بخش تئوری

سوال اول

به سوالات زیر پاسخ دهید.

الف) داده‌ی پرت[1] با نویز[2] را با یکدیگر مقایسه کنید.

نویزها به هیچ عنوان مطلوب ما نیستند و می توانند ما را گمراه کنند. اما outlier ها با اینکه به عنوان داده پرت شناسایی میشوند میتوانند برای ما مفید باشند و اطلاعات ارزشمندی را از آنها بدست آوریم و در برخی مسائل به دنبال آنها هستیم. البته در بعضی مواقع هم آن ها کم اهمیت هستند و ترجیح داده میشود که حذف شوند.

ب) یک سناریو بیان کنید که در آن داده‌های پرت برای ما مفید هستند و اطلاعات ارزشمندی از آن دریافت می‌کنیم.

برای مثال، یک نوسان شدید قیمت در بازارهای مالی. این نوسان شدید میتواند در دسته‌ی outlier قرار بگیرد ولی در عین حال نکات ارزشمندی را در خود گنجانده است.

ج) مشخص کنید که آیا یک نویز می‌تواند داده‌ی پرت باشد یا خیر؟

بله؛ noise میتواند داده ها را به صورت غیر طبیعی تغییر دهد و آنها را outlier کند؛ البته همچنین داده های noisy میتوانند مانند داده های طبیعی در محدوده عادی باشند، بنابراین noise ها نمیتوانند همیشه outlier باشند ولی این این امکان وجود دارد.

سوال دوم

در حوزه‌ی داده‌کاوی، انبار داده[3] چیست و چه تفاوت و شباهتی با پایگاه داده[4] دارد؟

انبار داده مجموعه بزرگی از داده ها از مکان های مختلف است که به گونه ای سازماندهی شده است که جستجو و تجزیه و تحلیل آن را آسان می کند. برای کمک به کسب و کارها در تصمیم گیری بر اساس داده ها استفاده می شود. داده ها اغلب اطلاعات قدیمی تری هستند که پاکسازی شده و برای تجزیه و تحلیل مفید هستند. انبارهای داده معمولاً در مناطق مختلف، مانند داده‌های فروش یا مشتری سازمان‌دهی می‌شوند و می‌توانند در فضای ابری ذخیره شوند.

انبار داده و پایگاه داده هر دو برای ذخیره و مدیریت داده ها استفاده می شوند، اما اهداف و ویژگی های متفاوتی دارند.

یک پایگاه داده معمولاً برای اهداف عملیاتی استفاده می شود و برای پشتیبانی از پردازش تراکنش های روزانه طراحی شده است. برای بازیابی و به روز رسانی سریع داده ها بهینه شده است و معمولاً با استفاده از مدل داده های رابطه ای

---

[1] outlier
[2] noise
[3] data warehouse
[4] database

سازماندهی می شود. پایگاه های داده اغلب برای پشتیبانی از برنامه هایی مانند مدیریت موجودی، پردازش سفارش و مدیریت ارتباط با مشتری استفاده می شود .از سوی دیگر، یک انبار داده برای مقاصد تحلیلی طراحی شده و برای پرس و جو و تجزیه و تحلیل کارآمد مقادیر زیادی از داده ها بهینه شده است. معمولاً با استفاده از مدل داده های بعدی سازماندهی می شود و برای پشتیبانی از هوش تجاری و فعالیت های تجزیه و تحلیل داده ها، مانند تجزیه و تحلیل روند، پیش بینی و تصمیم گیری استفاده می شود.

انبارهای داده معمولاً بزرگ‌تر و پیچیده‌تر از پایگاه‌های داده هستند، زیرا باید داده‌ها را از منابع متعدد ادغام کنند و آن‌ها را به قالبی تبدیل کنند که به راحتی قابل تجزیه و تحلیل باشد. آنها همچنین ممکن است از فناوری‌های مختلف پایگاه داده مانند پایگاه‌های داده ستونی یا پایگاه‌های داده NoSQL برای پشتیبانی از پرس و جو و تحلیل کارآمد استفاده کنند.

سوال سوم

یکی از روش‌های یافتن داده‌های پرت استفاده از توزیع نرمال[5] و percentile ها است. در مورد این روش تحقیق کرده و آن را توضیح دهید.

صدک یا percentile مقداری است در یک توزیع نرمال که درصد مشخصی از مشاهدات در زیر آن قرار دارد .در اصل روش توزیع نرمال و صدک یک روش آماری است که برای تشخیص داده های پرت در یک مجموعه داده استفاده می شود. در این روش، ابتدا توزیع نرمال برای داده های مجموعه تعریف می شود. توزیع نرمال یک توزیع پیوسته است که به شکل منحنی بلند و باریکی می باشد . سپس با استفاده از توزیع نرمال، با استفاده از صدک ها می توان داده های پرت را تشخیص داد. صدک یک مقدار است که n درصد از داده های مجموعه کوچکتر یا مساوی آن هستند. به عبارت دیگر، اگر n=95 صدک ۹۵ برابر با مقداری است که 95 درصد از داده های مجموعه کوچکتر یا مساوی آن هستند. برای تشخیص داده های پرت، داده هایی که خارج از محدوده صدک مشخصی هستند، به عنوان داده های پرت شناخته می شوند . به عنوان مثال، نقاط داده ای که از 99 percentile درصد فاصله دارند و کمتر از 1 percentile هستند، نقطه پرت در نظر گرفته می شوند .

در روش دیگر ابتدا داده ها را مرتب میکنیم سپسQ1 و Q2 و Q3 را در بین داده های مرتب شده پیدا میکنیم (که به ترتیب برابرند با داده ای که یک چهارم داده ها از آن کوچکترند - داده ای که نیمی از داده ها از آن کوچکترند - داده ای که سه چهارم داده ها از آن کوچکترند) و در مرحله بعد Q3-Q1=IRQ را محاسبه میکنیم .در مقدار Q3 + 1.5IRQ و Q1 – 1.5IRQرا به عنوان ابتدا و انتهای بازه محاسبه میکنیم .داده هایی که خارج از بازه مرحله قبل یعنی کوچکتر از Q1 – 1.5IRQ او بزرگتر از Q3 + 1.5IRQ باشند را به عنوان داده پرت معرفی میکنیم.

روش بعد می‌تواند بر روی داده‌هایی با توزیع نرمال اعمال شود. هر دادهای که از میانگین منهای 3 برابر انحراف معیار کوچکتر و یا از میانگین به‌علاوه 3 برابر انحراف معیار بزرگتر باشد داده پرت محسوب می‌شود و در این روش ۹۹.۸۷ درصد داده‌ها قابل قبول خواهند بود.

[5] Normal distribution

سوال چهارم

فرایند پاکسازی داده‌ها[6] و نمایش داده‌ها[7] را در نظر بگیرید:

الف) فرایند پاکسازی داده‌ها را تعریف کنید.

پاکسازی داده ها فرآیند شناسایی و تصحیح خطاها و ناهماهنگی ها در داده ها است. این شامل حذف یا تصحیح داده‌های ناقص، نادرست یا نامربوط از یک مجموعه داده و تبدیل آن به یک قالب سازگار و قابل استفاده است. هدف از پاکسازی داده ها بهبود دقت و کیفیت داده ها و قابل اعتمادتر کردن آن برای تجزیه و تحلیل و تصمیم گیری است. پاکسازی داده‌ها می‌تواند شامل تکنیک‌های مختلفی از جمله تکرار، استانداردسازی و عادی‌سازی باشد و ممکن است به روش‌های دستی یا خودکار نیاز داشته باشد.

ب) اهمیت نمایش داده‌ها را بیان کنید و به یک مورد از چالش‌های آن اشاره کنید.

تجسم داده ها فرآیند ارائه داده ها در قالب گرافیکی یا بصری است که به راحتی قابل درک و تفسیر است. این شامل استفاده از نمودارها، نمودارها، نقشه ها و سایر ابزارهای بصری برای نمایش داده ها و الگوهای موجود در داده ها است.

هدف تجسم داده ها انتقال داده های پیچیده به روشی است که به راحتی قابل درک باشد و شناسایی الگوها و بینش هایی که ممکن است در داده های خام به سختی دیده شوند. همچنین می تواند برای برجسته کردن روندها، مقایسه داده ها و شناسایی ناهنجاری ها یا نقاط پرت در یک مجموعه داده استفاده شود. تجسم داده ها ابزار مهمی برای تجزیه و تحلیل داده ها و ارتباطات است، زیرا کمک می کند تا داده ها را برای مخاطبان گسترده، از جمله کاربران غیر فنی، در دسترس و قابل درک تر کند.

ج) چرا پاکسازی داده‌ها یک فرایند مهم و پیشنیاز برای نمایش داده‌ها می‌باشد؟

پاکسازی داده‌ها قبل از تجسم داده‌ها مهم است، زیرا داده‌های نادرست، متناقض یا ناقص می‌توانند منجر به نتایج گمراه‌کننده شوند. با تمیز کردن داده ها از قبل، تحلیلگران می توانند اطمینان حاصل کنند که تجسم ها بر اساس داده های دقیق و قابل اعتماد است.

سوال پنجم

در یک آزمایشگاه ژنتیک مقدار فعالیت دو ژنوم مختلف مورد بررسی قرار گرفته و در ۱۰ بازه زمانی مختلف در به صورت زیر ثبت شده است:

Gen\time	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10

[6] data cleaning/cleansing
[7] data visualization

G1	-3	5	8	-2	1	2	3	-5	10	**-1**
G2	9	20	16	8	2	10	-6	-15	25	**-2**

الف) با استفاده از معیار شباهت Cosine Similarity، Correlation، Mutual Information شباهت این دو ژن را مقایسه کنید.

طبق روابطی که داشتیم شباهت کسینوسی برابر است با:

$$\frac{((-3 * 9) + (5 * 20) + (8 * 16) + (-2 * 8) + (1 * 2) + (2 * 10) + (3 * -6) + (-5 * -15) + (10 * 25) + (-1 * -2))}{\sqrt{(9 + 25 + 64 + 4 + 1 + 4 + 9 + 25 + 100 + 1)} * \sqrt{(81 + 400 + 256 + 64 + 4 + 100 + 36 + 225 + 625 + 4)}}$$
$$= ۵۱۶/۶۵۹.۰۸ = ۰.۷۸۲۹$$

معیار همبستگی برابر است با:

میانگین G1 : ۱.۸ = ۱۰ / ۱۸

میانگین G2 : ۶.۷ = ۱۰ / ۶۷

G1 – avg($G1$) = [

G1 X G2 = [-11.04, 42.56, 57.66, -4.96, 3.76, 0.66, -15.24, 147.56, 150.06, 24.36] = 395.4

G1 X G1 = [23.04, 10.24, 38.44, 14.44, 0.64, 0.04, 1.44, 46.24, 67.24, 7.84] = 209.6

G2 X G2 = [5.29, 176.89, 86.49, 1.69, 22.09, 10.89, 161.29, 470.89, 334.89, 75.69] = 1346.1

در نهایت پاسخ برابر است با:

$$\frac{۳۹۵.۴}{\sqrt{۲۰۹.۶ * ۱۳۴۶.۱}} = ۰.۷۴$$

معیار اطلاعات متقابل برابر است با:

P(G1, G2)	-3	5	8	-2	1	2	3	-5	10	-1	P(G2)
9	0.1	0	0	0	0	0	0	0	0	0	0.1
20	0	0.1	0	0	0	0	0	0	0	0	0.1
16	0	0	0.1	0	0	0	0	0	0	0	0.1
8	0	0	0	0.1	0	0	0	0	0	0	0.1
2	0	0	0	0	0.1	0	0	0	0	0	0.1
10	0	0	0	0	0	0.1	0	0	0	0	0.1
-6	0	0	0	0	0	0	0.1	0	0	0	0.1
-15	0	0	0	0	0	0	0	0.1	0	0	0.1
25	0	0	0	0	0	0	0	0	0.1	0	0.1
-2	0	0	0	0	0	0	0	0	0	0.1	0.1
P(G1)	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

بنابراین خواهیم داشت که:

$$MI = 10\,(0.1\,log\,(\frac{0.1}{0.1 * 0.1})) = log\,10 = 1$$

ب) طبق نتایج هر معیار مشخص کنید آیا دو ژنوم از یکدیگر مستقل هستند یا خیر.

هرچه similarity cosine دو سری از اعداد به یک نزدیکتر باشد شباهت آنها بیشتر است. بنابراین طبق این معیار دو ژنوم شباهت نسبتاً بالایی دارند. همچنین معیار correlation نیز وابستگی خطی دو ژنوم را به یکدیگر نشان خواهد داد و هر چه به یک نزدیکتر باشد این وابستگی قویتر است. لذا طبق این معیار نیز دو ژنوم شباهت نسبتاً باالایی دارند. هر چه مقدار information mutual باالاتر باشد نیز به این معناست که با داشتن مقادیر ژنوم اول با دقت باالاتری میتوان مقادیر ژنوم دوم را بدست آورد. بنابراین بالا بودن این معیار نیز نشان‌دهنده وابستگی زیاد دو ژنوم است.

ج) آیا نتایج به دست آمده متفاوت است؟ اگر پاسخ مثبت است علت آن را توضیح دهید.

نتایج بدست آمده همه نشانگر شباهت نسبتاً بالای دو ژنوم هستند اما تا حدی متفاوت‌اند. دلیل این تفاوت این است که هر یک از معیارهای فوق یک جنبه از شباهت دو ژنوم را بررسی میکنند similarity cosine میزان شباهت دو ژنوم، correlation میزان وابستگی خطی میان آن دو و information mutual هر گونه وابستگی میان دو ژنوم و نه فقط وابستگی خطی، را نشان میدهد.

سوال ششم

دو مورد از روش‌های data preprocessing روش‌های aggregation و sampling هستند. این دو روش را توضیح داده و مزایا و معایب هر یک را بنویسید.

دو روش پردازش داده ها در داده کاوی شامل تجمیع داده و نمونه برداری داده است.

تجمیع داده فرایندی است که در آن داده ها به صورت خلاصه یا تجمیعی ترکیب می شوند که می تواند شامل گروه بندی داده ها بر اساس ویژگی های خاصی مانند زمان، مکان یا دسته بندی باشد و آماره های خلاصه ای مانند میانگین، جمع ها، شماره ها یا درصدها را محاسبه کند. تجمیع داده می تواند پیچیدگی و اندازه مجموعه داده را کاهش داده و در عین حال به بینش انتزاعی تر و خلاصه تر از داده ها کمک کند.

مزایا:

- ساده سازی داده های پیچیده
- کاهش اندازه داده
- افزایش عملکرد

معایب:

- از دست دادن جزئیات: تجمیع داده ممکن است منجر به از دست دادن جزئیات خاصی شود که در آماره های خلاصه شده وجود داشته و این ممکن است توانایی تجزیه و تحلیل داده را در سطح دقیق کاهش دهد
- امکان از دست دادن دقت
- امکان به وجود آمدن bias

نمونه‌برداری داده یک روش دیگر از پردازش داده است که در آن یک زیرمجموعه کوچکتر از داده ها از مجموعه داده اصلی برداشته می شود. این زیرمجموعه ممکن است به صورت تصادفی یا با الگوریتم های خاصی انتخاب شود. نمونه‌برداری داده می تواند در تجزیه و تحلیل داده ها، کاهش حجم داده ها، و بررسی اولیه داده ها به کار رود.

مزایا:

- کاهش حجم داده
- امکان تجزیه و تحلیل سریعتر و کارآمدتر
- کاهش هزینه و زمان

معایب:

- از دست دادن جزئیات
- امکان به وجود آمدن bias sampling
- نیازمند طراحی دقیق نمونه برداری

سوال هفتم

در رابطه با کاهش بعد تحقیق کرده و به سوالات زیر پاسخ بدهید.

الف) مفاهیم انتخاب ویژگی[8]، استخراج ویژگی[9] و مهندسی ویژگی[10] را توضیح و تفاوت‌های بین آن‌ها را بیان کنید.

مهندسی ویژگی فرآیند انتخاب و تبدیل متغیرها و ویژگیها از داده خام جمع‌آوری‌شده است، به نحوی که بتوان مدل را با این ویژگیها آموزش داد. بنابراین هدف مهندسی ویژگی این است که با استفاده از اطلاعات موجود در داده‌های خام یکسری ورودی برای مدل تهیه کند و این ورودیها را به نحوی انتخاب کند و یا از روی دادهها بسازد که کارآیی آن را افزایش دهد .انتخاب ویژگی فرآیندی است که در آن دستهای از مرتبط‌ترین ویژگیها با هدف مورد نظر انتخاب میشوند تا به عنوان ورودی به مدل داده شوند (مانند روش انتخاب حریصانه ویژگیها). با این کار ابعاد داده ورودی کاهش پیدا

[8] Feature selection
[9] Feature extraction
[10] Feature engineering

میکند و هدف آن بهبود عملکرد مدل با استفاده از حذف ویژگیهای نامربوط و اضافه است که میتوانند باعث اضافه شدن نویز به مدل و یا overfit شدن آن شوند.

استخراج ویژگی اطلاعاتی را از ویژگیهای اصلی استخراج کرده و از طریق آن ویژگیهای جدیدی میسازد (مانند روشPCA). هدف اصلی این کار فشرده‌سازی اطلاعات مفید ویژگیهای مرتبط است، به نحوی که بیشترین استفاده از این اطلاعات شود. همچنین مانند فرآیند انتخاب ویژگی در این فرآیند نیز ابعاد داده کاهش پیدا کرده و با کاهش پیچیدگی مدل و overfitting در آن کارایی مدل افزایش پیدا میکند. بنابراین در انتخاب ویژگی بر خالف استخراج ویژگی، ویژگیهای اولیه همانطور که هستند باقی میمانند و لذا زمانی که قابل توضیح بودن مدل برای ما اهمیت دارد بهتر است از روش انتخاب ویژگی استفاده کنیم. در مقابل استخراج ویژگی میتواند اطلاعات بیشتری را برای آموزش مدل حفظ کند.

مهندسی ویژگی نیز معمولاً پیش از دو مورد دیگر انجام میگیرد تا ویژگیهای کلی از دادههای خام استخراج شوند. سپس با استفاده از دو روش دیگر میتوان ویژگیهای مطلوب را از میان تمام ویژگیها، انتخاب و یا استخراج کرد.

ب) الگوریتم‌های کاهش بعد به دو دسته خطی و غیرخطی تقسیم می‌شوند. تفاوت این دو دسته را توضیح داده و روش کار الگوریتم PCA از دسته خطی و الگوریتم t-sne از دسته غیرخطی را توضیح دهید.

به طور کلی هدف الگوریتمهای کاهش بعد این است که یکسری بردار با ابعاد بالا را به یکسری بردار با ابعاد پایینتر نگاشت کنند؛ یعنی هر یک از بردارها با ابعاد بالاتر با یک بردار با ابعاد پایینتر نمایندگی شود. در کاهش بعد خطی تبدیل این بردارها به یکدیگر توسط یک تبدیل خطی انجام میگیرد، در حالی که در الگوریتمهای کاهش بعد غیرخطی این تبدیل نیز غیرخطی است .به طور کلی الگوریتم PCA بردارهایی را استخراج میکند که داده‌ها در جهت آن واریانس بالایی دارند. پیشنیاز الگوریتم PCA نرمالسازی داده‌هاست .پس از آن k بردار orthonormal استخراج میشود (components principal)، به صورتی که هر یک از دادههای ورودی ترکیب خطیای از این k بردار باشند. این بردارهای یکه در حقیقت بردارهای ویژه ماتریس کوواریانس دادههای نرمالسازی شده هستند (این ماتریس زمانی که fویژگی داشته باشیم یک ماتریس f × f است و لذا میتواند f بردار ویژه متمایز داشته باشد). پس از آن، این k بردار بر حسب اهمیت مرتب میشوند و بردارهایی که اهمیت کمتری داشته و اطلاعات کمتری به ما میدهند (یعنی مربوط به مقدار ویژه کوچکتری بوده و دادهها در آن جهت واریانس کمتری دارند) میتوانند حذف شوند. تعداد PC هایی که در نهایت انتخاب میشوند، تعداد feature های استخراجشده خروجی است. حال اگر این بردارها را درکنار یکدیگر قرار داده و ماتریسی به نام V ایجاد کنیم، و اگر ماتریس دادهها را D بنامیم، ماتریس DV مقدار دادههای تبدیل شده و یا scores component principle را در فضای PC ها میدهد.

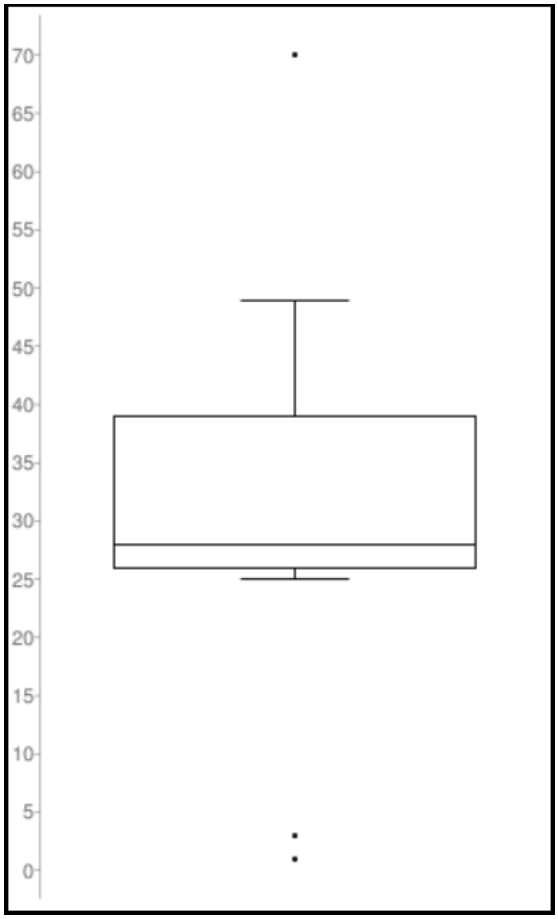
در روش t-sne در ابتدا بردارهای با ابعاد پایینتر به صورت تصادفی مقداردهی میشوند .سپس در فضای با ابعاد بالاتر میزان شباهت تمام نقاط با یکدیگر محاسبه میشود. محاسبه این میزان شباهت به این صورت است که ابتدا یکی از نقاط در نظر گرفته میشود، سپس فاصله میان این نقطه با هر یک از نقاط دیگر محاسبه میشود. پس از آن این فاصله بر روی محور یک نمودار با توزیع نرمال (به مرکزیت نقطه انتخاب شده) قرار داده شده و مقدار نمودار به ازای فاصله فوق خوانده میشود. سپس تمام این مقادیر (به ازای فواصل نقطه انتخاب شده و دیگر نقاط) نرمالسازی میشوند تا مجموع 1 داشته باشند. این عملیات به ازای تمام نقاط تکرار میشود تا میزان شباهت تمام نقاط با یکدیگر محاسبه شوند (از آنجایی که میزان شباهت بدست آمده برای نقطه A با B و B با A میتواند متفاوت باشد میان دو عدد بدست آمده میانگین گرفته

میشود). در مرحله بعد همین کار برای نقاط بر روی فضای با ابعاد پایینتر تکرار میشود، با این تفاوت که به جای توزیع نرمال از توزیع t-student استفاده میشود. حال با استفاده از روش گرادیان کاهشی نقاط بر روی فضای با ابعاد پایینتر را طوری حرکت میدهیم که ماتریس شباهت‌ها برای نقاط در ابعاد پایینتر شبیه به ماتریس شباهت‌ها برای نقاط در ابعاد بالاتر شود. با این کار در انتها نقاطی که در ابعاد بالاتر به یکدیگر نزدیکتر هستند در ابعاد پایین‌تر نیز به یکدیگر نزدیکتر خواهند بود. همانطور که واضح است در این روش دیگر نگاشت میان نقاط با ابعاد بالاتر و پایینتر خطی نیست.

سوال هشتم

برای داده‌های عددی زیر نمودار جعبه[11] را رسم کنید.

41, 28, 39, 25, 49, 36, 27, 33, 26, 70, 27, 29, 1, 3, 27



سوال نهم

همانطور که می‌دانید، یکی از روش‌های مقایسه دو توزیع آماری استفاده از روش q-q plot است.

[11] Box plot

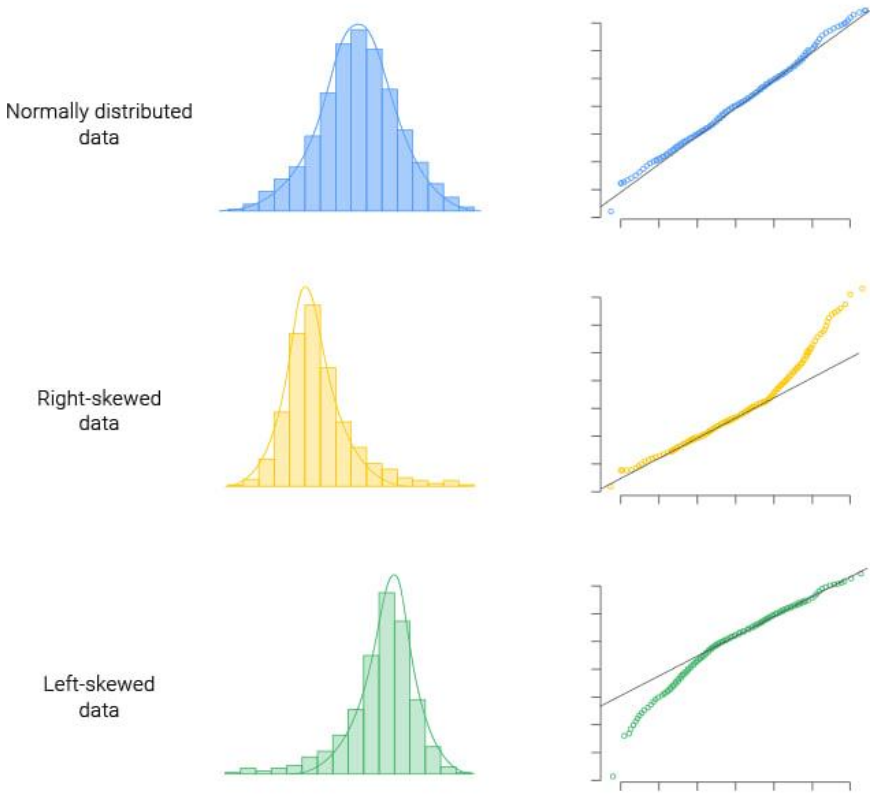
الف) نحوه کار این روش را توضیح دهید.

نمودار quantile-quantile یک روش بصری است برای اینکه بررسی کنیم که آیا یک مجموعه داده دارای یک توزیع خاص (مانند نرمال، یا یکنواخت یا...) است یا خیر، یا به طور کلی‌تر آیا دو مجموعه داده توزیع یکسانی دارند یا خیر. البته این روش یک پاسخ دقیق به ما نخواهد داد، اما کمک میکند که بررسی کنیم که آیا فرضی که در رابطه با توزیع یک مجموعه داده داشته‌ایم درست است یا خیر، یا در صورتی که درست نیست کدام بخش از داده‌ها باعث نقض آن شده‌اند.Q-Q plot در حقیقت یک scatterplot است که از رسم quantile های متناظر دو مجموعه داده، هر یک به عنوان یک نقطه از نمودار، بدست می‌آید. در صورتی که این quantile ها از یک توزیع یکسان آمده باشند، مجموعه نقاط فوق تقریباً یک خط راست را تشکیل میدهند (یا در صورتی که توزیع‌ها یکسان نباشند اما با یکدیگر رابطه خطی داشته باشند).

ب) نمودار q-q plot می‌تواند به شکل‌های متفاوتی نمایان شود: به طور مثال شبیه یک خط راست مورب. سه نوع از این شکل‌های متفاوت را بررسی کنید و تحلیل خود داده‌های توزیع‌های آماری ورودی به آن را بنویسید. به نظر شما از روی شکل q-q plot چه مواردی در مورد توزیع‌های آماری اولیه قابل استنتاج است؟

همانطور که در شکل زیر دیده میشود، اگر سه نمودار توزیع نرمال، چوله به چپ و چوله به راست را طوری زیر هم قرار دهیم که مقدار میانه برای آنها برابر باشد، برای نمودار چوله به چپ مقادیر چارک‌های اول و سوم نسبت به توزیع نرمال کمتر بوده (در حالی که میزان اختالف برای Q3 کمتر و برای Q1 بیشتر است) و برای یک نمودار چوله به راست مقادیر چارک‌های اول و سوم نسبت به توزیع نرمال بیشتر است (در حالی که میزان اختلاف برای Q3 بیشتر و برای Q1 کمتر است).

حال فرض کنید نمودار Q-Q را برای یک مجموعه داده با هر یک از توزیع های فوق و توزیع نرمال رسم کرده باشیم. با توجه به توضیحات داده شده این نمودارها مانند زیر خواهند بود:



بنابراین میتوان با رسم نمودار Q-Q نرمال اطلاعاتی در رابطه با شکل کلی و چولگی توزیع داده‌ها بدست آورد. در اینجا این مقایسه برای توزیع یک مجموعه داده دلخواه و توزیع نرمال انجام شد، اما با همین روش برای هر دو نمودار میتوان به طور تقریبی بررسی کرد که فرم کلی توزیع آنها چه ارتباطی با هم دارد و quantile های آنها چقدر و چگونه با یکدیگر فاصله دارند.

با استنتاج از روی شکلهای نمودار q-q، میتوان تحلیلهای مختلفی انجام داد. به طور کلی، اگر نقاط بر روی خط قطری متمایل با زاویه 45 درجه قرار داشته باشند، میتوان نتیجه گرفت که دادهها به طور کامل با توزیع مقایسهای تطابق دارند. اگر نقاط بر روی خط راست افقی یا عمودی قرار داشته باشند، میتوان نتیجه گرفت که تطابق کامل بین یکی از کوانتیل‌ها دو توزیع وجود دارد، اما تطابق کامل با بقیه کوانتیل‌ها وجود ندارد. اگر نقاط در شکل منحنی ظاهر شوند، میتوان نتیجه گرفت که تفاوتهای غیرخطی و احتمالی بین دو توزیع وجود دارد .استنتاج‌های ممکن از روی نمودار q-q بستگی به شکل نمودار و تطابق میزان دادهها با توزیع مقایسه‌ای (معمولاً توزیع نرمال) دارد. از نمودار q-qبرای بررسی تطابق دادهها با توزیعهای مختلف، تشخیص تفاوت‌های غیرخطی بین دو توزیع، و همچنین تشخیص نقاط پرتی و ناهنجاریها در دادهها استفاده می شود. این نمودار میتواند یک ابزار قدرتمند برای تحلیل و بررسی توزیعهای آماری ورودی و مقایسه ی آنها با توزیع نرمال باشد

سوال دهم

برای هر یک از روش‌های نرمال‌سازی زیر تحقیق کرده و بازه‌ی اعداد را مشخص کنید.

الف) نرمال‌سازی min-max

داده ها را در محدوده ای بین 0 و 1 مقیاس می کند، که در آن حداقل مقدار در مجموعه داده ها به 0 و حداکثر مقدار به 1 نگاشت می شود. همه مقادیر دیگر به تناسب بین این دو مقدار مقیاس بندی می شوند.

ب) نرمال‌سازی z-score

داده ها را با میانگین 0 و انحراف استاندارد 1 مقیاس می کند. فرمول آن به صورت $\frac{x-mean}{standard\ deviation}$، که در آن x مقدار اصلی است.

ج) نرمال‌سازی با مقیاس‌دهی[12]

داده ها را با تقسیم هر مقدار بر توان 10 مقیاس می کند. توان 10 بر اساس بزرگترین مقدار مطلق در مجموعه داده انتخاب می شود. به عنوان مثال، اگر بزرگترین مقدار مطلق 500 باشد، داده ها با تقسیم هر مقدار بر 1000 مقیاس بندی می شوند و در نتیجه مقادیری بین 0.5- و 0.5 به دست می آیند.

سوال یازدهم

با توجه به مقادیر ورودی X و مقادیر هدف Y می‌توان یک برازش خطی یا غیرخطی بر روی بسیاری از دادگان‌ها ایجاد کرد. با توجه به این مقادیر، به سوالات زیر پاسخ دهید.

$$X = [2, 4, 1, 3, 2, 6], \qquad Y = [5, 6, 3, 6, 3, 10]$$

الف) روش محاسبه معادله نرمال[13] را با استفاده از روش محاسبه مشتق جزئی باقی‌مانده[14] کامل شرح دهید.

در رگرسیون، مدل به شکل زیر می‌باشد.

$$h_\theta(x) = x^T\boldsymbol{\theta} = x_0\theta_0 + x_1\theta_1 + \cdots + x_n\theta_n$$

$$\hat{\boldsymbol{y}} = X\boldsymbol{\theta}$$

که در اینجا، ماتریس X همان ماتریس ویژگی‌ها، بردار $\boldsymbol{\theta}$ ، بردار وزن و بردار $\boldsymbol{y}$ بردار هدف می‌باشد. بنابراین می‌توان خطا را به صورت زیر نمایش داد:

$$L(\boldsymbol{\theta}) = \|\boldsymbol{y} - X\boldsymbol{\theta}\|^2 = (\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta}) = \boldsymbol{y}^T\boldsymbol{y} - (X\boldsymbol{\theta})^T\boldsymbol{y} - \boldsymbol{y}^TX\boldsymbol{\theta} + (X\boldsymbol{\theta})^TX\boldsymbol{\theta}$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2(X\boldsymbol{\theta})^T\boldsymbol{y} + \boldsymbol{\theta}^TX^TX\boldsymbol{\theta}$$

[12] decimal scaling
[13] Normal Equation
[14] residual

همانطور که از قبل می‌دانیم برای پیدا کردن مینیمم تابع، در صورت محدب بودن تابع L می‌توان با برابر صفر قرار دادن مشتق جزئی، نقطه اکسترمم تابع را یافت. حال مشتق جزئی این تابع را نسبت به بردار $\boldsymbol{\theta}$ می‌نویسیم:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\theta} = 0$$

$$X^TX\boldsymbol{\theta} = X^T\boldsymbol{y}$$

$$\boldsymbol{\theta} = (X^TX)^{-1}X^T\boldsymbol{y}$$

ب) یک برازش خطی ($Y = \beta_1 X + \beta_0$) را برای این دادگان محاسبه کنید. (مقدار خطای برازش را نیز به دست آورید)

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 1 \\ 1 & 3 \\ 1 & 2 \\ 1 & 6 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 5 \\ 6 \\ 3 \\ 6 \\ 3 \\ 10 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 1 & 3 & 2 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 1 \\ 1 & 3 \\ 1 & 2 \\ 1 & 6 \end{bmatrix} = \begin{bmatrix} 6 & 18 \\ 18 & 70 \end{bmatrix}$$

$$(X^TX)^{-1} = \frac{1}{420 - 324} \begin{bmatrix} 70 & -18 \\ -18 & 6 \end{bmatrix}$$

$$X^T\boldsymbol{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 1 & 3 & 2 & 6 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \\ 3 \\ 6 \\ 3 \\ 10 \end{bmatrix} = \begin{bmatrix} 33 \\ 121 \end{bmatrix}$$

$$\boldsymbol{\theta} = (X^TX)^{-1}X^T\boldsymbol{y} = \frac{1}{96} \begin{bmatrix} 70 & -18 \\ -18 & 6 \end{bmatrix} \begin{bmatrix} 33 \\ 121 \end{bmatrix} = \begin{bmatrix} 1.375 \\ 1.375 \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = X\boldsymbol{\theta} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 1 \\ 1 & 3 \\ 1 & 2 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} 1.375 \\ 1.375 \end{bmatrix} = \begin{bmatrix} 4.125 \\ 6.875 \\ 2.75 \\ 5.5 \\ 4.125 \\ 9.625 \end{bmatrix}$$

$$residual = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} 0.875 \\ -0.875 \\ 0.25 \\ 0.5 \\ -1.125 \\ 0.375 \end{bmatrix}$$

ج) یک برازش غیرخطی ($Y = \beta_2 X^2 + \beta_1 X + \beta_0$) برای این دادگان محاسبه کنید. (مقدار خطای برازش را نیز به دست آورید)

$$X = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & 2 & 4 \\ 1 & 6 & 36 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 5 \\ 6 \\ 3 \\ 6 \\ 3 \\ 10 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 1 & 3 & 2 & 6 \\ 4 & 16 & 1 & 9 & 4 & 36 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & 2 & 4 \\ 1 & 6 & 36 \end{bmatrix} = \begin{bmatrix} 1 & 18 & 70 \\ 18 & 70 & 324 \\ 70 & 324 & 1666 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 3.270787 & -2.052809 & 0.261798 \\ -2.052809 & 1.431461 & -0.192135 \\ 0.261798 & -0.192135 & 0.026966 \end{bmatrix}$$

$$X^T \boldsymbol{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 1 & 3 & 2 & 6 \\ 4 & 16 & 1 & 9 & 4 & 36 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \\ 3 \\ 6 \\ 3 \\ 10 \end{bmatrix} = \begin{bmatrix} 33 \\ 121 \\ 545 \end{bmatrix}$$

$$\boldsymbol{\theta} = (X^T X)^{-1} X^T \boldsymbol{y} = \begin{bmatrix} 3.270787 & -2.052809 & 0.261798 \\ -2.052809 & 1.431461 & -0.192135 \\ 0.261798 & -0.192135 & 0.026966 \end{bmatrix} \begin{bmatrix} 33 \\ 121 \\ 545 \end{bmatrix} = \begin{bmatrix} 2.225843 \\ 0.750562 \\ 0.087640 \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = X\boldsymbol{\theta} = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & 2 & 4 \\ 1 & 6 & 36 \end{bmatrix} \begin{bmatrix} 2.225843 \\ 0.750562 \\ 0.087640 \end{bmatrix} = \begin{bmatrix} 4.0775 \\ 6.6303 \\ 3.0640 \\ 5.2663 \\ 4.0775 \\ 9.8843 \end{bmatrix}$$

$$residual = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} 0.922472 \\ -0.630337 \\ -0.064045 \\ 0.733708 \\ -1.077528 \\ 0.115730 \end{bmatrix}$$