داده کاوی

تمرین اول

آرش حریرپوش
9731505

آرش حریرپوش
9731505

❖ بخش اول – مباحث تئوری و مسائل تشریحی

1. ویژگی های زیر را به صورت دودویی، گسسته یا پیوسته طبقه بندی کنید. همچنین آنها را به صورت کیفی (اسمی یا ترتیبی) یا کمی (بازه یا نسبت) طبقه بندی کنید. برخی از موارد ممکن است بیش از یک حالت داشته باشند، بنابراین اگر فکر می کنید ابهامی وجود دارد به طور خلاصه استدلال خود را نشان دهید.

مثال: سن به سال

پاسخ: گسسته، کمی-نسبی

الف) زاویه اندازه گیری شده بین 0 و 360 درجه

پیوسته، کمی – نسبت

ب) ارتفاع از سطح دریا

پیوسته، کمی – بازه

ج) تراکم ماده به گرم در سانتی متر مکعب

پیوسته، کمی – نسبت

د) مدل های طلا و نقره و برنز در بازی المپیک

گسسته، کیفی – ترتیبی

ح) روشنایی اندازه گیری توسط نور سنج

پیوسته، کمی – بازه

ت) روشنایی اندازه گیری توسط قضاوت ناظر انسانی

گسسته، کیفی – ترتیبی

2. با توجه به تفاوت های نویز و outlier به سوالات زیر با دلیل پاسخ دهید.

الف) آیا نویز می تواند مطلوب باشد؟ outlier چطور؟

به صورت پیش فرض نویز در ویژگی ها مطلوب نمی باشد، به این دلیل که مقادیر اصلی ویژگی ها را تحریف می کند. Outlier ها می توانند به طور بالقوه اشیا (یا مقادیر) داده های درستی باشند؛ برای مثال مشخص کردن آن ها می تواند هدف اصلی برخی از کار های داده کاوی باشد. (در anomaly detection هدف یافتن Outlier ها می باشد) بنابراین، Outlier ها می توانند به طور بالقوه مطلوب باشند، اما نویز نمی تواند.

ب) آیا اشیا نویز می توانند outlier باشند؟

بله، در ویژگی هایی که در مقدار آن ها نویز وجود دارد داده بیشتر به صورت تصادفی (randomized) یا غیر طبیعی (unusual) دیده می شود. بنابراین، ممکن است برخی از داده های نویز به صورت Outlier هم خود را نشان داده باشند.

آرش حریرپوش
9731505

ج) آیا اشیا نویز همیشه outlier هستند؟

خیر، داده نویز دار می تواند شبیه به داده معمولی باشد. بنابراین اشیا نویز همیشه با Outlier ها یکی نیستند.

د) آیا Outlier ها همیشه نویز هستند؟

خیر، Outlier ها می توانند اشیا داده های واقعی باشند که ظاهرا به این مجموعه داده ها تعلق ندارند. این Outlier ها معمولا در دسته اشیا نویز قرار نمی گیرند.

3. با توجه به وکتور های x,y معیار های گفته شده برای آنها را محاسبه کنید.

الف) x= (0,-1,0,1), y = (1, 0, -1, 0) معیار های cosine, correlation, Euclidean

$$Euclidean\ distance: d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \Rightarrow d(x,y)$$

$$= \sqrt{(0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2} = \sqrt{1+1+1+1}$$

$$= \sqrt{4} = \mathbf{2} \Rightarrow \boldsymbol{d(x,y) = 2}$$

$$Correlation: corr(x,y) = \frac{covariance(x,y)}{standard\_deviation(x) * standard\_deviation(y)}$$

$$= \frac{S_{xy}}{S_x S_y}$$

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

$$= \sqrt{\frac{1}{3}(0+1+0+1)} = \sqrt{\frac{2}{3}} \simeq 0.816 \Rightarrow \boldsymbol{S_x \simeq 0.816}$$

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

$$= \sqrt{\frac{1}{3}(1+0+1+0)} = \sqrt{\frac{2}{3}} \simeq 0.816 \Rightarrow \boldsymbol{S_y \simeq 0.816}$$

آرش حریرپوش
9731505

$$covariance(x,y) = S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{3}\{(0-0)(1-0) + (-1-0)(0-0) + (0-0)(-1-0)$$

$$+ (1-0)(0-0)\} = 0 \Rightarrow \boldsymbol{S_{xy} = 0}$$

$$corr(x,y) = \frac{S_{xy}}{S_x S_y} = \frac{0}{\frac{2}{3}} = \mathbf{0} \Rightarrow \boldsymbol{corr(x,y) = 0}$$

$$Cosine\ Similarity: \cos(x,y) = \frac{< x,y >}{||x||\ ||y||} = \frac{x.y}{\sqrt{(x.x)(y.y)}} \Rightarrow \cos(x,y)$$

$$= \frac{0*1 + (-1)*0 + 0*(-1) + 1*0}{\sqrt{(0+1+0+1)(1+0+1+0)}} = \frac{0}{2} = \mathbf{0} \Rightarrow \boldsymbol{cos(x,y) = 0}$$

ب) x = (2, -1, 0, 2, 0, -3), y = (-1, 1, -1, 0, 0, -1) معیار های cosine, correlation, Euclidean

$$Euclidean\ distance: d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \Rightarrow d(x,y)$$

$$= \sqrt{(2+1)^2 + (-1-1)^2 + (0+1)^2 + (2-0)^2 + (0-0)^2 + (-3+1)^2}$$

$$= \sqrt{9+4+1+4+4} = \sqrt{22} \simeq \mathbf{4.7} \Rightarrow \boldsymbol{d(x,y) \simeq 4.7}$$

$$Correlation: corr(x,y) = \frac{covariance(x,y)}{standard\_deviation(x) * standard\_deviation(y)}$$

$$= \frac{S_{xy}}{S_x S_y}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(2-1+0+2+0-3) = 0 \Rightarrow S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{5}(4+1+0+4+0+9)} = \sqrt{\frac{18}{5}} \simeq 1.9 \Rightarrow \boldsymbol{S_x \simeq 1.9}$$

آرش حریرپوش
9731505

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}(-1+1-1+0+0-1) = -\frac{1}{3} \Rightarrow S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$= \sqrt{\frac{1}{5}\left(\left(-1+\frac{1}{3}\right)^2 + \left(1+\frac{1}{3}\right)^2 + \left(-1+\frac{1}{3}\right)^2 + \left(0+\frac{1}{3}\right)^2 + \left(0+\frac{1}{3}\right)^2 + \left(-1+\frac{1}{3}\right)^2\right)}$$

$$= \sqrt{\frac{1}{5}\left(3*\frac{4}{9}+\frac{16}{9}+2*\frac{1}{9}\right)} = \sqrt{\frac{1}{5}*\frac{30}{9}} = \sqrt{\frac{6}{9}} = \sqrt{\frac{2}{3}} \simeq 0.816 \Rightarrow \boldsymbol{S_y \simeq 0.816}$$

$$covariance(x,y) = S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{5}\left\{(2-0)\left(-1+\frac{1}{3}\right) + (-1-0)\left(1+\frac{1}{3}\right) + (0-0)\left(-1+\frac{1}{3}\right)\right.$$
$$\left. + (2-0)\left(0+\frac{1}{3}\right) + (0-0)\left(0+\frac{1}{3}\right) + (-3-0)(-1+\frac{1}{3})\right\}$$

$$= \frac{1}{5}\left\{-\frac{4}{3}-\frac{4}{3}+\frac{2}{3}+2\right\} = \frac{1}{5}*0 = 0 \Rightarrow \boldsymbol{S_{xy} = 0}$$

$$corr(x,y) = \frac{S_{xy}}{S_x S_y} = \frac{0}{0.816*1.9} = \mathbf{0} \Rightarrow \boldsymbol{corr(x,y) = 0}$$

$$Cosine\ Similarity: \cos(x,y) = \frac{< x,y >}{||x||\,||y||} = \frac{x.y}{\sqrt{(x.x)(y.y)}} \Rightarrow \cos(x,y)$$

$$= \frac{2*(-1)+(-1)*1+0*(-1)+2*0+0*0+(-3)*(-1)}{\sqrt{(4+1+0+4+0+9)(1+1+1+0+0+1)}}$$

$$= \frac{-2-1+3}{\sqrt{18*4}} = \frac{\mathbf{0}}{\sqrt{72}} \Rightarrow \boldsymbol{cos(x,y) = 0}$$

ج) x= (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) معیار های cosine, correlation, Jaccard

$$Jaccard\ Coefficient: J(x,y) = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} \Rightarrow J(x,y) = \frac{3}{1+1+3} = \frac{3}{5} = \mathbf{0.6}$$

$$\Rightarrow \boldsymbol{J(x,y) = 0.6}$$

آرش حریرپوش
9731505

$$Correlation: corr(x,y) = \frac{covariance(x,y)}{standard\_deviation(x) * standard\_deviation(y)}$$

$$= \frac{S_{xy}}{S_x S_y}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(1+1+0+1+0+1) = \frac{4}{6} = \frac{2}{3} \Rightarrow S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{5}(4 * \left(1-\frac{2}{3}\right)^2 + 2 * \left(-\frac{2}{3}\right)^2)} = \sqrt{\frac{1}{5}(\frac{4}{9}+\frac{8}{9})} = \sqrt{\frac{1}{5} * \frac{12}{9}} = \sqrt{\frac{4}{15}}$$

$$\simeq 0.51 \Rightarrow \boldsymbol{S_x \simeq 0.51}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}(1+1+1+0+0+1) = \frac{4}{6} = \frac{2}{3} \Rightarrow S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \sqrt{\frac{1}{5}\left(4 * \left(1-\frac{2}{3}\right)^2 + 2 * \left(0-\frac{2}{3}\right)^2\right)} = \sqrt{\frac{1}{5}\left(\frac{4}{9}+\frac{8}{9}\right)} = \sqrt{\frac{1}{5} * \frac{12}{9}} = \sqrt{\frac{12}{45}}$$

$$= \sqrt{\frac{4}{15}} \simeq 0.51 \Rightarrow \boldsymbol{S_y \simeq 0.51}$$

$$covariance(x,y) = S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{5}\left\{\left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(0-\frac{2}{3}\right)\left(1-\frac{2}{3}\right)\right.$$

$$\left. + \left(1-\frac{2}{3}\right)\left(0-\frac{2}{3}\right) + \left(0-\frac{2}{3}\right)\left(0-\frac{2}{3}\right) + (1-\frac{2}{3})(1-\frac{2}{3})\right\}$$

$$= \frac{1}{5}\left\{\frac{1}{9}+\frac{1}{9}-\frac{2}{9}-\frac{2}{9}+\frac{4}{9}+\frac{1}{9}\right\} = \frac{1}{5} * \frac{3}{9} = \frac{1}{15} \Rightarrow \boldsymbol{S_{xy} = \frac{1}{15}}$$

$$corr(x,y) = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{15}}{\sqrt{\frac{4}{15}} * \sqrt{\frac{4}{15}}} = \frac{\frac{1}{15}}{\frac{4}{15}} = \frac{1}{4} = \mathbf{0.25} \Rightarrow \boldsymbol{corr(x,y) = 0.25}$$

آرش حریرپوش
9731505

$$Cosine\ Similarity: \cos(x,y) = \frac{< x,y >}{||x||\,||y||} = \frac{x.y}{\sqrt{(x.x)(y.y)}} \Rightarrow \cos(x,y)$$

$$= \frac{1*1+1*1+0*1+1*0+0*0+1*1}{\sqrt{(1+1+0+1+0+1)(1+1+1+0+0+1)}}$$

$$= \frac{1+1+0+0+0+1}{\sqrt{4*4}} = \frac{3}{4} = \mathbf{0.75} \Rightarrow \boldsymbol{cos(x,y) = 0.75}$$

4. برای رفع هر یک از چالش های زیر دو راه حل ارائه دهید.

الف) وجود تاپل هایی با مقادیر ویژگی های بخصوص حذف شده در مجموعه داده.

برای رفع این چالش می توان از راه حل های زیر استفاده نمود:

- حذف اشیا داده یا متغیر های آن: این کار را در صورتی می توان انجام داد که داده های زیادی برای آموزش مدل وجود داشته و حذف یک داده باعث کاهش کارآیی مدل نشود.
- تخمین داده حذف شده: در این روش می توان به کمک برخی از الگوریتم ها داده حذف شده را تخمین زد؛ برای مثال در داده های سری زمانی مربوط به دما می توان در صورتی که دما برای یک زمان مشخص نشده باشد با تحلیل دما در نقاط اطراف آن مقدار حذف شده را تخمین زد.
- ویژگی که مقدار آن در برخی از تاپل ها حذف شده را کلا بررسی نکنیم: این روش نیز زمانی می تواند موثر باشد که ویژگی های کافی برای بررسی آن داده وجود داشته باشد و حذف یک ویژگی از تحلیل باعث کاهش کارآیی نشود. (برای مثال در صورتی که 2 ویژگی داشته باشیم حذف یک ویژگی ممکن است باعث ایجاد مشکل در تحلیل آن داده شود اما در حالتی که 100 ویژگی داشته باشیم احتمالا بتوان یک ویژگی را در بررسی تاثیر نداد و آن را از تحلیل حذف کرد)

ب) پردازش مجموعه داده های با ابعاد بسیار بالا تا هزار ویژگی و نحوه انتخاب ویژگی های پر اهمیت.

برای رفع این چالش می توان از روش های کاهش ابعاد استفاده نمود که برخی از آن ها عبارتند از:

- تولید ویژگی های جدید که از ترکیب متغیر های قدیمی ایجاد شده است. (Feature Creation)
- کاهش دادن ابعاد مجموعه داده ها به کمک الگوریتم PCA.
- از بین همه ویژگی ها برخی از آن ها را انتخاب می کنیم (Feature Selection)؛ ویژگی های تکراری (ویژگی هایی بیشتر یا همه اطلاعات آن ها در یک یا چند ویژگی دیگر وجود دارد) و غیر مرتبط (ویژگی هایی که شامل اطلاعات مفیدی در زمینه ای که کار داده کاوی انجام می شود نمی باشد) حذف می گردد.

ج) مجموعه داده های نامتوازن (اختلاف بالا میان تعداد تاپل های با برچسب متفاوت)

این چالش معمولا به مشکلی در طبقه بندی (Classification) اشاره دارد که کلاس ها را به طور مساوی نشان نمی دهد؛ برای رفع این چالش می توان از روش های زیر استفاده نمود:

- جمع آوری داده بیشتر: یک مجموعه داده بزرگ تر ممکن است متفاوت تر عمل کند و شاید دیدگاه متعادل تری را نسبت به کلاس ها نشان دهد.
- عوض کردن معیار عملکرد (Performance Metric): دقت معیار مناسبی برای مجموعه داده های نا متعادل نمی باشد؛ به این دلیل که گمراه کننده می باشد. همچنین می توان از معیار های عملکرد دیگری چون Confusion Matrix، Precision، Recall یا F-score ا ستفاده کرد که نسبت به دقت شهود بیشتری از مدل را نشان می دهد.
- استفاده از الگوریتم های مختلف: از الگوریتم های یکسان برای حل هر نوع مشکلی نباید استفاده کرد. حداقل باید الگوریتم های مختلفی را در مورد یک مسئله مشخص بررسی کرد.
- نمونه برداری مجدد از داده: می توان مجموعه داده ای که برای ساخت مدل پیشبینی استفاده می شود را تغییر داد تا داده های متعادل تری داشته باشیم.

(منبع: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/)

د) رخداد بیش برازش در مدل های یادگیری

بیش برازش زمانی رخ می دهد که مدل نسبت میزان نویز و داده های آموزشی بسیار پیچیده باشد؛ برای حل آن می توان از روش های زیر استفاده نمود:

- ساده کردن مدل
- جمع آوری تعداد بیشتری داده آموزشی
- کاهش نویز موجود در داده های آموزشی

هـ) رخداد کم برازش در مدل های یادگیری

کم برازش زمانی رخ می دهد که مدل برای یادگیری ساختار موجود در داده ها بسیار ساده باشد؛ برای حل آن می توان از روش های زیر استفاده نمود:

- استفاده از مدل قوی تر
- انجام مهندسی feature (برای مثال اعمال یک تابع مشخص به ویژگی ها بر اساس دانشی که از آن موضوع داریم و extract کردن feature های جدید)

5. به سوالات زیر در ارتباط با رگرسیون خطی پاسخ دهید.

الف) آیا این الگوریتم نسبت بهoutlier ها حساسیت دارد؟ توضیح دهید.

Outlier تنها در صورتی در رگرسیون تاثیر دارند که در معادله رگرسیون تاثیر زیادی داشته باشند. بعضی مواقع Outlier ها تاثیر زیادی بر روی معادله رگرسیون ندارند. به عنوان مثال زمانی که مجموعه داده ها خیلی بزرگ باشد، یک Outlier تاثیر چندانی بر روی معادله رگرسیون ندارد.

(منبع: https://stattrek.com/regression/influential-points.aspx)

ب) معیار اصلی اندازه گیری خطا در این الگوریتم چیست و چرا؟

معیار اصلی اندازه گیری خطا در این الگوریتم محاسبه عددی است که بتواند فاصله داده ها با خط رسم شده توسط این الگوریتم را محاسبه کند که این کار را به کمک روش هایی مانند Mean Squared Error (MSE) یا Root-Mean-Squared-Error (RMSE) می باشد؛ MSE به سادگی برابر مربع میانگین بین فاصله مقدار هدف و مقدار پیشبینی شده در این الگوریتم می باشد. به دلیل اینکه مربع فاصله را محاسبه می کند، حتی خطا های کوچک را نیز مجازات می کند، که باعث بیش از حد بد بودن مدل می شود. و با توجه به اینکه قابل تغییر است و بهتر می تواند بهینه شود مناسب می باشد؛ همچنین به دلیل convex بودن آن می توان با استفاده از Gradient Descent بدون ماندن در در min محلی مقدار خطا را کمینه کرد. همچنین RMSE نیز یکی از معیار های پر کاربرد در رگرسیون می باشد.که ریشه مربع اختلاف بین مقدار هدف و مقدار پیشبینی شده توسط مدل را نشان می دهد؛ در برخی موارد به دلیل اینکه ابتدا خطا ها قبل از اینکه میانگین گرفته شوند مربع می شوند و که باعث می شود مجازات بالایی برای خطا های بزرگ ایجاد شود. این بدان معنی است که RMSE زمانی که خطا های بزرگ ناخواسته ایجاد می شوند مناسب می باشد.

$$MSE = ||X\beta - y||_2^2 = \frac{1}{n}\sum(y - \hat{y})^2 \, , y \; is \; target \; data, \hat{y} \; is \; predicted \; data$$

$$RMSE = \sqrt{||X\beta - y||_2^2} = \sqrt{\frac{1}{n}\sum(y - \hat{y})^2} \, , y \; is \; target \; data, \hat{y} \; is \; predicted \; data$$

(منبع: https://towardsdatascience.com/regression-an-explanation-of-regression-metrics-and-what-can-go-wrong-a39a9793d914#:~:text=Mean%20Squared%20Error%3A%20MSE%20or,predicted%20by%20the%20regression%20model.)

آرش حریرپوش
9731505

ج) جدول زیر مقادیر متناظر قد و وزن ده نفر را نشان می دهد. معادلات خطوط کمترین مربعات را به طور تقریبی بدست آورید

قد (m)	1.58	1.60	1.62	1.65	1.68	1.70	1.74	1.75	1.77	1.80
وزن (kg)	57.5	58.2	59.5	62.1	63.4	64.5	66.2	67.7	69.4	71.3

برای محاسبه خط در این رگرسیون باید ابتدا خط را در فرم $y = mx + b$ محاسبه کرد تا پیشبینی ها بر اساس آن صورت گیرد. در اینجا به دنبال خطی هستیم که فاصله آن با تمام نقاط تا جایی که ممکن است کوچک باشد.

خط کمترین مربعات دارای دو بخش می باشد: شیب m و intercept برابر b. ابتدا m محاسبه خواهد شد، و پس از آن b. معادلات m و b برابر است با:

$$\sum xy = 1.58 * 57.5 + 1.6 * 58.2 + 1.62 * 59.5 + 1.65 * 62.1 + 1.68 * 63.4 + 1.7 * 64.5 + 1.74 * 66.2 + 1.75 * 67.7 + 1.77 * 69.4 + 1.8 * 71.3$$
$$= 90.85 + 93.12 + 96.39 + 102.465 + 106.512 + 109.65 + 115.188 + 118.475 + 122.838 + 128.34 = 1{,}083.828$$

$$\sum x = 1.58 + 1.6 + 1.62 + 1.65 + 1.68 + 1.7 + 1.74 + 1.75 + 1.77 + 1.8 = 16.89$$

$$\sum y = 57.5 + 58.2 + 59.5 + 62.1 + 63.4 + 64.5 + 66.2 + 67.7 + 69.4 + 71.3 = 639.8$$

$$\sum x^2 = 1.58^2 + 1.6^2 + 1.62^2 + 1.65^2 + 1.68^2 + 1.7^2 + 1.74^2 + 1.75^2 + 1.77^2 + 1.8^2$$
$$= 2.4964 + 2.56 + 2.6244 + 2.7225 + 2.8224 + 2.89 + 3.0276 + 3.0625 + 3.1329 + 3.24 = 28.5787$$

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \Rightarrow m = \frac{10(1083.828) - 16.89 * 639.8}{10 * 28.5787 - 16.89^2}$$

$$= \frac{10838.28 - 10806.222}{285.787 - 285.2721} = \frac{32.058}{0.5149} \simeq 62.26 \Rightarrow \boldsymbol{m = 62.26}$$

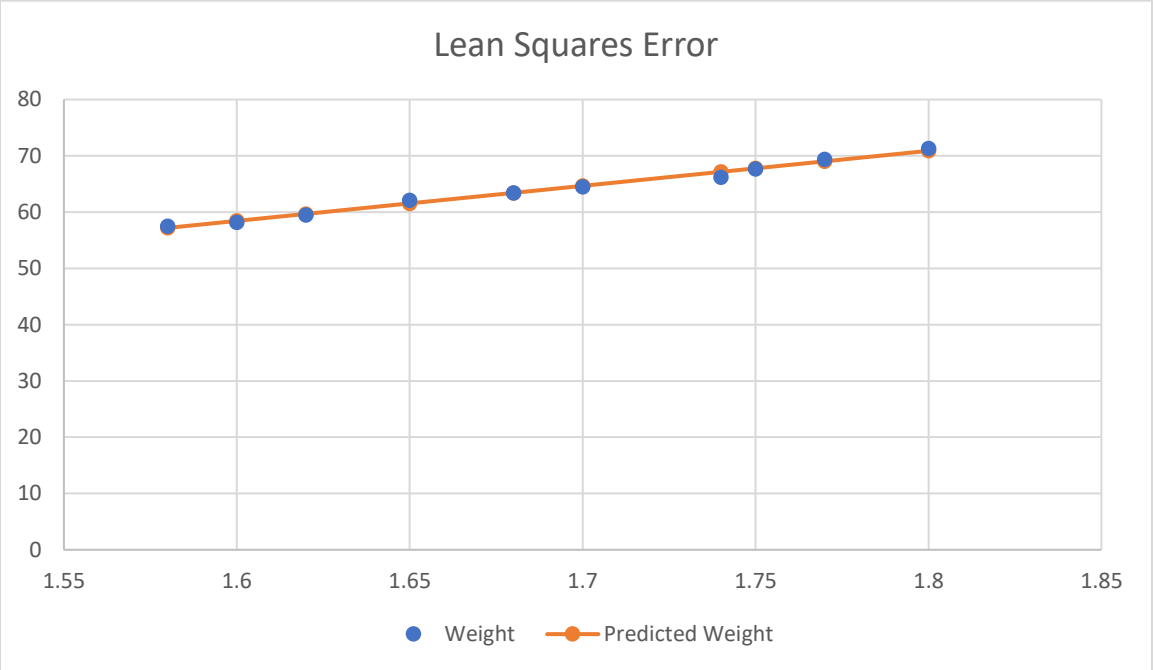
$$b = \frac{\sum y - m(\sum x)}{n} \Rightarrow b = \frac{639.8 - 62.26 * 16.89}{10} = \frac{639.8 - 1{,}051.5714}{10}$$

$$= -\frac{411.7714}{10} = -41.17714 \Rightarrow \boldsymbol{b = -41.17714}$$

بنابراین معادله خط کمترین مربعات برابر است با:

$$y = mx + b \Rightarrow \boldsymbol{y = 62.26 \times x - 41.17714}$$

در شکل زیر نقاط آبی رنگ برابر داده ها و خط نارنجی رنگ برابر خط رسم شده توسط regression می باشد:



(منبع: https://towardsdatascience.com/linear-regression-by-hand-ee7fe5a751bf)

آرش حریرپوش
9731505

6. در یک نظرسنجی انجام شده از 200 هزار نفر در آمریکا، تمایل آنها به یکی از دو حزب جمهوری‌خواه و دموکرات پرسیده شده است. در این نظرسنجی افراد متعلق به سه طبقه اقتصادی اصلی ضعیف، متوسط و مرفه حضور داشته‌اند و تعداد افراد هر طبقه که به یکی از دو حزب رای داده‌اند در جدول زیر مشخص شده است.

تعداد نفر	طبقه اقتصادی	حزب مورد علاقه
20,000	ضعیف	دموکرات
35,000	متوسط	دموکرات
45,000	مرفه	دموکرات
50,000	ضعیف	جمهوری‌خواه
20,000	متوسط	جمهوری‌خواه
30,000	مرفه	جمهوری‌خواه

الف) مقدار Entropy هر یک از متغیرهای طبقه اقتصادی و حزب مورد علاقه را بدست آورید.

مقدار Entropy برای طبقه اقتصادی متوسط برابر است با:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \Rightarrow H(Economic\ Class = Medium)$$

$$= -\frac{35000}{200000}\log_2\frac{35000}{200000}$$

$$-\frac{20000}{200000}\log_2\frac{20000}{200000} = -(0.175\log_2 0.175 + 0.1\log_2 0.1)$$

$$\simeq -(0.175 * -2.5 + 0.1 * -3.3) \simeq -(-0.4375 - 0.33) \simeq \mathbf{0.7675}$$

$$\Rightarrow \boldsymbol{H(Economic\ Class = Medium) = 0.7675}$$

مقدار Entropy برای حزب دموکرات برابر است با:

آرش حریرپوش
9731505

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \Rightarrow H(Favorite\ Faction = Democrat)$$

$$= -\frac{20000}{200000}\log_2\frac{20000}{200000} - \frac{35000}{200000}\log_2\frac{35000}{200000} - \frac{45000}{200000}\log_2\frac{45000}{200000}$$

$$= -(0.1\log_2 0.1 + 0.175\log_2 0.175 + 0.225\log_2 0.225)$$

$$\simeq -(0.1 * -3.3 + 0.175 * -2.5 + 0.225 * -2.1)$$

$$\simeq -(-0.33 - 0.4375 - 0.4725) \simeq \mathbf{1.24}$$

$$\Rightarrow \boldsymbol{H(Favorite\ Faction = Democrat) = 1.24}$$

ب) مقدار Mutual Information دو متغیر طبقه اقتصادی و حزب مورد علاقه را بدست آورید.

مقدار Mutual Information برای طبقه متوسط که به حزب دموکرات علاقه دارند:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \Rightarrow I(Favorite\ Faction = Democrat, Economic\ Class = Medium) = H(Favorite\ Faction = Democrat) + H(Economic\ Class = Medium) - H(Favorite\ Faction = Democrat, Economic\ Class = Medium)$$

با توجه به بخش الف می دانیم:

$$\boldsymbol{H(Favorite\ Faction = Democrat) = 1.24}$$

$$\boldsymbol{H(Economic\ Class = Medium) = 0.7675}$$

بنابراین برای محاسبه این مقدار به محاسبه عبارت زیر پرداخته می شود:

$$H(Favorite\ Faction = Democrat, Economic\ Class = Medium) = -\sum_{i} -\sum_{j} p_{ij}\log_2 p_{ij}$$

$$\Rightarrow H(Favorite\ Faction = Democrat, Economic\ Class = Medium)$$

$$= -\left(\frac{35000}{200000}\log_2\frac{35000}{200000}\right) = -(0.175\log_2 0.175) = -0.175 * -2.5$$

$$= 0.4375$$

$$\Rightarrow \boldsymbol{H(Favorite\ Faction = Democrat, Economic\ Class = Medium) = 0.4375}$$

بنابراین حاصل مقدار Mutual Information برای طبقه متوسط که به حزب دموکرات علاقه دارند برابر است با:

$$I(Favorite\ Faction = Democrat, Economic\ Class = Medium)$$
$$= H(Favorite\ Faction = Democrat) + H(Economic\ Class = Medium)$$
$$- H(Favorite\ Faction = Democrat, Economic\ Class = Medium)$$
$$= 1.24 + 0.7675 - 0.4375 = 1.57$$
$$\Rightarrow \boldsymbol{I(Favorite\ Faction = Democrat, Economic\ Class = Medium)}$$
$$\boldsymbol{= 1.57}$$

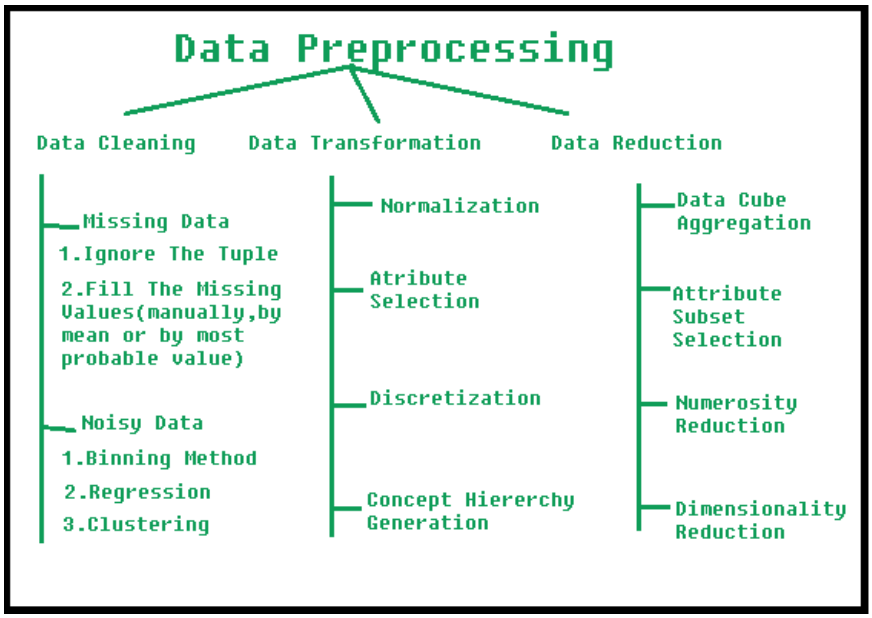
ج) آیا این دو متغیر از هم مستقل‌اند؟ تحلیل خود را از جواب بدست آمده ارائه دهید.

خیر، با توجه به مقدار Mutual Information که برای طبقه متوسط که به حزب دموکرات علاقه دارند که در بخش ب محاسبه شده است، به این دلیل که در صورتی که مقدار Mutual Information برابر صفر باشد می توان نتیجه گرفت که دو متغیر مستقل هستند؛ و با توجه به اینکه این مقدار در بخش ب برابر 1.57 محاسبه شده است بنابراین این دو متغیر از هم مستقل نیستند.

.7

الف) انواع مختلف پیش پردازش داده و موارد استفاده آنها را به اختصار توضیح دهید.

در داده کاوی از تکنیک های پیش پردازش استفاده می شود تا داده خام را به یک فرمت کارآمد و مفید تبدیل کرد.



گام هایی در پیش پردازش نیاز است برابر است با:

1. مرتب کردن داده ها (Data Cleaning): داده ها ممکن است قسمت های نامرتبط و گمشده زیادی داشته باشند. برای برطرف کردن این مشکل از Data Cleaning استفاده می شود. این بخش شامل مدیریت داده های از بین رفته، داده های نویزی و غیره می باشد.

- داده های از بین رفته (Missing Data): این وضعیت زمانی رخ می دهد که برخی از داده ها از بین رفته باشد. از طریق روش های گوناگونی می تواند برطرف شود. برخی از این روش ها عبارتند از:
  - در نظر نگرفتن تایل ها(Ignore the tuples): این روش زمانی استفاده می شود که مجموعه داده های ما تقریبا بزرگ باشد و چندین مقدار در یک تاپل ها از بین رفته باشد.
  - پر کردن مقادیر از بین رفته (Fill the Missing values): روش های زیادی برای انجام این کار وجود دارد. می توان مقادیر از بین رفته را به صورت دستی پر کرد، به کمک ویژگی میانگین (mean) یا محتمل ترین مقدار.
- داده های نویزی (Noisy Data): داده های نویزی داده های بی معنی می باشند که نمی تواند توسط ماشین تفسیر شود. می تواند از طریق جمع آوری مجموعه داده های اشتباه، خطا های داده های ورودی و غیره تولید شده باشد. به کمک روش های زیر می توان این مشکل را برطرف کرد:
  - روش Binning (Binning Method): این روش بر روی داده های مرتب شده کار می کند تا آن ها را ساده (smooth) کند. کل داده ها به بخش هایی با اندازه برابر تقسیم می شود و سپس روش های گوناگونی اجرا می شود تا کار کامل شود. هر بخش به صورت جداگانه بررسی می شود. می توان همه داده های یک بخش را با میانگین جایگزین کرد یا می توان مقادیر مرزی برای کامل کردن کار استفاده کرد.
  - Regression: در اینجا داده ها را می توان با قرار دادن در یک تابع Regression ساده کرد. Regression استفاده شده می تواند خطی (یک متغیر مستقل داشته باشد) یا چند تایی (چند متغیر مستقل داشته باشد) باشد
  - Clustering: این روش داده های یکسان را در یک خوشه (cluster) قرار می دهد. Outlier ها ممکن است تشخیص داده نشوند یا بیرون از خوشه ها افتاده باشند.

2. تبدیل داده ها (Data Transformation): در این بخش داده ها به فرم مناسب برای انجام عملیات استخراج تبدیل می شوند. این بخش شامل روش های زیر می باشد:
   - نرمال سازی (Normalization): این کار برای مقیاس گذاری مقادیر در یک محدوده مشخص انجام می شود.
   - انتخاب ویژگی ها (Attribute Selection): در این استراتژی ، ویژگی های جدیدی از طریق مجموعه ویژگی های داده شده ساخته می شود تا به فرآیند استخراج کمک کند.

- گسسته سازی (Discretization): این کار انجام می شود تا مقادیر خام ویژگی numeric با سطح های interval یا سطح های مفهومی (conceptual) جایگزین شود.
- تولید سلسله مراتب مفهومی (Concept Hierarchy Generation): در اینجا ویژگی ها به سطح های بالاتری سلسله مراتب تبدیل می شوند. برای مثال- ویژگی شهر می تواند به کشور تبدیل شود. (در اسلاید ها این روش در بخش Aggregation، Change of scale عنوان شده است.)

3. کاهش داده ها (Data Reduction): از آنجا که داده کاوی روشی است که برای کنترل حجم زیادی از داده ها استفاده می شود. زمانی که با حجم زیادی از داده کار می شود، در این موارد تجزیه و تحلیل سخت تر می شود. برای برطرف کردن این مشکل، از روش های کاهش داده ها استفاده می شود. هدف آن افزایش کارآیی حافظه و کاهش داده ها در حافظه و هزینه تجزیه و تحلیل ها می باشد. برای انجام این کار می توان از روش های زیر استفاده کرد:
   - Data Cube Aggregation: عملیات تجمع (Aggregation) برای ساخت مکعب داده (data cube) بر روی داده ها اعمال می شود
   - انتخاب زیر مجموعه ویژگی (Attribute Subset Selection): داده هایی که با یک دیگر ارتباط زیادی دارند باید استفاده شوند، بقیه می توانند کنار گذاشته شوند. برای انجام انتخاب زیر مجموعه ویژگی ها، می توان از سطح معنی دار (level of significance) و مقدار p ویژگی استفاده کرد. ویژگی که دارای مقدار p بزرگ تر از سطح معنی باشد می تواند کنار گذاشته شود.
   - Sampling: نمونه برداری خوب است که نمایانگر تمام داده ها باشد و تقریبا ویژگی ها داده اصلی را داشته باشد. این کار به دو روش انجام می شود:
     - نمونه برداری بدون جایگزینی (Sampling without replacement): در صورتی که یک نمونه انتخاب شود نمی تواند دوباره انتخاب و از مجموعه داده های انتخابی حذف می شود.
     - نمونه برداری با جایگزینی (Sampling with replacement): در صورتی که یک نمونه انتخاب شود از مجموعه داده های انتخابی حذف نمی گردد و می تواند دوباره انتخاب شود. در این روش یک نمونه می تواند چندین بار انتخاب شود.
   - کاهش عدد (Numerosity Reduction): با انجام این کار می توان بجای کل داده مدل داده را ذخیره کرد، برای مثال مدل های Regression.
   - کاهش بعد (Dimensionality Reduction): این روش با استفاده از مکانیسم های رمز گذاری (encoding) اندازه داده را کاهش می دهد. می تواند زیان آور یا بدون زیان باشد. اگر پس از بازسازی داده های فشرده شده باشد، داده های قبلی می تواند بازیابی شود، این کاهش

ها را کاهش بدون زیان می نامند در غیر این صورت کاهش زیان آور می شود. دو روش موثر در کاهش ابعاد برابر است با: تبدیل Wavelet و PCA.

(منبع: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/)

ب) با دو روش بیشینه کمینه (min=0, max=1) و z-score داده های زیر را نرمال کنید.

1000 ,600 ,400 ,300 ,200

به کمک فرمول زیر می توان داده های بالا را به روش بیشینه کمینه نرمال کرد :

$$A = \{1000, 600, 400, 300, 200\}, \qquad new\_max_A = 1, \qquad new\_min_A = 0$$

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \Rightarrow v_i'$$

$$= \frac{v_i - 200}{1000 - 200}(1 - 0) + 0 \Rightarrow v_i' = \frac{v_i - 200}{800}$$

بنابراین مجموعه داده های نرمال شده از طریق روش بیشینه کمینه برابر است با:

$$A' = \left\{\frac{800}{800}, \frac{400}{800}, \frac{200}{800}, \frac{100}{800}, \frac{0}{800}\right\} \Rightarrow \boldsymbol{A' = \{1, 0.5, 0.25, 0.125, 0\}}$$

برای محاسبه مقدار نرمال شده داده ها به روش *z-score* ابتدا مقادیر زیر محاسبه می گردد:

$$\bar{A} = \frac{1000 + 600 + 400 + 300 + 200}{5} = \frac{2500}{5} = 500$$

$$S_A = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{A})^2} = \sqrt{\frac{1}{4}(500^2 + 100^2 + 100^2 + 200^2 + 300^2)}$$

$$= \sqrt{\frac{1}{4}(25 + 1 + 1 + 4 + 9) \times 10^4} = \sqrt{\frac{1}{4}(40) \times 10^4} = \sqrt{10^5} \simeq 316.2$$

حال به کمک فرمول زیر می توان داده های بالا را به روش *z-score* نرمال کرد:

$$v_i' = \frac{v_i - \bar{A}}{S_A} \Rightarrow v_i' = \frac{v_i - 500}{316.2}$$

بنابراین مجموعه داده های نرمال شده از طریق روش *z-score* برابر است با:

$$A' = \left\{\frac{500}{316.2}, \frac{100}{316.2}, -\frac{100}{316.2}, -\frac{200}{316.2}, -\frac{300}{316.2}\right\} \Rightarrow \boldsymbol{A'}$$

$$\boldsymbol{= \{1.58, 0.31, -0.31, -0.63, -0.95\}}$$

.8

الف) در منظم سازی tikhonov معادله نرمال مربوطه را به دست آورید.

$$f(\beta) = \left|\left|X\beta - y\right|\right|_2^2 + \alpha\left|\left|\beta\right|\right|_2^2 = (X\beta - y)^T(X\beta - y) + \alpha\beta^T\beta$$
$$= \beta^T X^T X\beta - \beta^T X^T y - y^T X\beta + y^T y + \alpha\beta^T\beta$$

اگر β مینیمم محلی f باشد، بنابراین $\nabla f(\beta)$ برابر یک بردار صفر می باشد. ابتدا گرادیان f محاسبه می گردد، می دانیم:

$$\nabla f(\beta) = \begin{pmatrix} \frac{\partial f}{\partial \beta_1} \\ \vdots \\ \frac{\partial f}{\partial \beta_n} \end{pmatrix}$$

ابتدا گرادیان های زیر محاسبه می شود:

$$\nabla(\beta^T X^T y) = X^T y, \quad \nabla(y^T X\beta) = X^T y, \quad \nabla(\beta^T X^T X\beta) = 2X^T X\beta, \quad \nabla(y^T y) = 0,$$
$$\nabla(\alpha\beta^T\beta) = 2\alpha\beta$$

بنابراین $\nabla f(\beta)$ برابر است با:

$$\nabla f(\beta) = \nabla(\beta^T X^T X\beta) - \nabla(\beta^T X^T y) - \nabla(y^T X\beta) + \nabla(y^T y) + \nabla(\alpha\beta^T\beta)$$
$$= 2X^T X\beta - X^T y - X^T y + 0 + 2\alpha\beta = 2X^T X\beta - 2X^T y + 2\alpha\beta$$
$$\Rightarrow \quad \boldsymbol{\nabla f(\beta) = 2X^T X\beta - 2X^T y + 2\alpha\beta}$$

حال برای محاسبه معادله نرمال مربوطه $\nabla f(\beta) = 0$ قرار داده می شود:

$$\nabla f(\beta) = 0 = 2X^T X\beta - 2X^T y + 2\alpha\beta \Rightarrow 2X^T X\beta - 2X^T y + 2\alpha\beta = 0$$
$$\Rightarrow X^T X\beta - X^T y + \alpha\beta = 0 \Rightarrow \boldsymbol{X^T X\beta + \alpha\beta = X^T y}$$

بنابراین معادله نرمال مربوطه برابر است با:

$$\boldsymbol{X^T X\beta + \alpha\beta = X^T y}$$

ب) رابطه هر گام روش GD را برای این روش بنویسید.

با توجه به محاسبات بخش الف رابطه هر گام روش GD برای این روش برابر است با:

$$\nabla f(\beta) = 2X^T X\beta - 2X^T y + 2\alpha\beta \Rightarrow \quad \boldsymbol{\nabla f(\beta) = X^T X\beta - X^T y + \alpha\beta}$$

آرش حریرپوش
9731505

ج) با استفاده از تعریف تابع محدب نشان دهید ترم منظم ساز اضافه شده در این روش یک ترم محدب است.

$$\boldsymbol{f(\beta) = ||X\beta - y||_2^2 + \alpha||\beta||_2^2}$$

همانطور که در صورت سوال نیز گفته شد با توجه به اینکه $||X\beta - y||_2^2$ می باشد؛ برای اثبات محدب بودن تابع بالا نیاز است تا محدب بودن $||\beta||_2^2$ اثبات شود؛ برای اثبات محدب بودن $||\beta||_2^2$ باید اثبات کرد که تابع $||\beta||_2$ محدب است؛ حال با توجه به اینکه در صورتی که همه $||\beta||$ محدب باشد $||\beta||_2$ نیز محدب است به اثبات محدب بودن $||\beta||$ پرداخته می شود:

برای اثبات محدب بودن $||\beta||$ ابتدا به تعریف آن پرداخته می شود:

اگر V یک فضای برداری، $||.||: V \to R$ یک نرم است: $\leftrightarrow$

- $\forall v \in V : ||v|| \geq 0 \; and \; ||v|| = 0 \leftrightarrow v = 0$ (مثبت/قطعی)
- $\forall v \in V, \gamma \in R : |\gamma|||v|| = ||\gamma v||$ (کاملا مقیاس پذیر)
- $\forall v, w \in V : ||v + w|| \leq ||v|| + ||w||$ (نابرابری مثلثی)

حال به تعریف تابع محدب پرداخته می شود:

$f : V \to R$ محدب است در صورتی که: $\leftrightarrow$

$$\forall v, w \in V, \gamma \in [0,1] : f(\gamma v + (1-\gamma)w) \leq \; \gamma \in \; \gamma f(v) + (1-\gamma)f(w)$$

بنابراین با استفاده از قانون نابرابر مثلثی و حقیقتی که نرم کاملا مقیاس پذیر می باشد، می توان اثبات کرد که هر نرمی محدب می باشد:

$$||\gamma v + (1-\gamma)w|| \leq ||\gamma v|| + ||(1-\gamma)w|| = \gamma||v|| + (1-\gamma)||w||$$

بنابراین $||\beta||$ محدب می باشد پس در نتیجه $||\beta||_2$ نیز محدب می باشد بنابراین $||\beta||_2^2$ محدب است.

همچنین به صورت دیگری نیز می توان نشان داد که $||\beta||_2^2$ محدب می باشد؛ با توجه به اینکه

$\nabla^2 \left(||\beta||_2^2 \right) = 2 > 0$ می باشد؛ بنابراین:

$$f\left(\theta||x|| + (1-\theta)||y||\right) \leq \theta f\left(||x||\right) + (1-\theta) f\left(||y||\right) \; for \; \theta \in [0,1].$$

برقرار می باشد به همین دلیل تابع $||\beta||_2^2$ محدب است.

(منابع: https://math.stackexchange.com/questions/2280341/why-is-every-p-norm-convex و https://math.stackexchange.com/questions/546945/prove-convexity-of-squared-euclidean-norm/2120314)

❖ بخش دوم – مسائل برنامه نویسی و پیاده سازی

قسمت اول

فایل csv موجود در پوشه data با نام covid.csv شامل اطلاعات افراد مبتلا به COVID-19 در کره جنوبی می باشد.

1. این فایل را خوانده و در یک جدول نمایش دهید.

نمونه ای از این جدول برابر است با:

	id	sex	birth_year	country	region	infection_reason	infected_by	confirmed_date	state
0	1	female	1984.0	China	filtered at airport	visit to Wuhan	NaN	1/20/2020	released
1	2	male	1964.0	Korea	filtered at airport	visit to Wuhan	NaN	1/24/2020	released
2	3	male	1966.0	Korea	capital area	visit to Wuhan	NaN	1/26/2020	released
3	4	male	1964.0	Korea	capital area	visit to Wuhan	NaN	1/27/2020	released

2. داده ها را با مشاهده سطر و ستون های آن شرح دهید. تعداد داده ها و نام ستون ها را نمایش دهید.

این داده ها اطلاعات افراد مبتلا به COVID-19 را نشان می دهد، در زیر به بررسی هر یک از ستون ها پرداخته می شود:

- id: شماره بیمار در جدول را نمایش می دهد که یک عدد یکتا و افزایشی می باشد.
- sex: جنسیت بیمار را نشان می دهد.
- birth_year: سال تولد بیمار را نشان می دهد.
- country: کشوری که بیمار در آن قرار دارد را نشان می دهد.
- region: منطقه ای که بیمار در آن قرار دارد را نشان می دهد.
- infection_reason: دلیل مبتلا شدن بیمار را نشان می دهد.
- infected_by: این ستون در صورتی که بیمار از فرد دیگری که در جدول قرار داشته باشد بیماری را گرفته باشد، id بیماری که باعث مبتلا شدن این بیمار شده است را نشان می دهد.
- confirmed_date: تاریخی که بیماری فرد مبتلا تایید شده است را نشان می دهد.
- State: وضعیت بیمار را نشان می دهد. (مثلا قرنطینه یا مرخص بودن یا فوت شدن بیمار)

در این فایل هر سطر اطلاعات مربوط به یک بیمار را نشان می دهد؛ و به طور کلی در این فایل تعداد داده ها برابر 176 می باشد.

3. مقادیر max, mean و std را در ستون birth_year به دست آورده و نمایش دهید.

این مقادیر برابر است با:

Mean = 1973.3855421686746

Max = 2009.0

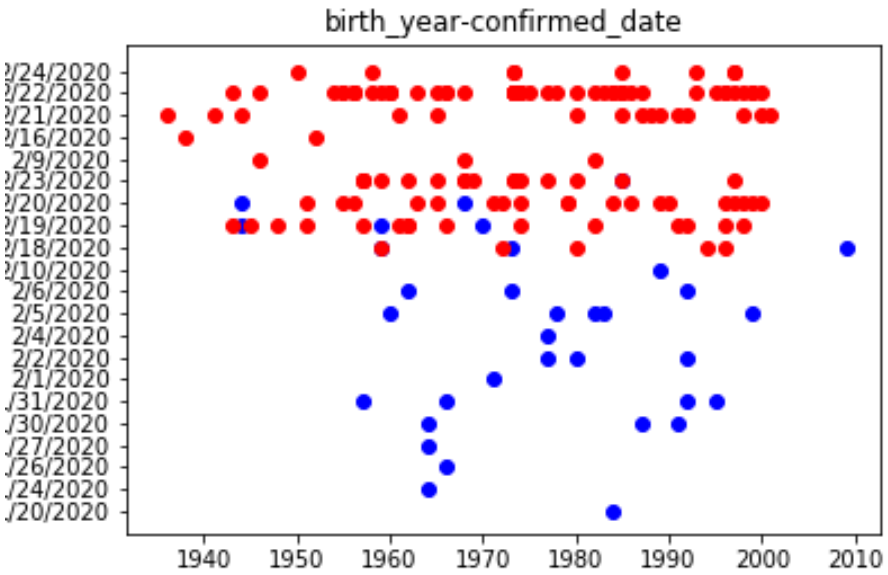
STD = 16.98144368201156

4. بررسی کنید که مقدار null در داده ها وجود دارد یا خیر. در صورت وجود با استفاده از متد مناسب آن را از بین ببرید.

در ستون های region، infected_by، infection_reason و birth_year در داده ها مقدار null وجود دارد؛ مقدار داده های null در ستون birth_year برابر مقدار میانه که در بخش 3 محاسبه شده است قرار داده می شود و بقیه ستون هایی که دارای مقدار null می باشند حذف شده اند.

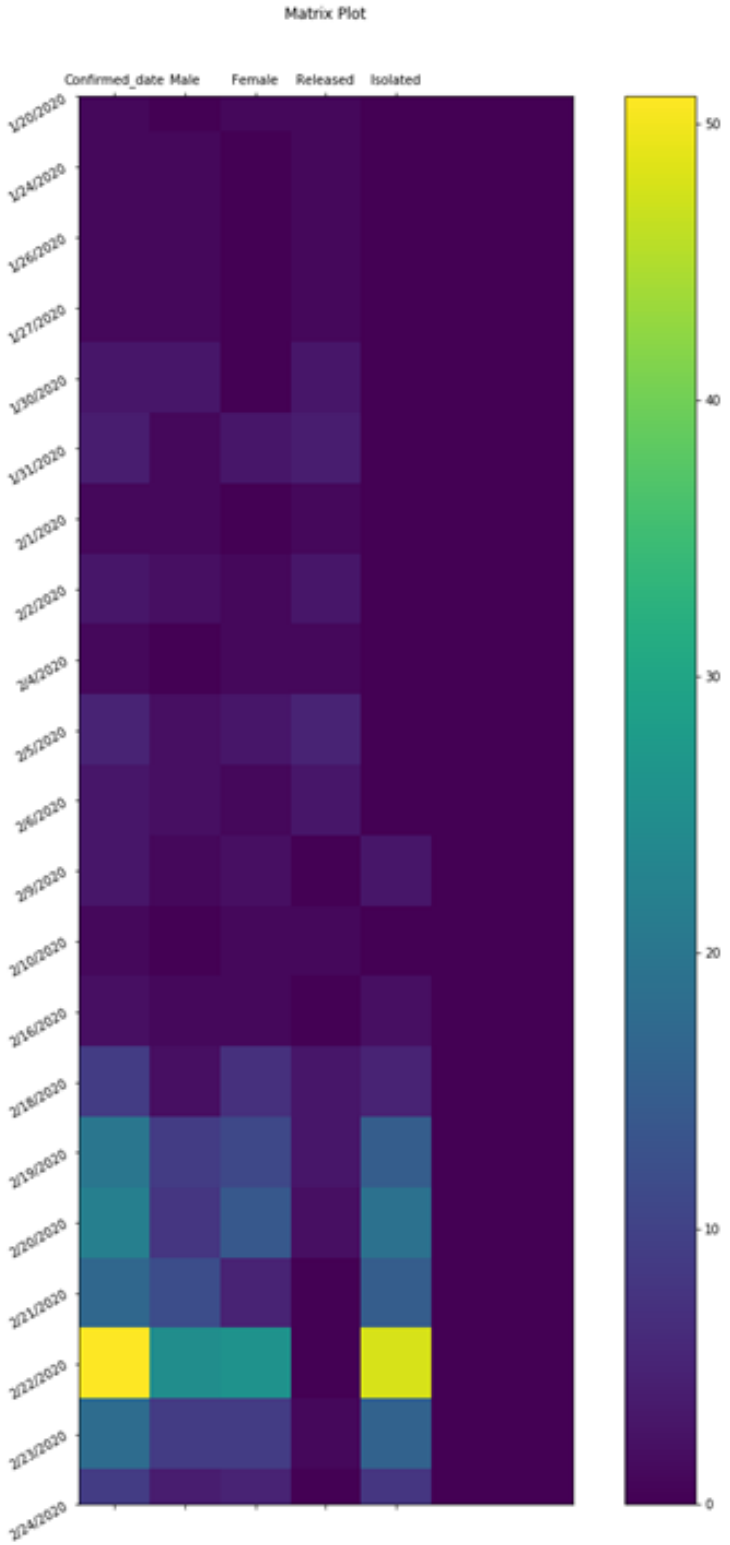
5. در این بخش مصور سازی داده ها را انجام می دهید. با انتخاب ستون مناسب از داده ها، scatter plot, histogram plot و matrix plot را نمایش دهید.

ابتدا scatter plot رسم می شود؛ در این تابع نقاط بر اساس birth_year و confirmed_date ترسیم شده اند نقاط آبی افرادی را نشان می دهد که مرخص شده اند و نقاط قرمز افرادی را نشان می دهد که قرنطینه شده اند:

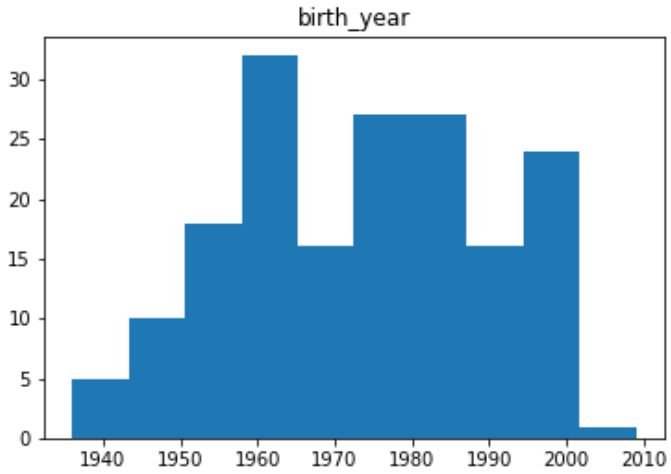


آرش حریرپوش
9731505

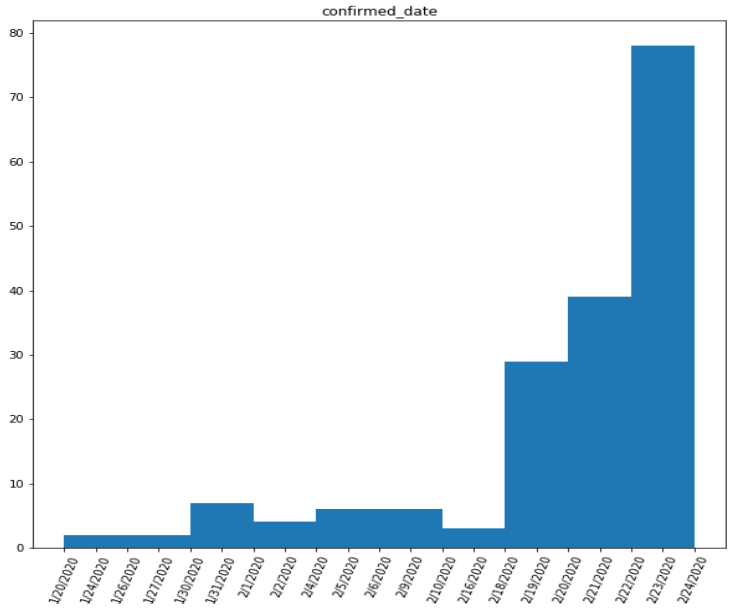
نمودار matrix plot برابر است با (محور y برابر تعداد مبتلایان در یک روز خاص می باشد):



Histogram plot های متفاوتی می تواند رسم شود؛ به عنوان مثال در صورتی که این نمودار برای birth_year رسم شود می توان دید که مبتلایان بیشتر متولدین چه سال هایی هستند؛ شکل زیر این نمودار را نمایش می دهد:



در صورتی که نمودار برای confirmed_date رسم شود؛ نشان می دهد که چه روز هایی مبتلایان بیشتری را به خود اختصاص داده اند؛ شکل زیر این نمودار را نمایش می دهد.



6. بررسی کنید که آیا این مجموعه داده دارای outlier هست یا خیر. در صورت وجود علت خود را بیان کنید و برای روشی برای حل آن ارائه دهید

با توجه به scatter plot رسم شده در سوال 5 این داده ها دارای نویز می باشند اما شامل داده Outlier نمی باشد.

آرش حریرپوش
9731505

قسمت دوم

برای این بخش یک مجموعه داده از تعدادی دانش آموز در پوشهdata با نامstudent.csv قرار دارد. هدف از این قسمت پیش بینی نمره نهایی دانش آموز (G3 attribute) با استفاده از رگرسیون خطی می باشد. اطلاعات مربوط به این مجموعه داده را می توانید در این لینک مشاهده کنید.

داده ها باید به دو بخشtrain و test تقسیم کنید (نسبت تقسیم 80 به 20 باشد و می توانید از متد های آماده استفاده کنید) و روی داده های train رگرسیون خطی انجام دهید. برای سادگی این قسمت فقط ستون هایی که مقادیر عددی دارند را استفاده کنید (در حالت کلی میتوان ستون هایی که مرتبط هستند و مقدار عددی ندارند را به عدد تبدیل کرد).

سپس نمره نهایی را(G3) برای داده های test پیش بینی کنید و دقت(accuracy) مدل آموزش داده شده را به دست آورید

با مطالعه لینک گفته شده در صورت سوال این نتیجه گرفته می شود نمره G3 یک correlation قوی با نمره های G1 و G2 دارد؛ به همین دلیل در این رگرسیون داده های ورودی برابر G1, G2 و داده ای که قرار است پیشبینی شود G3 می باشد.

با انجام رگرسیون بر روی این داده ها دقت مدل در داده های آموزشی برابر تقریبا برابر 0.82 و دقت داده های تست تقریبا برابر 0.79 می باشد.

(کد بخش اول در فایل Visualization.ipynb و بخش دوم در فایل Linear_Regression.ipynb به پیوست ارائه می گردد.)