



دانشگاه صنعتی امیرکبیر

(پلیتکنیک تهران)

دانشکده مهندسی کامپیوتر

درس داده کاوی

تمرین اول ۱

امیرمهدی زرین نژاد

۹۷۳۱۰۸۷

نیمسال دوم ۱۴۰۰

بخش نوشتاری

سوال اول(۱) مفاهیم زیر را تعریف کنید.

۱-Dimension: یا ابعاد داده خصیصه‌هایی (ویژگی-variable-attribute-feature) از داده است که برای شناسایی، تحلیل و ... در نظر می‌گیریم و بررسی می‌کنیم. درواقع یک بُعد، دیتا فیلدی است که یک خصیصه (فیچر) از دیتا آبجکت را بیان می‌کند. و می‌تواند به شکل‌های مختلفی ظاهر شود از جمله اسمی (Nominal)، دو دویی (Binary)، ترتیبی و عددی. هم‌چنین مجموعه مقادیرش می‌تواند گسته باشد.

۲-Outlier: داده‌ی پرت درواقع داده‌ای (Data Object) است که همانند رفتار عمومی و معمول داده‌ها عمل نمی‌کند و تفاوت قابل توجهی با دیگر داده‌های مجموعه داده دارد. پس از دیگر داده‌ها فاصله دارد و نمی‌توان آن را به هیچ گروهی از داده‌ها نسبت داد و اگر در گروهی هم قرار بگیرد از بقیه داده‌های گروه دور است. این داده‌ها گاه‌ها هدف ما هستند و از آن‌ها در تحلیل‌هایمان استفاده می‌کنیم و گاهی هم نیازی نداریم و حذف‌شان می‌کنیم.

۳-Independent variable: متغیر مستقل متغیری است که محاسبه مقدارش نیازمند محاسبه هیچ متغیر دیگری نیست و اصطلاحاً مستقل از هر متغیر دیگری است. پس از ترکیب خطی متغیر(ها) دیگر محاسبه نمی‌شود.

۴-Dependent variable: متغیر وابسته برای محاسبه‌اش به متغیر(ها) دیگری نیاز است و به تنها‌ی محاسبه نمی‌شود. پس از ترکیب خطی متغیرهای دیگر به دست می‌آید.

۵-Stratified Sampling: یک روش نمونه‌برداری است که در آن ابتدا داده‌ها را به زیرگروه‌های (طبقه-پارتیشن-stratum) یک‌جور تقسیم می‌کنیم (یا خوش‌بندی می‌کنیم) و سپس از هر زیرگروه نمونه برداری (simple random sampling) می‌کنیم. در این روش هر عنصر (داده) باید فقط و فقط در یک زیرگروه قرار بگیرد و مجموع زیرگروه‌ها باید برابر کل مجموعه داده شود. با پیاده‌سازی این روش توازن بین پارتیشن‌های مختلف برقرار می‌شود و نمونه‌برداری بین زیرگروه‌ها پخش می‌شود.

سوال دوم(۲) پیش‌پردازش داده‌ها از جمله موارد پراهمیت در انجام پروژه‌های مبتنی بر یادگیری است و نرمال‌سازی داده‌ها یکی از مهم‌ترین مراحل پیش‌پردازش است. سه مورد از روش‌های نرمال‌سازی را با ذکر مثال توضیح دهید. سپس محدوده‌ی نرمال‌سازی هر یک را مشخص کنید.

:Min-max normalization

در این روش مقادیر را به بازه‌ای جدید $[new_min, new_max]$ و دلخواه می‌بریم و با این کار محدوده داده‌هایمان عوض می‌شود (محدوده جدید = $[new_min, new_max]$) اما تناسب قرارگیریشان تغییری نمی‌کند. این کار را با فرمول زیر انجام می‌دهیم:

V_{new} = مقدار داده، V = مقدار جدید داده پس از نرمال سازی، \min و \max = کمینه و بیشنهای مجموعه داده، new_max و new_min = کمینه و بیشنهای بازه‌ی جدید موردنظر برای نرمال سازی

$$V_{new} = \frac{(V - \min)}{(\max - \min)} (new_max - new_min) + new_min$$

یک کاربرد رایج آن نرمال سازی از طریق بردن به بازه‌ی $[0, 1]$ است:
مثالاً عدد ۳۷ را در بازه‌ی ۳۱ تا ۹۱ به بازه‌ی ۰ تا ۱ ببریم.

$$V_{new} = \frac{(31-31)}{(91-31)} (1 - 0) + 0 = 0 \quad \text{برای ۳۱}$$

$$V_{new} = \frac{(37-31)}{(91-31)} (1 - 0) + 0 = 0.1 \quad \text{برای ۳۷}$$

$$V_{new} = \frac{(91-31)}{(91-31)} (1 - 0) + 0 = 1 \quad \text{برای ۹۱}$$

Z-score normalization

در این روش نرمال سازی را با محاسبه و استفاده‌ی فاصله‌ی داده‌ها از میانگین انجام می‌دهیم. مقدار مطلق z , فاصله بین مقدار آن داده (v) و میانگین جمعیت را در همان یکای انحراف معیار نشان می‌دهد. وقتی نمره خام زیر میانگین باشد، z منفی است و وقتی بالاتر باشد، مثبت است.

مقدار اولیه داده v ، انحراف معیار جمعیت σ ، میانگین جمعیت μ در واقع z-score مضری از انحراف معیار است که مقدار آن نشان می‌دهد مقدار یک داده چقدر بیشتر یا کمتر از میانگین است. مقادیر داده‌ی بالاتر از میانگین دارای نمره استاندارد مثبت و مقادیر پایین‌تر از میانگین دارای نمره استاندارد منفی هستند.

محدوده جدید: $[\mu - 3\sigma, \mu + 3\sigma]$

مثال:

برای داده‌های ۳۱, ۳۳, ۴۶, ۷۸, ۹۱

$$\mu = 55.8, \quad \sigma = 24.34$$

$$Z\text{-score}(31) = -1.01879, \quad Z\text{-score}(33) = -0.93663, \quad Z\text{-score}(46) = -0.40259,$$

$$Z\text{-score}(78) = 0.91198, \quad Z\text{-score}(91) = 1.44603$$

:Decimal scaling

در این روش بزرگ‌ترین عدد از نظر اندازه در میان داده‌ها را پیدا می‌کنیم، تعداد ارقامش را می‌شماریم و ۱۰ را به توان تعداد ارقامش می‌رسانیم. نهایتاً همه اعداد را بر این توان از ۱۰ تقسیم می‌کنیم:

$$V_{new} = \frac{v}{10^j},$$

j = کوچک‌ترین عدد ممکن که اگر آن را در فرمول قرار دهیم، V_{new} برای بزرگ‌ترین (از نظر اندازه مطلق و بدون علامت) داده، از ۱ کمتر شود.

برای اعداد $-10, -401, 301, 201, 501, 601$:

$$V_{new}(-10) = -0.01, V_{new}(201) = 0.201, V_{new}(301) = 0.301, V_{new}(-401) = -0.401,$$

$$V_{new}(501) = 0.501, V_{new}(601) = 0.601, V_{new}(701) = 0.701$$

محدوده‌ی جدید: $\left[\frac{\min}{10^j}, \frac{\max}{10^j} \right]$ و از آنجایی که زبرابر با تعداد ارقام بزرگ‌ترین عدد است؛ تعریف این بازه را می‌توان به $[-1, 1]$ محدود کرد.

سوال سوم (۳) تکنیک ChiMerge یک الگوریتم خودکار گسته‌سازی تحت ناظارت، مبتنی بر ادغام از پایین به بالا است که با استفاده از آماره‌ی χ^2 کار خود را انجام می‌دهد. فواصل مجاور با حداقل مقادیر χ^2 با هم ادغام می‌شوند تا زمانی که معیار توقف انتخاب شده برآورده شود. به طور خلاصه نحوه عملکرد ChiMerge را شرح دهید.

این تکنیک، یک الگوریتم گسته‌سازی است که با استفاده از آماره‌ی χ^2 همراه با ادغام (merge) از پایین به بالا پیاده سازی می‌شود. و بیان می‌کند که دو ویژگی مورد بررسی وابستگی دارند یا خیر.

به این صورت پیاده می‌شود که یکسری تقسیمات بازه‌های تعریف می‌کنیم. سپس این بازه‌ها را به بازه‌هایی کوچک‌تر تبدیل می‌کنیم. و سپس میزان شباهت این بازه‌ها را با استفاده از آماره‌ی χ^2 محاسبه می‌کنیم. نهایتاً جفت‌های مجاور که کمترین مقدار χ^2 را دارند با یکدیگر ادغام می‌کنیم. این روند ادغام را تا جایی ادامه می‌دهیم که یک معیار از پیش تعریف شده برآورده شود و دیگر ادامه نمی‌دهیم (شرط پایان که می‌تواند یک مقدار آستانه‌ای برای χ^2 باشد).

نحوه محاسبه آماره χ^2 :

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

سوال چهارم (۴) برای بردارهای داده شده، موارد خواسته شده را بدست آورید.

$$x = [1, 1, 1, 1], y = [2, 2, 2, 2]$$

Cosine similarity, Correlation, Euclidean distance.

$$x = [0, 1, 0, 1], y = [1, 0, 1, 0]$$

Cosine similarity, Correlation, Euclidean distance, Jaccard distance.

$$x = [1, 1, 0, 1, 0, 1], y = [1, 1, 1, 0, 0, 1]$$

Correlation, Manhattan distance, [Bhattacharya distance](#).

$$x = [2, -1, 0, 2, 0, -3], y = [-1, 1, -1, 0, 0, -1]$$

Cosine similarity, Correlation.

$$\text{iii) } x = [1, 1, 1, 1] \quad , \quad y = [1, 1, 1, 1]$$

$\text{Corr} = ?$

$$\text{Cosine Sim} \Rightarrow \frac{(x, y)}{\|x\| \cdot \|y\|} = \frac{x_1 + x_2 + x_3 + x_4}{\sqrt{1^2 + 1^2 + 1^2 + 1^2} + \sqrt{1^2 + 1^2 + 1^2 + 1^2}} = \frac{4}{4\sqrt{2}} = \frac{1}{\sqrt{2}}$$

$$\text{Cov} = \frac{\text{Covariance}(x, y)}{\text{std-deviation}(x) \times \text{std-deviation}(y)} = \frac{s_{xy}}{s_x \times s_y}$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad \bar{x} = \frac{1+1+1+1}{4} = \frac{4}{4} = 1 \Rightarrow s_x = 0$$

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1+1+1+1}{4} = \frac{4}{4} = 1 \Rightarrow s_y = 0$$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad \bar{x} = 1, \bar{y} = 1 \Rightarrow s_{xy} = \frac{1}{4} \cdot 4 = 1 \quad \text{Correlation}(x, y) = \frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}}$$

$$\text{Euclidean Dist} \Rightarrow \sqrt{\sum_{i=1}^n (x_i - y_i)^2} = \sqrt{(1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2} = 0$$

$$\Rightarrow x = [0, 1, 0, 1], \quad y = [1, 0, 1, 0]$$

$$\text{Cos Sim} = \frac{0+1+0+1}{\sqrt{2} + \sqrt{2}} = \frac{2}{2\sqrt{2}} = \frac{1}{\sqrt{2}}$$

$$\text{Correlation} = \frac{s_{xy}}{s_x \times s_y} = \frac{-\frac{1}{\sqrt{2}}}{\sqrt{\frac{1}{2}} \times \sqrt{\frac{1}{2}}} = \frac{-1}{\sqrt{2}}$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad \bar{x} = \frac{1}{4} \Rightarrow s_x = \sqrt{\frac{1}{2} \times \left(\frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right)} = \sqrt{\frac{1}{2}}$$

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{4} \Rightarrow s_y = \sqrt{\frac{1}{2}}$$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{4} \left(-\frac{1}{4} \times \frac{1}{4} + \left(\frac{1}{4} \times -\frac{1}{4} \right) + \left(-\frac{1}{4} \times \frac{1}{4} \right) + \left(\frac{1}{4} \times -\frac{1}{4} \right) \right) = -\frac{1}{8}$$

$$\text{Euclidean Dist} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} = \sqrt{2} = \sqrt{1+1+1+1}$$

$$\text{Jaccard Dist} = \frac{0}{4+1+1} = \frac{0}{6} = 0$$

$$2) \quad x = [1, 1, 0, 1, 0, 1], \quad y = [1, 1, 1, 0, 0, 1]$$

$$\text{Correlation} \Rightarrow \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{10}}{\frac{1}{\sqrt{10}} \times \frac{1}{\sqrt{10}}} = \underline{\underline{\frac{1}{10}}}$$

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad \bar{x} = \frac{1}{6} \Rightarrow S_x = \sqrt{\frac{1}{6} \left(\frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} \right)} = \sqrt{\frac{1}{6} \times \frac{6}{9}} = \underline{\underline{\frac{1}{\sqrt{6}}}}$$

$$S_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{1}{\sqrt{10}}, \quad \bar{y} = \frac{1}{6}$$

$$S_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{10} \times \left(\left(\frac{1}{6} \times \frac{1}{6} \right) + \left(\frac{1}{6} \times \frac{1}{6} \right) \right) = \underline{\underline{\frac{1}{10}}}$$

$$\text{Manhattan Dist} \Rightarrow \sum_{i=1}^n |x_i - y_i| = |1| + |-1| + |0| + |1| + |0| + |1| = \underline{\underline{4}}$$

$$\text{Bhattacharya Dist} \Rightarrow D_B(x, y) = -\ln(B_C(x, y))$$

$$B_C(x, y) = \sum_{i=1}^n \sqrt{x_i y_i} = \frac{1}{2} = \sqrt{1} + \sqrt{1} + \sqrt{0} + \sqrt{0} + \sqrt{0} + \sqrt{1}$$

$$\Rightarrow D_B(x, y) = -\ln(\omega) = \underline{\underline{1.98}}$$

$$2) \quad x = [1, -1, 0, 1, 0, -1], \quad y = [-1, 1, -1, 0, 0, -1]$$

$$\text{Cosine Sim} \Rightarrow \frac{(x \cdot y)}{\|x\| \|y\|} = \frac{-1 - 1 + 0 + 1 + 0 - 1}{\sqrt{10} \times \sqrt{10}} = \underline{\underline{0}}$$

$$\text{Correlation} \Rightarrow \frac{S_{xy}}{S_x S_y} \quad \leftarrow S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad S_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad \bar{x} = 0, \quad \bar{y} = -\frac{1}{4}$$

$$\Rightarrow S_x = \sqrt{\frac{1}{6} \times 10} = \sqrt{\frac{10}{6}}, \quad S_y = \sqrt{\frac{1}{6}}, \quad S_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{6} \left(\left(\frac{1 \times -1}{6} \right) + \left(-1 \times \frac{1}{6} \right) + \left(0 \times -\frac{1}{4} \right) + \left(1 \times \frac{1}{6} \right) + \left(0 \times -\frac{1}{4} \right) + \left(-1 \times -\frac{1}{4} \right) \right) = 0$$

$$\Rightarrow \text{Corr} = \frac{0}{\frac{1}{\sqrt{6}} \times \frac{1}{\sqrt{6}}} = \underline{\underline{0}}$$

سوال پنجم(۵) کاهش داده یکی از عملیات‌های اصلی در پیش‌پردازش داده در داده‌کاوی به‌شمار می‌رود. هدف از انجام تکنیک کاهش داده چیست؟ راهبردهای آن را توضیح دهید.

در داده‌کاوی به هنگام جمع‌آوری داده لازم است تا جای ممکن آن‌ها را جمع کنیم چراکه فرصتی است که شاید تکرار نشود و یا تکرارش هزینه‌بر باشد پس با حجم بالایی از دیتا سروکار داریم. حال کار با این حجم از داده و تحلیلشان می‌تواند بسیار زمان بر و پرهزینه باشد و یا اصلاً نیاز نباشد. پس خیلی اوقات لازم است داده‌ها را کاهش دهیم به طوری که تحلیل‌های ما را تغییر ندهد و برروی قسمتی از داده‌ها تمرکز کنیم. با تکنیک‌های data reduction می‌توانیم این نیاز را برآورده کنیم و با توجه به مورد استفاده و کاربرد، تعداد رکوردهای داده را کاهش دهیم و یا خلاصه‌ی مناسبی از داده‌ها تامین کنیم.

چند راهبرد برای این پیاده‌سازی این تکنیک تعریف شده:

:که در آن ابعاد داده (فیچرهای تعریف شده) را با توجه به کاربردمان کاهش می‌دهیم و موارد کم اهمیت‌تر را حذف می‌کنیم. با این کار حجم داده و پردازش‌هایمان سبک می‌شود اما بر تحلیل‌هایمان اثر نامطلوبی نمی‌گذارد. این راهبرد خود به روش‌های مختلفی می‌تواند پیاده‌سازی شود از جمله Wavelet transform

:Rahberd دیگری است که برای کاهش داده استفاده می‌شود و از طریق جایگزین کردن داده‌های حجیم با فرم آلترنیت و کوچک‌ترشان به این هدف می‌رسد. این راهبرد به دو گروه non-parametric و parametric تقسیم می‌شود. پارامتریک که در آن با استفاده از داده‌ها مدل ساخته می‌شود و به جای داده استفاده می‌شود و از روش‌هایی مانند رگرسیون استفاده می‌کند. غیر پارامتریک که در آن مدل ایجاد نمی‌کنیم و از روش‌هایی مانند خوشبندی، نمونه‌برداری و ... استفاده می‌کنیم.

:که به معنای فشرده‌سازی داده‌ها است و از الگوریتم‌های فشرده سازی برای کاهش حجم داده استفاده می‌شود. در واقع عملکرد اصلی فشرده‌سازی داده از طریق کاهش تعداد بیت‌های لازم برای بیان‌کردن داده است و از طریق روش‌هایی مانند تغییر کدگذاری، تبدیل داده و تغییراتی از این قبیل عملی می‌شود.

سوال ششم(۶) کاهش بعد یکی از تکنیک‌های رایج در داده‌کاوی است و روش‌های گوناگونی برای آن وجود دارد. تبدیل موجک یکی از تکنیک‌هایی است که برای راهبرد کاهش بعد انجام می‌گیرد. آن را به اختصار توضیح دهید. در ادامه تفاوت feature extraction و feature selection را بیان کنید.

تبدیل موجک یا Wavelet transform یک تکنیک پردازش سیگنال خطی است که وقتی روی بردار یک مجموعه داده اعمال شود، آن را به یک بردار عددی جدید از ضرایب موجک تبدیل می‌کند که طول این دو بردار برابر است. این روش را می‌توان برای داده‌های چند بعدی استفاده کرد. به این صورت که ابتدا تبدیل به بعد اول و سپس به بعد دوم و الی آخر انجام شود. این تبدیلات در دنیای واقعی نیز کاربردهای فراوانی از جمله در تجزیه و تحلیل داده‌های سری زمانی، پاکسازی داده‌ها و فشرده‌سازی تصاویر اثر انگشت.

برای پیدا‌سازی این روش به شکل زیر عمل می‌شود:

طول بردار داده باید توانی از ۲ باشد و اگر نبود با اضافه کردن ۰ به انتهایش آن را به توانی از ۲ تبدیل می‌کنیم.

سپس ۲ عمل(تابع) برای اعمال به داده‌ها داریم. هموارسازی(smoothing) داده‌ها (اعمالی مانند محاسبه میانگین وزن دار داده‌ها و ...). و محاسبه اختلاف وزن دار که ویژگی‌های مهم بردار را استخراج می‌کند.

این دو تابع را به صورت جفت جفت بر داده‌هایمان اعمال می‌کنیم که در نهایت دو مجموعه داده با طول $L/2$ خواهیم داشت که اولی نسخه با فرکانس پایین و دومی نسخه با فرکانس بالای داده‌هایمان خواهد بود.

این دو تابع تا زمانی که بردار داده‌هایمان به طول ۲ برسند، به صورت بازگشته فراخوانده می‌شوند.

نهایتاً وزن‌های موجک به بردار تبدیل شده‌ی داده‌ها اعمال می‌شوند.

در feature selection ویژگی‌های مهم‌تر و با ارزش‌تر را انتخاب می‌شوند و آن‌هایی که بی‌ربط و کم‌اهمیت هستند حذف می‌شوند و با این روش از ابعاد و حجم داده کاسته می‌شود. در feature extraction از چند ویژگی یک ویژگی جدید ایجاد می‌شود که اهمیت بیشتری برای ما خواهد داشت و از طرفی نماینده آن چند ویژگی خواهد بود و از حجم داده‌هایمان کم می‌کند. پس در انتخاب ویژگی، فیچرها بدون اینکه تغییر کنند زیر مجموعه‌ای از کل فیچرها هستند اما در استخراج ویژگی، فیچرهای جدیدی خواهیم داشت که در داده‌های اولیه‌مان وجود نداشتند.

سوال هفتم ۷) اگر داده‌های زیر را به روش box plot نمایش دهیم، \min و \max چه اعدادی خواهد بود؟ همچنین داده‌های پرت را نیز مشخص کنید.

70, 56, 71, 73, 74, 144, 89, 80, 90, 143, 89, 80, 90, 143, 146, 100, 20, 44, 74

ابتدا داده‌ها را به ترتیب صعودی مرتب می‌کنیم (۱۹ تا داده):

20, 44, 56, 70, 71, 73, 74, 80, 80, 89, 89, 90, 90, 100, 143, 143, 144, 146

برای پیدا کردن میانه اگر تعداد داده‌ها فرد باشد عنصر وسط (۱۰امین عنصر) را انتخاب می‌کنیم:

Median = 80

که در box plot برابر Q2 نیز می‌باشد:

Q2 = Median = 80

حال برای محاسبه‌ی Q1 میانه‌ی عناصر پیش از میانه‌ی کل (که می‌شود ۱۰امین عنصر) و برای محاسبه‌ی Q3، میانه‌ی عناصر پس از میانه‌ی کل (که می‌شود ۱۵امین عنصر) را پیدا می‌کنیم:

Q1 = 71

Q3 = 100

در ادامه برای پیدا کردن محدوده‌ی box plot و عناصر outlier، به مقدار IQR و Lower bound و Upper bound نیاز داریم:

IQR = Q3 - Q1 = 100 - 71 = 29

Upper bound = Q3 + 1.5 * IQR = 100 + 43.5 = 143.5

Lower bound = Q1 - 1.5 * IQR = 71 - 43.5 = 27.5

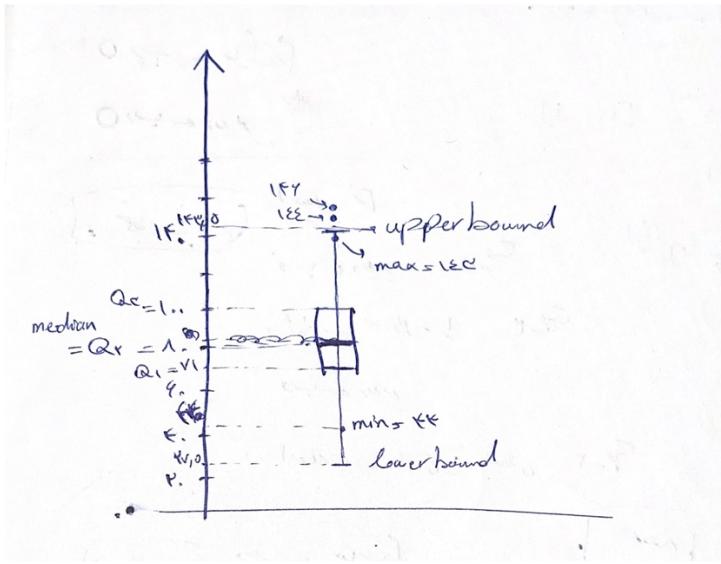
حال \min و \max مربوط به box plot را می‌شوند اولین عنصر بعد از Lower bound و اولین عنصر پیش از Upper bound:

Box plot Min: 44

Box plot Max: 143

(مینیمم و مаксیمم کلی هم که طبق اعداد مرتب شده 20 و 146 هستند)

حال داده‌های پرت را نیز می‌توان پیدا کرد که می‌شوند داده‌های قبل از مینیمم و بعد از مаксیمم باکس پلات: 44 > Outliers > 143 ➔ Outliers: 20, 144, 146



سوال هشتم) در رابطه با noise و outlierها به سوالات زیر پاسخ کامل همراه با توضیح ارائه دهید.

۱- آیا noise همیشه مطلوب هست؟ outlierها چطور؟

خیر نویزها عموماً داده‌های همراه با error هستند. یعنی در آن‌ها خطای رخ داده یا فرمت نادرست دارند یا ... و از حالت مطلوب و مورد استفاده ما خارج شده‌اند و هندل کردن‌شان (تصحیح یا جایگزینی یا ...) هزینه مضاعف برای ما ایجاد می‌کند و خود داده‌ی نویز هم خاصیتی و فایده‌ای در فرآیند داده‌کاوی به همراه ندارد.

اما outlierها اینگونه نیستند و ارور یا فرمت نادرست و ... ندارند. بلکه داده‌هایی هستند که از عموم دیگر داده‌ها دور هستند و فاصله دارند. و گرنه خود داده مشکلی ندارد بلکه گاهما می‌تواند مفید و داده‌ی هدف باشد و در تحلیل‌ها از آن‌ها استفاده کرد. و یا گاهما داده‌ی هدف نیستند و مورد استفاده قرار نمی‌گیرند و می‌توان حذف‌شان کرد.

۲- آیا noise objects می‌توانند outlier باشند؟

بله ممکن است یک داده‌ی پرت نویز هم داشته باشد به این صورت یک نویزآجکت خواهیم داشت که داده‌ی پرت هم هست. و یا اینکه داده‌ای تحت تاثیر نویز به داده‌پرت تبدیل شود.

۳- آیا outlierها همیشه noise objects هستند؟

خیر outlierها داده‌هایی هستند که پرت هستند و لزوماً نویزی نیستند. داده‌های پرت هم ممکن است مانند بقیه داده (چه پرت چه غیر پرت) نویز داشته باشند و یا نداشته باشند.

۴- آیا noise می‌تواند یک مقدار معمولی را به یک مقدار غیرمعمول تبدیل کند یا برعکس؟

بله ممکن است. نویز واردہ بر داده، همیشه آن را غیرقابل تفسیر نمی‌کند. بلکه ممکن است مقدارش را به گونه‌ای تغییر دهد که هم‌چنان معنی‌دار باشد اما از مقدار اصلی‌اش فاصله گرفته باشد. به این صورت ممکن است یک داده‌ی معمولی را به داده‌ای غیرمعمول تبدیل کند و یا برعکس.

سوال نهم) در رابطه با correlation و cosine measure به سوالات زیر پاسخ دهید.

۱- محدوده مقادیر ممکن برای cosine measure چقدر است؟

طبق فرمول محاسبه‌ی شباهت کسینوسی؛ محدوده مقادیر ممکن شباهت برابر با محدوده مقادیر ممکن برای کسینوس می‌شود که بازه $[1, -1]$ است. اما از آنجایی که عموماً در محاسبه شباهت‌ها زوایای بیشتر از 90° بین بردارها درنظر گرفته نمی‌شود، محدوده مقادیر cosine measure بین 0 تا 1 است. که صفر به معنی کمترین شباهت و 1 به معنای بیشترین شباهت است. (شباهت منفی کمی بی‌معنی است و بیشتر بازه مثبت درنظر گرفته می‌شود)

$$\cos(d1, d2) = (d1 \cdot d2) / \|d1\| \|d2\|$$

۲- اگر cosine measure دو object را برابر یک باشد، آیا آن‌ها یکسان هستند؟

خیر چراکه در این فرمول اندازه بردارها نرمال سازی و می‌شود. پس اگر بردارها ضربی از یکدیگر باشند(هم جهت باشند) هم شباهتشان برابر 1 می‌شود.

۳- چه رابطه‌ای بین correlation و cosine measure وجود دارد؟

در فرمول correlation از میانگین دو بردار نیز استفاده می‌شود و از مقادیر دو بردار(مجموعه داده) کم می‌شود. اگر این میانگین‌ها را صفر 0 درنظر بگیریم و درواقع مقدارش را از بردارها کم نکنیم و یا اگر میانگین هردو بردار برابر صفر باشد، مقدار correlation بینشان با مقدار cosine measure برابر خواهد شد و همان فرمول به دست می‌آید:

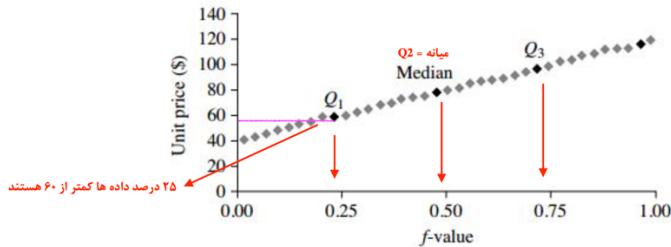
$$\begin{aligned} Corr(x, y) &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \\ &= \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{\|x - \bar{x}\| \|y - \bar{y}\|} \\ &= CosSim(x - \bar{x}, y - \bar{y}) \end{aligned}$$

$$corr(x, y) = \text{cos similarity}(x - \text{mean}(x), y - \text{mean}(y))$$

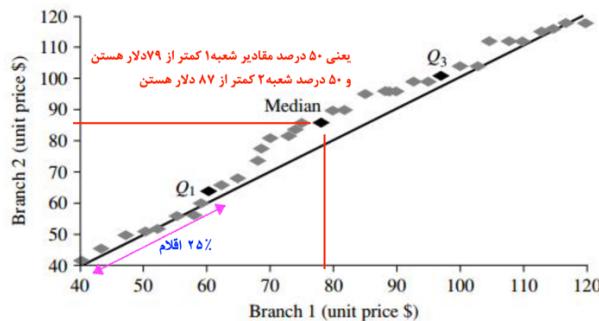
اما این دو مفهوم، به طور کلی معانی متفاوتی را بیان می‌کنند. شباهت کسینوسی میزان شباهت بین دو مجموعه را بیان می‌کند اما correlation میزان همبستگی‌شان را.

سوال دهم) نمودار quantile-quantile را با هم مقایسه کنید.

در نمودار quantile یا چارک رفتار داده‌های یک جمعیت را در چارک‌های مختلف بررسی می‌کنیم. در واقع یک محور را به چارک‌ها اختصاص می‌دهیم و محور دیگر را به مقادیر داده‌هایمان. و با توجه خطی که بین این دو محور رسم می‌شود چه نقاطی از محورها را قطع می‌کند؛ می‌توان بیان کرد که تقریباً چند درصد از توزیع کوچک‌تر چه مقداری هستند.



در نمودار quantile-quantile یا چارک-چارک، رفتار دو توزیع مختلف را در یک نمودار بررسی و مقایسه می‌کنیم. این کار را با رسم مقادیر دو توزیع در محور عمودی و افقی و چارک‌ها را به صورت خط در بین دو محور انجام می‌دهیم. پس اگر توزیع‌ها رابطه خطی داشته باشند، نقاط نمودار چندک-چندک، تقریباً روی یک خط راست قرار می‌گیرند و اگر دو توزیع مقایسه شده مشابه باشند، نقاط روی نمودار تقریباً روی خط $y=x$ قرار خواهند گرفت.



پس همان‌طور که می‌بینیم نمودار quantile تغییرات مقادیر یک جمعیت و توزیع را در چارک‌های مختلف نشان می‌دهد. اما می‌تواند مقادیر دو جمعیت و توزیع متفاوت را در چارک‌های مختلف نشان دهد و مقایسه کند.

سوال یازدهم(11) به طور خلاصه نحوه محاسبه عدم تشابه بین اشیا توصیف شده توسط دو ویژگی nominal و numeric را شرح دهید.

از آن جایی که بردارهایی برای مجموعه داده‌های numeric مان در نظر می‌گیریم؛ می‌توانیم از فرمول‌های محاسبه فاصله‌ی برداری برای محاسبه عدم تشابه در این ویژگی‌ها استفاده کنیم. پس فرمول محاسبه Minkowski Distance برای این ویژگی‌ها استفاده می‌شود. که برخی حالتهای خاص آن پرکاربردتر هستند:

$$d(i, j) = (|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|)^{\frac{1}{h}}$$

که در این فرمول i و j دو بردار(data object) با p بعد هستند. و h مرتبه‌ای است که برای محاسبه فاصله در نظر می‌گیریم(و می‌گوییم L-h norm). چند حالت خاص پرکاربردش:

- $H = 1$ (L1 norm): فاصله منهتن

- $H = 1$ (L2 norm): فاصله اقلیدوسی

- $H \rightarrow \infty$ (Lmax norm): فاصله سوپریمم

بیشترین فاصله بین آجکت او ز

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

در محاسبه عدم تشابه با استفاده از ویژگی nominal نسبت پارامترهای غیر مشترک به کل پارامترها را محاسبه می‌کنیم. درواقع اگر m را تعداد متغیرها(ویژگی) مشابه بین دو مجموعه در نظر بگیریم و p را تعداد کل متغیرها، فاصله‌ی بین دو دیتا آجکت i و j برابر تعداد متغیرهای غیر مشترک(p-m) تقسیم بر تعداد کل متغیرها می‌شود. (درحالت symmetric و برابر بودن ارزش حالت‌ها و متغیرهای مختلف مانند مرد یا زن بودن)

$$d(i, j) = \frac{p - m}{p}$$

اما درحالت asymmetric که مقادیر مختلف یک ویژگی ارزش‌های متفاوتی ایجاد می‌کنند(مانند مثبت یا منفی بودن جواب آزمایش) باید در فرمول‌ها هم این تفاوت‌ها را درنظر گرفت. پس هرکدام از حالتهای هر ویژگی آن‌ها را به فضای باینری می‌بریم و سپس عدم شباهت را محاسبه می‌کنیم. (مثلاً اگر برای رنگ قرمز و آبی و سبز ارزش‌های متفاوت داشته باشند، برای هرکدام مقدار ۰ یا ۱ درنظر می‌گیریم و محاسبات باینری را انجام می‌دهیم) فرمول‌های حالت باینری:

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
sum		$q+s$	$r+t$	p

- Distance measure for symmetric binary variables:

$$d(i, j) = \sqrt{\frac{r+s}{q+r+s+t}}$$

- Distance measure for asymmetric binary variables:

$$d(i, j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (similarity measure for asymmetric binary variables):

$$\text{sim}_{\text{Jaccard}}(i, j) = \frac{q}{q+r+s}$$

- Note: Jaccard coefficient is the same as "coherence":

$$\text{coherence}(i, j) = \frac{\text{sup}(i, j)}{\text{sup}(i) + \text{sup}(j) - \text{sup}(i, j)} = \frac{q}{(q+r)+(q+s)-q}$$

بخش پیاده‌سازی

- پیدا کردن داده‌های از دست رفته NaN با دستور `isna()`:

```
Number of sepal_length NaNs: 2
   sepal_length  sepal_width  petal_length  petal_width      target
73          NaN        2.2        4.5        1.5 Iris-versicolor
143         NaN        3.0        6.1        2.3 Iris-virginica

Number of sepal_width NaNs: 0
Empty DataFrame
Columns: [sepal_length, sepal_width, petal_length, petal_width, target]
Index: []

Number of petal_length NaNs: 2
   sepal_length  sepal_width  petal_length  petal_width      target
20          5.4        3.9        NaN        1.7 Iris-setosa
87          5.5        2.4        NaN        NaN Iris-versicolor

Number of petal_width NaNs: 3
   sepal_length  sepal_width  petal_length  petal_width      target
6           5.4        3.9        1.7        NaN Iris-setosa
27          5.0        3.0        1.6        NaN Iris-setosa
87          5.5        2.4        NaN        NaN Iris-versicolor

Number of target NaNs: 3
   sepal_length  sepal_width  petal_length  petal_width target
60          6.3        3.3        4.7        1.6     NaN
139         6.2        3.4        5.4        2.0     NaN
156         6.5        3.0        5.2        2.0     NaN
```

- حذف داده‌های از دست رفته NaN با دستور `dropna()`:

```
Number of rows = 159
NaNs were removed. Number of rows = 150
```

که می‌بینیم ۹ سطر حذف شدند. (در قسمت قبل ۱۰ مورد NaN پیدا کرده بودیم که دو تا از آن‌ها مشترک‌کا در سطر ۸۷ بودند. پس با حذف NaN‌ها ۹ سطر حذف شد که درست است)

- استفاده از label encoder که میبینیم مقادیر تارگت را تغییر داده و عددی کرده.

Before:					
	sepal_length	sepal_width	petal_length	petal_width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
..
153	6.7	3.0	5.2	2.3	Iris-virginica
154	6.3	2.5	5.0	1.9	Iris-virginica
155	6.5	3.0	5.2	2.0	Iris-virginica
157	6.2	3.4	5.4	2.3	Iris-virginica
158	5.9	3.0	5.1	1.8	Iris-virginica

[150 rows x 5 columns]

After:					
	sepal_length	sepal_width	petal_length	petal_width	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
..
153	6.7	3.0	5.2	2.3	2
154	6.3	2.5	5.0	1.9	2
155	6.5	3.0	5.2	2.0	2
157	6.2	3.4	5.4	2.3	2
158	5.9	3.0	5.1	1.8	2

[150 rows x 5 columns]

- ایرادی که این روش دارد این است که ممکن است در کدگذاری جدیدی که صورت می‌گیرد، الگوریتم برای مقادیر ترتیب و ارزش‌های متفاوت در نظر بگیرد. در صورتی که ممکن است ترتیبی وجود نداشته باشد این باعث اختلال شود. که این مسئله در OneHotEncoder وجود ندارد.

- توضیح همراه با مثال:

در روش label encoder دیدیم که به مقادیر مختلف target، مقادیر عددی مختلف نسبت داده شد:

Iris-setosa: 0, Iris-versicolor: 1, Iris-virginica: 2

که ممکن است ارزش‌های متفاوتی بین‌شان ایجاد کند و بتوان اعداد را با ترتیب در نظر گرفت.

اما در OneHotEncoder برای هر کتگوری یک ستون جداگانه در نظر گرفته می‌شود که هر کدام می‌تواند مقادیر بولین یا باینری داشته باشد(اینکه این ویژگی را دارد یا خیر). به اینصورت ارزش همگی برابر می‌شود. پس برای پیاده‌سازی این روش چند ستون Boolean برای مقادیر مختلف target اضافه می‌شود. در هر سطر(برای هر سطر داده)، هر کدام از مقادیر تارگت که وجود داشته باشد، Boolean آنرا True و دو تای دیگر False می‌شوند.

پس مثلا برای یک سطر که target اش Iris-setosa است، بولین Iris-setosa مقدار ۱ می‌گیرد و دو تای دیگر مقدار ۰ می‌گیرند. (به این صورت بولین‌ها جایگزین ویژگی تارگت می‌شوند)

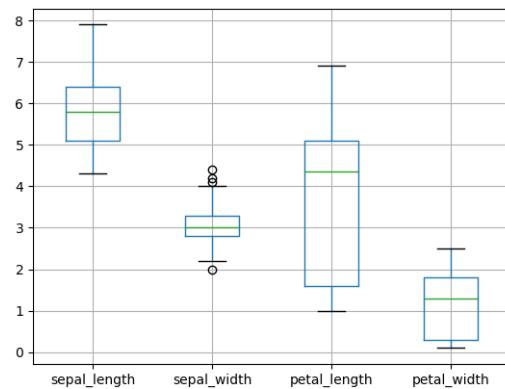
میانگین و واریانس ستون‌های مختلف قبل و بعد از نرمال‌سازی با Standardscalars:

```
Before normalization:
sepal_length:
    mean= 5.84333333333334, variance= 0.6856935123042507
sepal_width:
    mean= 3.054000000000003, variance= 0.1880040268456376
petal_length:
    mean= 3.758666666666666, variance= 3.113179418344519
petal_width:
    mean= 1.1986666666666668, variance= 0.582414317673378

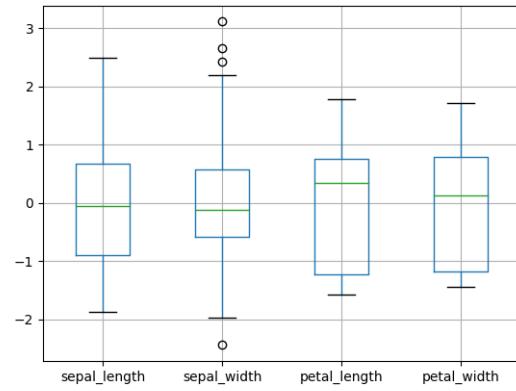
After normalization:
sepal_length:
    mean= -4.736951571734001e-16, variance= 1.0067114093959733
sepal_width:
    mean= -6.631732200427602e-16, variance= 1.006711409395973
petal_length:
    mean= 3.315866100213801e-16, variance= 1.0067114093959728
petal_width:
    mean= -2.842170943040401e-16, variance= 1.0067114093959733
```

با اعمال PCA می‌بینیم که از ۵ ویژگی اولیه، به ۳ ویژگی می‌رسیم:

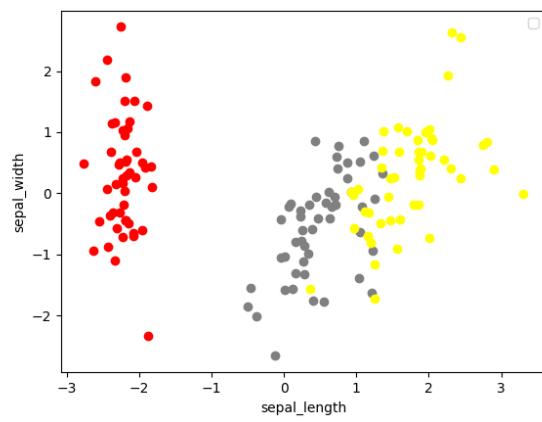
```
After PCA:
      col 1     col 2   target
0    -2.264542  0.505704    0.0
1    -2.086426 -0.655405    0.0
2    -2.367950 -0.318477    0.0
3    -2.304197 -0.575368    0.0
4    -2.388777  0.674767    0.0
..      ...
145   1.870522  0.382822    2.0
146   1.558492 -0.905314    2.0
147   1.520845  0.266795    2.0
148   1.376391  1.016362    2.0
149   0.959299 -0.022284    2.0
```



باکس پلات قبل از نرمال سازی:



باکس پلات بعد از نرمال سازی:



نمودار مت پلات نهایی: