نام و نام خانوادگی: حسنا اویارحسینی	تمرین اول داده کاوی
شماره دانشجویی: ۹۸۲۳۰۱۰	تاریخ: فروردین - ۱۴۰۲

سوال ۱)

الف) نویز را می توان به عنوان نمونه هایی با برچسب نادرست (نویز کلاس) یا نمونه های که در مقادیر ویژگی آن ها خطایی وجود دارد (نویز صفت) تعریف کرد. اما داده پرت مفهوم گسترده تری است که نه تنها شامل داده هایی می شود که دچار خطا شده اند بلکه شامل داده های معتبر اما ناسازگاری که ممکن است از تغییرات طبیعی در جامعه یا فرآیند ناشی شود هم می شود در واقع آنها اغلب خیلی کوچک یا خیلی بزرگ هستند و عموماً از پیروی از یک الگو که در بقیه داده ها برقرار است سرکشی می کنند.. شاید بتوان گفت داده پرت خارج از محدوده داده‌ای است که ما انتظار داریم اما نویز داده‌ای ناخواسته و اشتباه است و باید حذف شود.

ب) داده پرت می تواند اطلاعات ارزشمندی را در یک آزمایش در اختیار ما بگذارد. آنها چیزی منحصر به فرد در مورد یک موقعیت به ما می گویند. اگر ما بفهمیم که چرا یک امر دور افتاده رخ می دهد، به ما کمک می کند تا یک مشکل را به روشی بهتر حل کنیم. موارد پرت اغلب می توانند شروع چیز جدیدی را نشان دهند. برای مثال داده های پرت باعث می شوند تا به سوالات زیر فکر کنیم. آیا هنگام اندازه گیری این مشاهدات، اتفاق غیرعادی مانند قطع برق، شرایط آزمایشی غیرعادی یا هر چیز خارج از حد معمول رخ داده است؟ آیا خطاهای اندازه گیری یا ورود داده رخ داده است؟ و در صورت وجود ناهنجاری متوجه آن شویم و اگر آسیب زا است آن را برطرف کنیم. برای مثال در شناسایی تقلب با حالت خیلی خاص (مانند استعداد برتر) داده پرت مفید است.

ج) بله. یک نویز میتواند یک داده پرت باشد. برای مثال اگر در هنگام اندازه گیری مقادیر دستگاه مورد استفاده برای اندازه گیری دچار خطا شود و یک داده با مقدار خیلی کوچک یا خیلی بزرگ تولید کند یک نویز داریم که داده پرت نیز می باشد.

سوال ۲)

- انبار داده یا Data Warehouse مجموعه بزرگی از داده‌های تجاری است که اطلاعات بسیار ساختار یافته را از منابع مختلف برای ایجاد نوعی دیدگاه تجاری ذخیره میکند. انبار داده‌ها برای اتصال، گزارش دهی، بررسی و تحلیل داده‌های تجاری از منابع مختلف مورد استفاده قرار می‌گیرد و هسته اصلی سیستم هوش تجاری به شمار می‌رود. می‌توانید انبار داده‌ها را یک بانک اطلاعاتی در نظر بگیرید که داده‌های گذشته و فعلی را در یک مکان واحد جمع‌آوری می‌کند. حجم زیادی از داده‌های موجود در انبارهای داده از منابع مختلفی جمع‌آوری می‌شوند؛ برنامه‌های کاربردی داخلی مانند بازاریابی، فروش و امور مالی

نمونه‌هایی از این منابع هستند. در واقع میتوان گفت انبارهای داده، مخازن مرکزی داده‌های یکپارچه از یک یا چند منبع پراکنده هستند که برای پشتیبانی از طیف گسترده ای از تصمیم سازی ها در یک سازمان خاص طراحی شده است.

- تفاوت های کلیدی بین پایگاه داده و انبار داده عبارتند از:

تفاوت ها	Database	Data warehouse
تعریف	یک پایگاه داده داده های فعلی مورد نیاز برای یک برنامه را ذخیره می کند.	یک انبار داده، داده های جاری و قدیمی را از یک یا چند سیستم در یک طرح واره از پیش تعریف شده و ثابت ذخیره می کند، که به تحلیلگران تجاری و دانشمندان داده اجازه می دهد تا به راحتی داده ها را تجزیه و تحلیل کنند.
بار کاری	عملیاتی	تحلیلی
انعطاف پذیری طرحواره	طرحواره سفت یا انعطاف پذیر بسته به نوع پایگاه داده	تعریف طرحواره از پیش تعریف شده و ثابت
به روز بودن داده ها	Real time	ممکن است بر اساس فراوانی فرآیندهای ETL به روز نباشد
کاربران	توسعه دهنده ها	تحلیلگران کسب و کار و دانشمندان داده
کاربرد	پایگاه های داده برای تراکنش های کوچک و اتمی بیشترین کاربرد را دارند. انجام تغییرات و تعداد درخواست های بالا	انبارهای داده برای پرسش‌های تجاری بزرگ‌تر که به سطح بالاتری از تجزیه و تحلیل داده‌ها نیاز دارند، مناسب‌تر هستند. خواندن داده ها و درخواست های کمتر

انبارهای داده از ابتدا برای اهداف گزارش دهی و تجزیه و تحلیل با استفاده از داده های تجاری تاریخی که مرتبط است طراحی شده اند و داده های قدیمی را نیز نگهداری میکنند.	پایگاه‌های داده معمولاً فقط حاوی به‌روزترین اطلاعات هستند، که جستجوهای تاریخی را غیرممکن می‌سازد	داده ها
منبع داده ها معمولا چندین مورد بوده است.	داده ها معمولا از یک منبع گرداوری شده اند.	منبع داده ها

- شباهت ها:

- هر دو وظیفه نگهداری و ذخیره داده های ساختار یافته یا نیمه ساختار یافته را دارند.
- هر دو را می توان پرس و جو کرد و با تراکنش ها به روز کرد.
- هر دوی آنها حاوی داده هایی درباره یک یا چند نهاد مانند مشتریان و محصولات هستند.
- به طور کلی، لایه پایین انبار داده یک سیستم پایگاه داده رابطه ای است. پایگاه های داده نیز سیستم پایگاه داده رابطه ای هستند. سیستم های DB رابطه ای از سطرها و ستون ها و مقدار زیادی داده تشکیل شده است.
- هر دو از دسترسی چند کاربره پشتیبانی می کنند. یک نمونه واحد از پایگاه داده و انبار داده می تواند توسط بسیاری از کاربران در یک زمان قابل دسترسی باشد.

سوال ۳)

صدک یا percentile مقداری است در یک توزیع نرمال که درصد مشخصی از مشاهدات زیر آن است .در اصل روش توزیع نرمال و صدک یک روش آماری است که برای تشخیص داده‌های پرت در یک مجموعه داده استفاده می‌شود. در این روش، ابتدا توزیع نرمال (یا گاوسی) برای داده‌های مجموعه تعریف می‌شود. توزیع نرمال یک توزیع پیوسته است که به شکل منحنی بلند و باریکی می‌باشد.

سپس با استفاده از توزیع نرمال، با استفاده از صدک ها می‌توان داده‌های پرت را تشخیص داد. صدک ها یک مقداری است که n درصد از داده‌های مجموعه کوچکتر یا مساوی آن هستند. به عبارت دیگر، اگر $n=۹۵$ ، صدک

۹۵برابر با مقداری است که ۹۵ درصد از داده‌های مجموعه کوچکتر یا مساوی آن هستند. برای تشخیص داده‌های پرت، داده‌هایی که خارج از محدوده صدک مشخصی هستند، به عنوان داده‌های پرت شناخته می‌شوند.
به عنوان مثال، نقاط داده ای که از percentile ۹۹ درصد فاصله دارند و کمتر از percentile ۱ هستند، نقطه پرت در نظر گرفته می شوند.
و یا در حالتی دیگر روش ابتدا داه ها را مرتب میکنیم
سپس ۳Q۱, Q۲, Q 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 چهارک داده ها از آن کوچکترند می باشند)
در مرحله بعد ۱IRQ=Q۳-Q را محاسبه میکنیم.
در مقدار ۱,۵IRQ + Q۳ , Q۱ - ۱,۵IRQ را به عنوان ابتدا و انتهای بازه محاسبه میکنیم.
داده هایی که خارج از بازه مرحله قبل می باشند یعنی کوچکتر از Q۱ - ۱,۵IRQ و بزرگتر از Q۳ + ۱,۵IRQ را به عنوان داده پرت معرفی میکنیم.

سوال ۴)

الف) پاکسازی داده ها (Data cleaning)، شامل شناسایی و رفع خطاهای احتمالی داده‌ها برای بهبود کیفیت آنهاست. در این فرآیند، داده‌های «کثیف» را شناسایی، بررسی، تجزیه و تحلیل، اصلاح یا حذف می‌کنیم تا مجموعه داده‌های خود را پاکسازی کنیم. داده‌های کثیف به معنی ناهماهنگی‌ها و خطاها هستند که می‌توانند از هر بخش فرآیند تحقیق، مانند طراحی ضعیف، اندازه گیری غلط، ورود داده‌های ناقص و... به دست آیند.
این فرایند میتواند شامل مراحل زیر باشد:
- مشاهدات تکراری یا غیرمرتبط را حذف کنید
- رفع خطا های ساختاری
- داده های دور افتاده نامطلوب را فیلتر کنید
- رسیدگی به داده های گمشده
- اعتبارسنجی و اطمینان از کیفیت مجموعه داده

ب) اهمیت تجسم داده ها در تجزیه و تحلیل داده های پیچیده، شناسایی الگوها و استخراج بینش های ارزشمند است. ساده سازی اطلاعات پیچیده و ارائه آنها به صورت بصری، تصمیم گیرندگان را قادر می سازد تا تصمیمات

آگاهانه و موثر را سریع و دقیق اتخاذ کنند. اما یک از چالش های آن تعداد زیاد داده، محدوده گسترده برای هر ویژگی از داده و ابعاد بالای داده می باشد.

ج) زیرا با انجام پاکسازی داده دقت و صحت و مورد اعتماد بودن تجسم داده ها افزایش می یابد. برای مثال اگر پاکسازی داده انجام ندهیم و داده های پرت در میان داده های باقی بماند محدوده مقدار مربوط به یک ویژگی از داده ها به علت وجود چند داده پرت ممکن است گسترش بیابد و دقت در نمایش کاهش بیابد و اکثر داده ها که در بازه کوچک تری هستند به صورت فشرده تر و با دقت کمتری نمایش داده شوند. از طرفی اگر در پیش پردازش missing value ها را برطرق و یا حذف نکنیم در هنگام نمایش ممکن است به علت نداشتن مقادیر مورد نیاز به مشکل بخوریم.

سوال ۵)

الف)

- Cosine similarity:

$$\frac{[(-۳*۹)+(۵*۲۰)+(۸*۱۶)+(-۲*۸)+(۱*۲)+(۲*۱۰)+(۳*-۶)+(-۵*-۱۵)+(۱۰*۲۵)+(-۱*-۲)]}{||G1||.||G2||} =$$

$$\frac{۵۱۶}{[(۹+۲۵+۶۴+۴+۱+۴+۹+۲۵+۱۰۰+۱)(۸۱+۴۰۰+۲۵۶+۶۴+۴+۱۰۰+۳۶+۲۲۵+۶۲۵+۴)]^{۰.۵}} =$$

$$\frac{۵۱۶}{۶۵۹.۰۸۲۶۹۵۸۷۴} = ۰.۷۸۲۹۰۶۳۰۸$$

- Correlation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{-۳ + ۵ + ۸ + -۲ + ۱ + ۲ + ۳ + -۵ + ۱۰ + -۱}{۱۰} = ۱.۸$$

$$\bar{y} = \frac{۹ + ۲۰ + ۱۶ + ۸ + ۲ + ۱۰ + -۶ + -۱۵ + ۲۵ + ۲}{۱۰} = ۷.۱$$

$$\Sigma(x_i - \bar{x})^۲ = (-۳ - ۱.۸)^۲ + (۵ - ۱.۸)^۲ + (۸ - ۱.۸)^۲ + (-۲ - ۱.۸)^۲ + (۱ - ۱.۸)^۲ + (۲ - ۱.۸)^۲ + (۳ - ۱.۸)^۲ + (-۵ - ۱.۸)^۲ + (۱۰ - ۱.۸)^۲ + (-۱ - ۱.۸)^۲ = ۲۰۹.۶$$

$$\Sigma(y_i - \bar{y})^2$$
$$= (۹ - ۷.۱)^2 + (۲۰ - ۷.۱)^2 + (۱۶ - ۷.۱)^2 + (۸ - ۷.۱)^2 + (۲ - ۷.۱)^2$$
$$+ (۱۰ - ۷.۱)^2 + (-۶ - ۷.۱)^2 + (-۱۵ - ۷.۱)^2 + (۲۵ - ۷.۱)^2$$
$$+ (۲ - ۷.۱)^2 = ۱۲۹۰.۹$$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y})$$
$$= (-۳ - ۱.۸) * (۹ - ۷.۱) + (۵ - ۱.۸) * (۲۰ - ۷.۱) + (۸ - ۱.۸)$$
$$* (۱۶ - ۷.۱) + (-۲ - ۱.۸) * (۸ - ۷.۱) + (۱ - ۱.۸) * (۲ - ۷.۱)$$
$$+ (۲ - ۱.۸) * (۱۰ - ۷.۱) + (۳ - ۱.۸) * (-۶ - ۷.۱) + (-۵ - ۱.۸)$$
$$* (-۱۵ - ۷.۱) + (۱۰ - ۱.۸) * (۲۵ - ۷.۱) + (-۱ - ۱.۸) * (۲ - ۷.۱)$$
$$= ۳۸۴.۲$$

$$r = \frac{\Sigma(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})}{\sqrt{(\Sigma(\mathrm{x_i} - \bar{\mathrm{x}})^2 \Sigma(\mathrm{y_i} - \bar{\mathrm{y}})^2)}} = \frac{۳۸۴.۲}{\sqrt{(۲۰۹.۶ * ۱۲۹۰.۹)}} = ۰.۷۳۸۶$$

- Mutual Information:

Joint probabilityاین دو بردار در واقع فراوانی هر ترکیبی از مقادیر را نشان می دهد:

$$Joint\ probability =$$
[[۰. , ۰. , ۰. , ۰. , ۰. , ۰. , ۰.۰۵, ۰. , ۰. , ۰.],
[۰. , ۰. , ۰. , ۰. , ۰. , ۰.۱ , ۰. , ۰. , ۰. , ۰.],
[۰. , ۰. , ۰. , ۰. , ۰.۱ , ۰.۱ , ۰. , ۰. , ۰. , ۰.],
[۰. , ۰. , ۰. , ۰.۱ , ۰. , ۰. , ۰.۱ , ۰. , ۰. , ۰.],
[۰. , ۰. , ۰. , ۰. , ۰.۱ , ۰. , ۰. , ۰. , ۰. , ۰.],
[۰. , ۰. , ۰.۱ , ۰. , ۰.۱ , ۰.۱ , ۰. , ۰. , ۰. , ۰.],
[۰. , ۰. , ۰. , ۰.۱ , ۰. , ۰. , ۰.۱ , ۰. , ۰. , ۰.],
[۰. , ۰. , ۰. , ۰. , ۰. , ۰. , ۰. , ۰.۱ , ۰.۰۵, ۰.],
[۰. , ۰. , ۰.۱ , ۰. , ۰. , ۰. , ۰. , ۰.۰۵, ۰.۰۵, ۰.],
[۰. , ۰. , ۰. , ۰. , ۰. , ۰. , ۰. , ۰. , ۰. , ۰.۰۵]])

$$P(a) = [۰.۲, ۰.۱, ۰.۲, ۰.۱, ۰.۲, ۰.۲, ۰.۱, ۰.۱, ۰.۲, ۰.۱]$$
$$P(b) = [۰.۲, ۰.۱, ۰.۲, ۰.۱, ۰.۲, ۰.۲, ۰.۱, ۰.۱, ۰.۲, ۰.۱]$$

$$MI(a, b) = \sum_i \sum_j p(a_i, b_j) log \left[\frac{p(a_i, b_j)}{\left(p(a_i) p(b_j)\right)} \right] = ۱.۷۴۶۷$$

ب)

- Cosine similarity: بر اساس شباهت کسینوس، با توجه به اینکه شباهت متوسط تا زیاد بین G_1 و G_2 وجود دارد. نشان می‌دهد که این دو ژنوم مستقل از یکدیگر نیستند، زیرا تا حدی شباهت دارند.
- Correlation: با توجه به اینکه بازه correlation بین ۱و -۱ می باشد و correlation محاسبه شده برای این دو ژن به نسب زیاد است میتوان گفت این دو ژن با یکدیگر ارتباط دارند و تا حدی وابسته اند. البته به طور کلی، در حالی که همبستگی می تواند اطلاعات مفیدی در مورد رابطه بین دو متغیر ارائه دهد، برای ایجاد یک رابطه علی کافی نیست. تجزیه و تحلیل و تفسیر بیشتر برای تعیین ماهیت و قدرت رابطه بین دو متغیر ضروری است.
- Mutual Information: مقدار ۱.۷۴۶۷ نسبتاً بالا است که نشان می دهد وابستگی متوسط تا قوی بین دو بردار وجود دارد. با این حال، توجه به این نکته مهم است که این فقط یک تخمین بر اساس داده‌های موجود است و قدرت وابستگی می‌تواند با نمونه‌های مختلف یا در زمینه‌های مختلف تغییر کند. بنابراین، تحلیل و تفسیر اضافی برای درک کامل ماهیت و اهمیت رابطه بین دو بردار ضروری است.

ج) اطلاعات متقابل، شباهت کسینوس و همبستگی می تواند نتایج متفاوتی در مورد وابستگی بین دو بردار بدهد. زیرا:

- اطلاعات متقابل میزان اطلاعاتی را که یک بردار در مورد بردار دیگر ارائه می دهد اندازه گیری می کند. این می تواند هر نوع وابستگی، از جمله روابط غیر خطی را جذب کند. بر خلاف همبستگی، رابطه خطی بین دو بردار را در نظر نمی گیرد. بنابراین، اگر رابطه بین دو بردار غیر خطی باشد، می تواند نتایج متفاوتی نسبت به همبستگی بدهد.

- تشابه کسینوس، کسینوس زاویه بین دو بردار را اندازه گیری می کند. از ۱- تا ۱ متغیر است، که ۱ نشان می دهد که بردارها در یک جهت هستند، -۱ نشان می دهد که آنها در جهت مخالف هستند و ۰ نشان می دهد که آنها متعامد هستند. شباهت کسینوس اندازه بردارها را در نظر نمی گیرد، فقط جهت آنها را در نظر می گیرد. بنابراین، اگر بزرگی بردارها مهم باشد، می تواند نتایج متفاوتی نسبت به اطلاعات متقابل و همبستگی بدهد.

- همبستگی رابطه خطی بین دو بردار را اندازه گیری می کند. از ۱- تا ۱ متغیر است که ۱ نشان دهنده رابطه خطی مثبت کامل، ۰ نشان دهنده عدم وجود رابطه خطی و -۱ نشان دهنده رابطه خطی منفی کامل است. همبستگی فرض می کند که رابطه بین دو بردار خطی است. بنابراین، اگر رابطه بین دو بردار غیر خطی باشد، می تواند نتایج متفاوتی نسبت به اطلاعات متقابل بدهد.

به طور خلاصه، اطلاعات متقابل، شباهت کسینوس و همبستگی می توانند نتایج متفاوتی در مورد وابستگی بین دو بردار به دست دهند، زیرا آنها جنبه های مختلف رابطه بین بردارها را اندازه گیری می کنند. انتخاب این که از کدام معیار استفاده شود به ماهیت داده ها و سؤال خاص تحقیق بستگی دارد. البته که در این مثال خاص نتایج یکسان شده اند.

سوال ۶)

- Data aggregation در داده کاوی فرآیند یافتن، جمع آوری و ارائه داده ها در قالب خلاصه شده برای انجام تجزیه و تحلیل آماری طرح های تجاری یا تجزیه و تحلیل الگوهای انسانی است. هنگامی که داده های متعدد از مجموعه داده های مختلف جمع آوری می شود، جمع آوری داده های دقیق برای ارائه نتایج قابل توجه بسیار مهم است. تجمیع داده‌ها می‌تواند به تصمیم‌گیری محتاطانه در بازاریابی، مالی، قیمت‌گذاری محصول و غیره کمک کند. گروه‌های داده انباشته با استفاده از خلاصه‌های آماری جایگزین می‌شوند. داده های انبوه موجود در انبار داده می تواند به حل مسائل منطقی کمک کند که به نوبه خود می تواند فشار زمانی را در حل پرس و جو از مجموعه داده ها کاهش دهد.
- نمونه گیری داده ها یک تکنیک تجزیه و تحلیل آماری است که برای انتخاب، دستکاری و تجزیه و تحلیل زیرمجموعه ای نماینده از نقاط داده برای شناسایی الگوها و روندها در مجموعه داده های بزرگتر مورد بررسی استفاده می شود.

سوال ۷)

الف)

- انتخاب ویژگی: انتخاب ویژگی فرآیند کاهش تعداد متغیرهای ورودی هنگام توسعه یک مدل پیش‌بینی است. کاهش تعداد متغیرهای ورودی برای کاهش هزینه محاسباتی مدل سازی و در برخی موارد برای بهبود عملکرد مدل مطلوب است.

- استخراج ویژگی: استخراج ویژگی فرآیند استخراج خودکار ویژگی های معنی دار از داده های خام با استفاده از الگوریتم های یادگیری ماشین است. این روش اغلب در هنگام برخورد با داده های بدون ساختار، مانند تصاویر یا متن، که در آن داده های خام به شکلی نیستند که در حالت یادگیری ماشینی به راحتی قابل استفاده باشد، استفاده می شود.. استخراج ویژگی را می توان با استفاده از تکنیک هایی مانند تجزیه و تحلیل مؤلفه اصلی (PCA)، تجزیه و تحلیل تشخیص خطی (LDA) و t-SNE انجام داد.
- مهندسی ویژگی: مهندسی ویژگی یک تکنیک یادگیری ماشینی است که از داده ها برای ایجاد متغیرهای جدیدی که در مجموعه آموزشی نیستند استفاده می کند. می‌تواند ویژگی‌های جدیدی را هم برای یادگیری تحت نظارت و هم برای یادگیری بدون نظارت، با هدف ساده‌سازی و سرعت بخشیدن به تبدیل داده‌ها و در عین حال افزایش دقت مدل ایجاد کند. به عنوان مثال، ایجاد یک ویژگی جدید که نشان دهنده نسبت دو ویژگی دیگر است، یا رمزگذاری متغیرهای طبقه بندی شده با استفاده از one hot encoding.

به طور خلاصه، مهندسی ویژگی شامل ایجاد ویژگی های جدید از داده های موجود، انتخاب ویژگی شامل انتخاب مرتبط ترین ویژگی ها برای یک مدل، و استخراج ویژگی شامل استخراج خودکار ویژگی های معنی دار از داده های خام است.

ب)

- الگوریتم های کاهش بعد، یعنی روش هایی که برای کاهش تعداد ویژگی ها یا متغیرهای موجود در داده ها به کار می روند، به دو دسته الگوریتم های کاهش بعد خطی و غیر خطی تقسیم می شوند. این دو دسته الگوریتم به شکل زیر تفاوت دارند:

  الگوریتم های کاهش بعد خطی: این الگوریتم ها با استفاده از تبدیل خطی، مثل تجزیه ماتریسی (PCA)، عاملیت ماتریسی (FA) و تجزیه ماتریس مثبت-نیمه معین(NMF) ، ویژگی های موجود در داده ها را کاهش می دهند. در این روش ها، تغییر در مقدار هر ویژگی با ترکیب خطی ویژگی های دیگر بیان می شود.

  الگوریتم های کاهش بعد غیر خطی: این الگوریتم ها با استفاده از تبدیل غیر خطی، مثل کرنل PCA و تجزیه ماتریس مثبت-نیمه معین با کرنل(Kernel NMF) ، ویژگی های موجود در داده ها را کاهش می دهند. در این روش ها، تبدیل غیرخطی برای تبدیل داده ها به فضایی بیشتری استفاده می شود تا بتوان از مزایای غیر خطی آن ها بهره گرفت.

در کل، الگوریتم های کاهش بعد خطی و غیر خطی هر دو برای کاهش تعداد ویژگی های موجود در داده ها به کار می روند، اما الگوریتم های کاهش بعد خطی برای داده هایی که دارای ویژگی های خطی هستند مفید هستند، در حالی که الگوریتم های کاهش بعد غیر خطی برای داده هایی که دارای ویژگی های غیر خطی هستند مفید هستند.

- PCA:

۱. استانداردسازی داده‌ها
۲. به دست آوردن بردارهای ویژه و مقدارهای ویژه از ماتریکس کواریانس (Covariance matrix) یا ماتریس همبستگی(Correlation Matrix) ، یا انجام «تجزیه مقدارهای منفرد Singular) Vector Decomposition)
۳. مرتب‌سازی مقدارهای ویژه به ترتیب نزولی و انتخاب k بردار ویژه‌ای که متناظر با K بزرگ‌ترین مقدار ویژه هستند K .تعداد ابعاد زیرفضای ویژگی جدید است.($k \leq d$)
٤. ساخت ماتریکس تصویر W از K بردار ویژه انتخاب شده
٥. تبدیل مجموعه داده اصلی X به وسیله W، برای به دست آوردن زیرفضای Kبُعدی Y

- t-SNE:

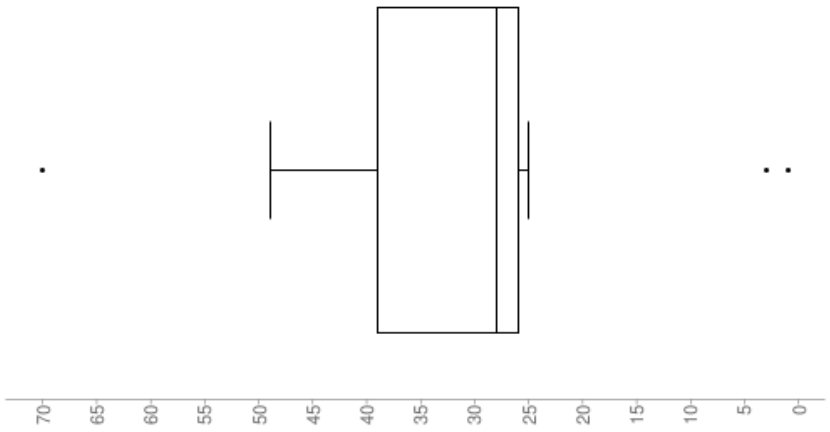
الگوریتم t-SNE شامل دو مرحله به شرح زیر است: مرحله اول بدین صورت است که، یک توزیع احتمال بر روی نقاط در ابعاد بالاتر ایجاد می‌کند به طوری که به اشیاء مشابه احتمال بالاتر و اشیا غیر مشابه احتمال کمتر اختصاص داده می شود.

مرحله دوم بدین صورت است که، همان توزیع احتمال را در ابعاد پایین‌تر به طور مکرر تکرار می‌کند تا زمانی که واگرایی کولبک-لیبلر به حداقل برسد. به بیانی دیگر، واگرایی کولبک-لیبلر معیاری برای اندازه‌گیری تفاوت بین توزیع‌های احتمال مرحله اول و دوم است.

در واقع برای انجام این کار، t-SNE از یک توزیع احتمالاتی برای مدل کردن روابط بین داده ها استفاده می کند. در این روش، ابتدا یک توزیع احتمالاتی چگالی برای داده ها در فضای اصلی محاسبه می شود، سپس توزیع احتمالاتی دیگری برای داده ها در فضای کاهش یافته (معمولاً دو یا سه بعدی) محاسبه می شود. سپس، با کمک گرادیان نزولی، پارامترهای این توزیع احتمالاتی به گونه ای تنظیم می شوند که توزیع احتمالاتی داده ها در فضای کاهش یافته، بهترین توزیعی باشد که داده ها را به خوبی در آن نمایش دهد.

سوال ۸)

Median: ۲۸
Minimum: ۱
Maximum: ۷۰
First quartile: ۲۶
Third quartile: ۳۹
Interquartile Range: ۱۳
Outliers: ۱, ۳, ۷۰



سوال ۹)

الف) Q-Q plot یا quantile-quantile plot یک نمودار است که برای بررسی تطابق توزیع داده های یک مجموعه با توزیع دیگری (مثلاً توزیع نرمال) استفاده می شود. در این روش، داده های مورد بررسی از کمترین تا بیشترین مقدار مرتب می شوند و برای هر یک از مقادیر، مقدار متناظر آن در توزیع دیگری که می خواهیم با آن مقایسه کنیم محاسبه می شود. سپس دو دنباله از این مقادیر در نمودار قرار داده می شوند: دنباله اول مقادیر مرتب شده از داده های واقعی است و دنباله دوم مقادیر متناظر در توزیع مورد نظر (مثلاً توزیع نرمال) است. در صورتی که داده ها با توزیع دیگری کاملاً تطابق داشته باشند، دو دنباله در یک خط قرار می گیرند.

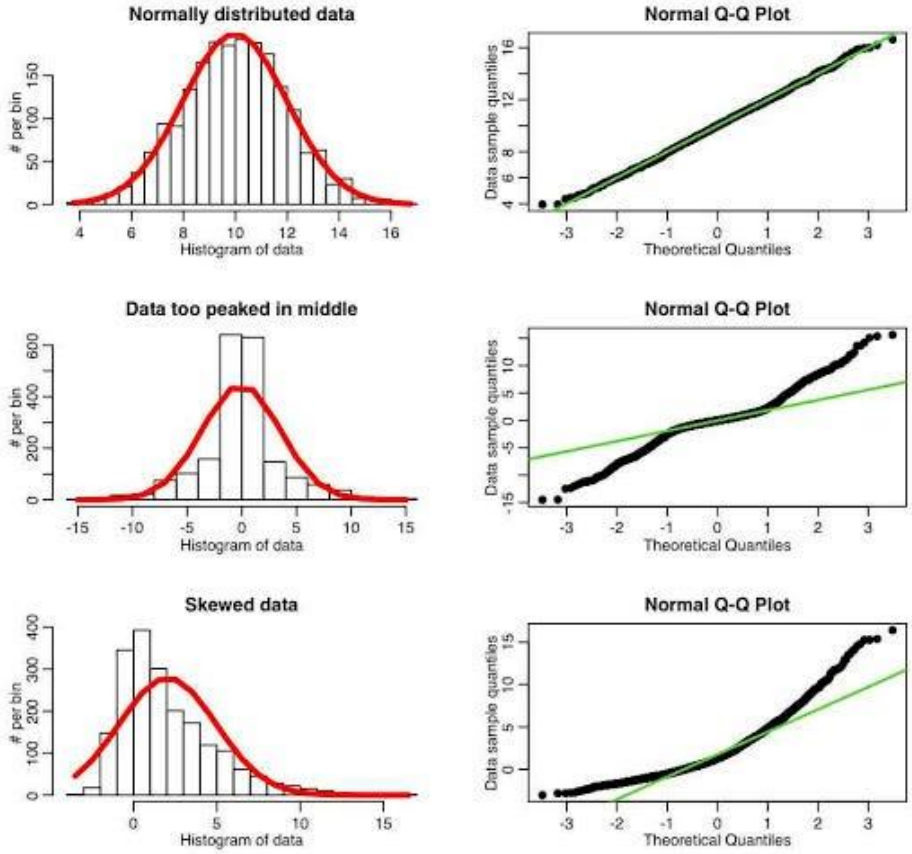
ب) همانطور که در شکل زیر دیده می شود ۳ حالت را بررسی میکنیم:

۱. در این نمودار، نقاط بر روی خط x=y قرار دارند که نماینده توزیع نرمال است. به عبارت دیگر، داده ها با توزیع نرمال کاملاً تطابق دارند.

۲. زمانی که هر دو انتهای نمودار Q-Q انحراف از خط مستقیم دارد و مرکز آن از یک خط مستقیم پیروی می کند، توزیع داده ها به نحوی است که در دو انتها تعداد داده ها کم و اکثر داده ها در میانه قرار گرفته اند.

۳. زمانی که، نقاط به طور کلی در یک خط قرار دارند، اما در برخی نقاط از خط کمی فاصله دارند. این نشان دهنده تطابق ناقص داده ها با توزیع نرمال است.

اگر انتهای پایین نمودار Q-Q از خط مستقیم منحرف شود اما انتهای بالایی منحرف نشود، به وضوح می توانیم بگوییم که توزیع دنباله بلندتری در سمت چپ خود دارد یا به سادگی به سمت چپ انحراف (یا منحنی منفی) است و برعکس.



از روی q-q plot میتوان:

- ارزیابی کنید که آیا نمونه ای از داده ها از توزیع نظری خاصی پیروی می کند یا خیر.
- انحرافات از یک توزیع نظری را شناسایی کنید، که ممکن است نشان دهنده نیاز به روش های آماری جایگزین باشد و درباره skewness داده ها نظر داد.

- تجسم میزان تناسب داده ها با توزیع نظری، که می تواند به محققان کمک کند فرآیند تولید داده های اساسی را درک کنند و تصمیمات آگاهانه ای در مورد تجزیه و تحلیل های آماری بگیرند.

سوال ۱۰)

الف) بازه اعداد: ۰ تا ۱

$$min - max: \frac{value - min}{\max - min}$$

ب) بازه اعداد: محدوده مقادیر پس از نرمال سازی امتیاز z به توزیع داده‌های اصلی بستگی دارد. اگر داده ها از توزیع نرمال پیروی کنند، محدوده معمولاً بین σ ۳- تا ۳σ می باشد.

$$z - score: \frac{value - \mu}{\sigma}$$

ج)

Decimal scaling normalizationیکی از روش های نرمال‌سازی داده‌ها است که در آن اعشار عدد را از محل آن به سمت چپ حرکت داده و تا جایی که لازم است، با صفر پر شده و عدد نهایی به دست می‌آید.
برای اعمال این روش، از فرمول زیر استفاده می‌شود:

$$x_{new} = \frac{x}{۱۰^{j}}$$

در این فرمول، x نشان دهنده عدد اولیه است و j تعداد ارقام اعشاری در x است. با استفاده از این فرمول، داده ها به صورتی نرمال شده و مقادیر آنها در محدوده [۰,۱] قرار می‌گیرند. این روش معمولاً برای داده‌هایی که با مقادیر بزرگ و کوچک زیادی سر و کار دارند و یا واحدهای مختلفی دارند، استفاده می‌شود.
بازه اعداد: محدوده نرمال سازی مقیاس دهی به تعداد ارقام اعشاری مورد استفاده در فرآیند نرمال سازی بستگی دارد. اگر از اعشار j استفاده شود، در صورت مثبت یا منفی بودن مقادیر داده اصلی، محدوده مقادیر نرمال شده بین -۱ و ۱ خواهد بود. با این حال، اگر مقادیر داده های اصلی همه مثبت باشند، محدوده مقادیر نرمال شده بین ۰ و ۱ خواهد بود. شایان ذکر است که محدوده مقادیر نرمال شده ممکن است دقیقاً بین -۱ و ۱ یا ۰ و ۱ نباشد، اما بسته به تعداد ارقام اعشاری استفاده شده در فرآیند عادی سازی، تقریباً در این محدوده خواهد بود.

سوال ۱۱)

الف)

۱۱) الف)

هدف ما پیدا کردن β به گونه‌ای است که S (همانطور که در بالا تعریف شده) مینیمم شود.

$$S = \|y - X\beta\|^2$$

برای این کار یک روش استفاده از مشتق است به این معنی که $\frac{\partial S}{\partial \beta}$ را محاسبه و برابر با صفر قرار دهیم.

S را می‌توان به صورت زیر نوشت:

$$S = (y - X\beta)^T (y - X\beta)$$

اگر تعریف کنیم

$$u = f(\beta) = y - X\beta \quad , \quad g(u) = u^T u$$

خواهیم داشت:

$$\frac{\partial S}{\partial \beta} = \underbrace{\frac{\partial g(u)}{\partial u}}_{2u^T} \cdot \underbrace{\frac{\partial f(\beta)}{\partial \beta}}_{-X} = (2u^T)(-X) = 0$$

$$-2(y - X\beta)^T X = 0$$

$$-2X^T y + 2X^T X\beta = 0$$

$$2X^T y = 2X^T X\beta \Rightarrow \boxed{\beta = (X^T X)^{-1} X^T y}$$

Normal Equation

ب)

$$X^T X B = X^T y$$

$$X = \begin{bmatrix} ۱ & ۲ \\ ۱ & ۴ \\ ۱ & ۱ \\ ۱ & ۳ \\ ۱ & ۲ \\ ۱ & ۶ \end{bmatrix}, y = \begin{bmatrix} ۵ \\ ۶ \\ ۳ \\ ۶ \\ ۳ \\ ۱۰ \end{bmatrix}$$

$$X^T X = \begin{bmatrix} ۱ & ۱ & ۱ & ۱ & ۱ & ۱ \\ ۲ & ۴ & ۱ & ۳ & ۲ & ۶ \end{bmatrix} \begin{bmatrix} ۱ & ۲ \\ ۱ & ۴ \\ ۱ & ۱ \\ ۱ & ۳ \\ ۱ & ۲ \\ ۱ & ۶ \end{bmatrix} = \begin{bmatrix} ۶ & ۱۸ \\ ۱۸ & ۷۰ \end{bmatrix}$$

$$X^T y = \begin{bmatrix} ۱ & ۱ & ۱ & ۱ & ۱ & ۱ \\ ۲ & ۴ & ۱ & ۳ & ۲ & ۶ \end{bmatrix} \begin{bmatrix} ۵ \\ ۶ \\ ۳ \\ ۶ \\ ۳ \\ ۱۰ \end{bmatrix} = \begin{bmatrix} ۳۳ \\ ۱۲۱ \end{bmatrix}$$

$$B = \frac{۱}{-۴۴} \begin{bmatrix} ۷۰ & -۱۸ \\ -۱۸ & ۶ \end{bmatrix} \begin{bmatrix} ۳۳ \\ ۱۲۱ \end{bmatrix} = \begin{bmatrix} ۱.۳۷۵ \\ ۱.۳۷۵ \end{bmatrix}$$

$$e = y - XB = \begin{bmatrix} ۰.۸۷۵ \\ -۰.۸۷۵ \\ ۰.۲۵ \\ ۰.۵ \\ -۱.۱۲۵ \\ ۰.۳۷۵ \end{bmatrix}$$

ج)

$$X^T X B = X^T y$$

$$X = \begin{bmatrix} ۱ & ۲ & ۴ \\ ۱ & ۴ & ۱۶ \\ ۱ & ۱ & ۱ \\ ۱ & ۳ & ۹ \\ ۱ & ۲ & ۴ \\ ۱ & ۶ & ۳۶ \end{bmatrix}, y = \begin{bmatrix} ۵ \\ ۶ \\ ۳ \\ ۶ \\ ۳ \\ ۱۰ \end{bmatrix}$$

$$X^T X = \begin{bmatrix} ۶ & ۱۸ & ۷۰ \\ ۱۸ & ۷۰ & ۳۲۴ \\ ۷۰ & ۳۲۴ & ۱۵۰۶ \end{bmatrix}$$

$$X^T y = \begin{bmatrix} ۳۳ \\ ۱۲۱ \\ ۵۴۵ \end{bmatrix}$$

$$B = \begin{bmatrix} ۲.۲۲۵۸۴ \\ ۰.۷۵۰۵۶ \\ ۰.۰۸۷۶۴ \end{bmatrix}$$

$$e = y - XB = \begin{bmatrix} ۰.۹۲۲۴۸ \\ -۰.۶۳۰۳۲ \\ -۰.۶۴۰۴. \\ ۰.۷۳۳۷۲ \\ -۱.۰۷۷۵۲ \\ ۰.۱۱۵۷۶ \end{bmatrix}$$

سوالات بخش عملی (در نوت بوک هم آمده):

Q۵)

داده‌افزایی (افزایش داده)، در داده‌کاوی، به سازوکارهایی برای افزایش شمارِ داده‌ها گفته می‌شود. داده‌های تازه با ساختن رونوشت‌ها و نمونه‌هایی از داده‌های کنونی (دردسترس)، یا ساخت داده‌هایی با الگوگرفتن از داده‌های کنونی (کمی ناهمسان)، ساخته می‌شوند. شیوه‌ی داده‌افزایی را می‌توان یک رگولارایزر (همترازکننده) دانست، که راه‌کاری است برای چالشِ بیش‌برازش در زمان آموزش یک الگوی (مدل) یادگیری ماشین.

-افزایش داده های صوتی

تزریق نویز: برای بهبود عملکرد مدل، نویز گاوسی یا تصادفی را به مجموعه داده صوتی اضافه کنید.

جابجایی: انتقال صدا به چپ (سریع به جلو) یا راست با ثانیه های تصادفی.

تغییر سرعت: سری های زمانی را با نرخ ثابتی افزایش می دهد.

تغییر زیر و بم: به طور تصادفی زیر و بم صدا را تغییر دهید.

-افزایش داده های متنی

تغییر کلمه یا جمله: تغییر تصادفی موقعیت یک کلمه یا جمله.

جایگزینی کلمه: جایگزین کلمات با مترادف.

دستکاری درخت نحو: جمله را با استفاده از همان کلمه بازنویسی کنید.

درج کلمه تصادفی: کلمات را به صورت تصادفی درج می کند

حذف تصادفی کلمه: کلمات را به صورت تصادفی حذف می کند.

-تقویت تصوی

تغییرات هندسی: به طور تصادفی ورق زدن، برش، چرخش، کشش و بزرگنمایی تصاویر. باید مراقب اعمال چندین تغییر شکل روی تصاویر مشابه باشید، زیرا این کار می تواند عملکرد مدل را کاهش دهد.

تغییر فضای رنگ: به طور تصادفی کانال های رنگ، کنتراست و روشنایی RGB را تغییر دهید

فیلترهای هسته: به طور تصادفی وضوح یا تاری تصویر را تغییر می دهند

پاک کردن تصادفی: بخشی از تصویر اولیه را حذف کنید.

اختلاط تصاویر: ترکیب و ترکیب چندین تصویر.

تقویت داده ها فقط در مجموعه آموزشی انجام می شود زیرا به تعمیم و استحکام مدل کمک می کند. بنابراین هیچ فایده ای برای افزایش مجموعه تست وجود ندارد.

(Q٦

- آپ سمپلسنگ رویه ای است که در آن نقاط داده تولید شده مصنوعی (مرتبط با کلاس اقلیت) به مجموعه داده تزریق می شود. پس از این فرآیند، تعداد هر دو برچسب تقریباً یکسان است. این روش یکسان سازی از تمایل مدل به سمت طبقه اکثریت جلوگیری می کند. علاوه بر این، تعامل (خط مرزی) بین کلاس های هدف تغییر می کند. و همچنین، مکانیسم upsampling به دلیل اطلاعات اضافی، سوگیری را به سیستم وارد می کند.
- پایین‌نمونه‌سازی مکانیزمی است که تعداد نمونه‌های آموزشی را که در طبقه اکثریت قرار می‌گیرند کاهش می‌دهد. زیرا به افزایش تعداد دسته‌های هدف کمک می‌کند. با حذف داده‌های جمع‌آوری‌شده، اطلاعات ارزشمند زیادی را از دست می‌دهیم.

(Q٧

- SMOTETomek

این روش ترکیبی از توانایی SMOTE برای تولید داده‌های مصنوعی برای کلاس اقلیت و توانایی Tomek Linksبرای حذف داده‌هایی است که به عنوان پیوندهای Tomek این روش روشی برای انجام down sampling, up samplingمی باشد. از کلاس اکثریت (یعنی نمونه‌هایی از داده‌های کلاس اکثریت نزدیک‌ترین به داده‌های کلاس اقلیت هستند)). فرآیند پیوندهای SMOTE-Tomek به شرح زیر است.

١ -شروع SMOTE داده های تصادفی را از کلاس اقلیت انتخاب کنید.

٢ -فاصله بین داده های تصادفی و k نزدیکترین همسایه آن را محاسبه کنید.

۳ -اختلاف را با یک عدد تصادفی بین ۰ و ۱ ضرب کنید سپس نتیجه را به عنوان نمونه مصنوعی به کلاس اقلیت اضافه کنید.

۳ -مرحله شماره ۲-۳ را تکرار کنید تا نسبت مورد نظر طبقه اقلیت برآورده شود) .پایان(SMOTE

۴) -شروع پیوندهای (Tomek داده های تصادفی را از کلاس اکثریت انتخاب کنید.

۵- اگر نزدیکترین همسایه داده های تصادفی، داده های کلاس اقلیت است (یعنی ایجاد پیوند Tomek)، سپس پیوند Tomek را حذف کنید.

- Smoteenn

این روش ترکیبی از توانایی SMOTE برای تولید نمونه‌های مصنوعی برای کلاس اقلیت و توانایی ENN برای حذف برخی مشاهدات از هر دو کلاس است که دارای کلاس متفاوتی بین کلاس مشاهده و کلاس اکثریت-K نزدیک‌ترین همسایه آن هستند. فرآیند SMOTE-ENN را می توان به صورت زیر توضیح داد.

۱) -شروع (SMOTE داده های تصادفی را از کلاس اقلیت انتخاب کنید.

۲ -فاصله بین داده های تصادفی و k نزدیکترین همسایه آن را محاسبه کنید.

۳ -اختلاف را با یک عدد تصادفی بین ۰ و ۱ ضرب کنید سپس نتیجه را به عنوان نمونه مصنوعی به کلاس اقلیت اضافه کنید.

۴ -مرحله شماره ۲-۳ را تکرار کنید تا نسبت مورد نظر طبقه اقلیت برآورده شود) .پایان(SMOTE

۵) -شروع K (ENN را به عنوان تعداد نزدیکترین همسایگان تعیین کنید. اگر تعیین نشد، =۳K.

۶-K -نزدیک ترین همسایه مشاهده را از بین مشاهدات دیگر در مجموعه داده پیدا کنید، سپس کلاس اکثریت را از K-nearest همسایه برگردانید.

۷ -اگر کلاس مشاهده و کلاس اکثریت از-K نزدیک ترین همسایه مشاهده متفاوت باشد، مشاهده و-K نزدیک ترین همسایه آن از مجموعه داده حذف می شود.

۸- مراحل ۲ و ۳ را تکرار کنید تا نسبت مورد نظر هر کلاس برآورده شود. (پایان ENN)

- در روش ترکیبی افزایش اندازه و کاهش اندازه معمولا به این صورت انجام می‌شود که ابتدا سیگنال ورودی با نرخ نمونه‌برداری بالا به نرخ نمونه‌برداری کمتری کاهش داده می‌شود، سپس پردازشی روی سیگنال کاهش یافته اعمال می‌شود و در نهایت سیگنال پردازش شده با نرخ نمونه‌برداری اولیه دوباره بزرگ شده و بازیابی می‌شود.

- دو روش SMOTETomek و SMOTEENN هر دو از ترکیب دو روش oversampling و undersampling برای مقابله با مشکل ایجاد تعادل در داده‌های دوتایی استفاده می‌کنند. در هر دو روش، ابتدا روش undersampling برای حذف نمونه‌های اکثریتی (مثلا نمونه‌های اقلیتی را حفظ می‌کنند و نمونه‌های اکثریتی را حذف می‌کنند) اعمال می‌شود و سپس با استفاده از روش oversampling (مثلا SMOTE)، نمونه‌های اقلیتی جدیدی ایجاد می‌شوند.