتری دارند و هرچه نمودار از این خط فاصله بگیرد، توزیع دو مجموعه متفاوت تر می باشند

ج:



(براي سهولت در تحلیل، خط y = x هم رسم شده است.)

نقاط آبی نمودار تا حد خوبی با خط y = xتطابق دارند و نشان دهنده توزیعی نزدیک به یک توزیع نرمال می باشند.

با توجه به قضیه حد مرکزي، هر چه مقدار n افزایش یابد نقاط خیلی بیشتر مطابق y = x خواهند شد.

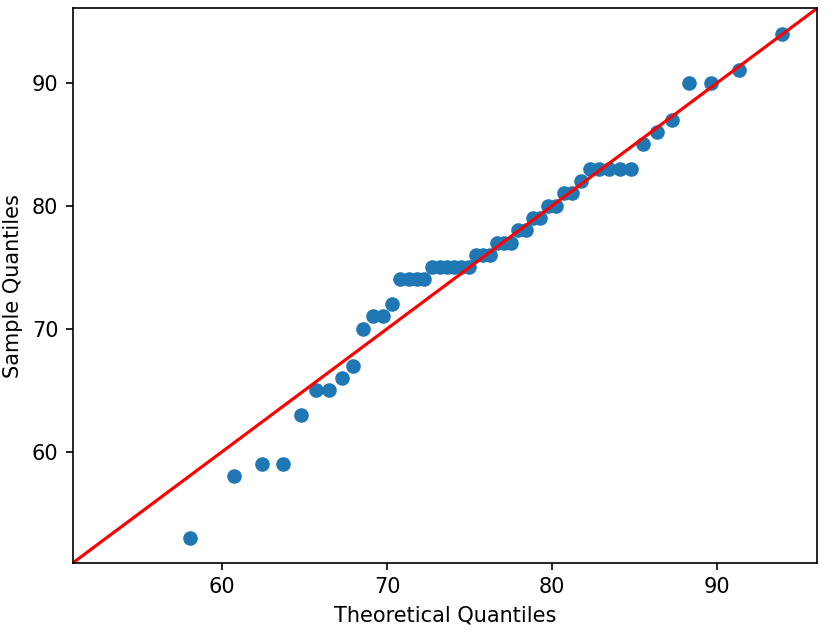
د:

به علت تصادفی بودن نمونه مقدار p\_value هربار متفاوت است. مقدار p\_value به دست آمده براي نمونه انتخاب شده در این اجرا از مقدار 0.05 کمتر بوده و در نتیجه فزض نرمال بودن توزیع این نمونه رد می شود.

ه:

با توجه به نمودار ها، با افزایش مقدار n، در نمودار Q-Q، نمودار به خط y = x میل می کند و بنابراین نمونه حاصل با توزیع نرمال تطابق بیشتري دارد.

این همانند قضیه حد مرکزی است و مشاهده کردیم که با افزایش n، نمونه ما به توزیع نرمال بیشتر نزدیک شد.



A graph with blue dots and red line

Description automatically generated

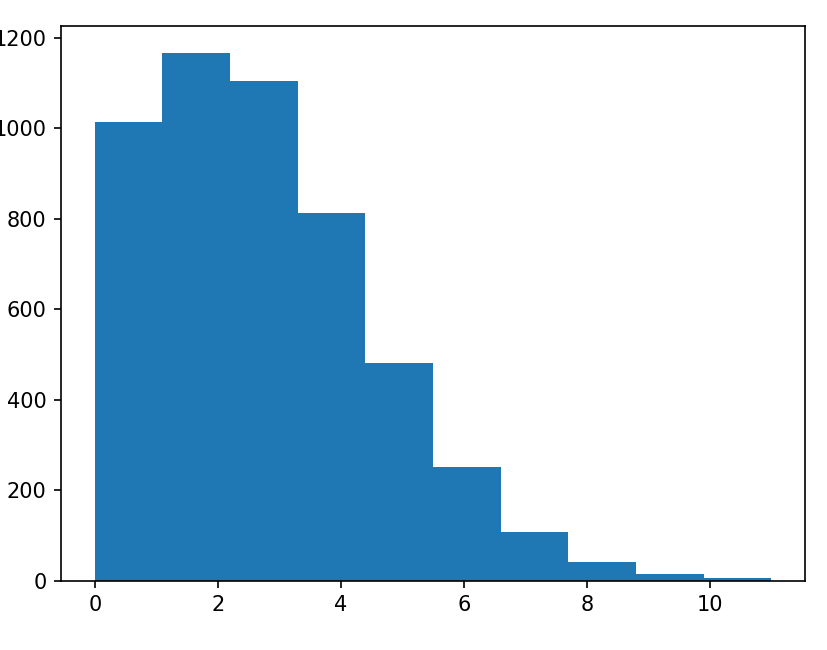
A red line with blue dots

Description automatically generated

بخش 4:

آ:

نمودار hist:



ب:

نمودارهای به دست آمده براي یک بار اجرا:

A red line with blue dots

Description automatically generated

A red and blue line with dots

Description automatically generated

A graph with blue dots and red line

Description automatically generated

با افزایش مقدار n، مقدارp\_value کاهش و نمودار از y = x فاصله گرفته به عبارت دیگر، با افزایش مقدار n، توزیع حاصل بیشتر به توزیع پواسون نزدیک می شود و از توزیع نرمال فاصله گرفته است که این از انتظار ما و قضیه حد مرکزی متفاوت است.

یکی از علت های آن می تواند خطاي آزمون شاپیرو ویلک باشد.

اعداد موجود در نمونه داراي توزیع پواسون می باشند و با افزایش تعداد دادهاي که از این توزیع برداشته میشود، بیشتر توزیع پواسون اولیه ترسیم می شود که متفاوت از توزیع نرمال خواهد بود و بنابراین از این لحاظ، نتایج نمودارها مطابق انتظار ما عمل کرده و تشکیل توزیع پواسون با افزایش تعداد داده کاملا مشخص است