به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

پاسخ تمرین اول

درس: داده کاوی

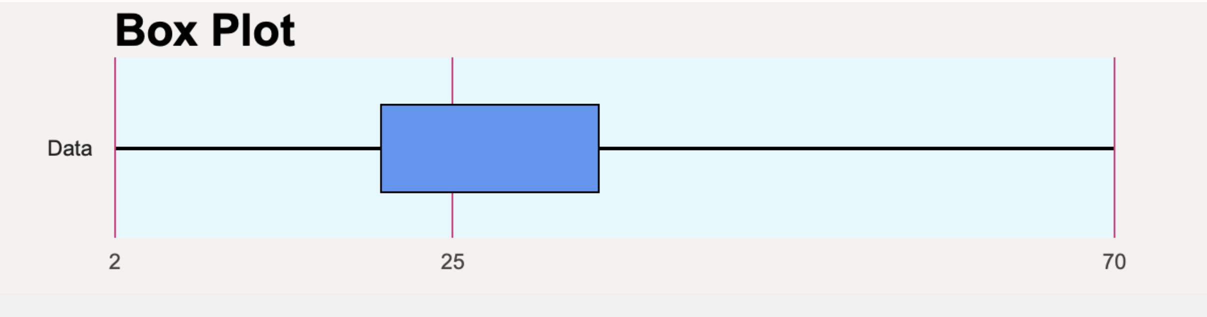
دانشجو: فرشید نوشی – ۹۸۳۱۰۶۸

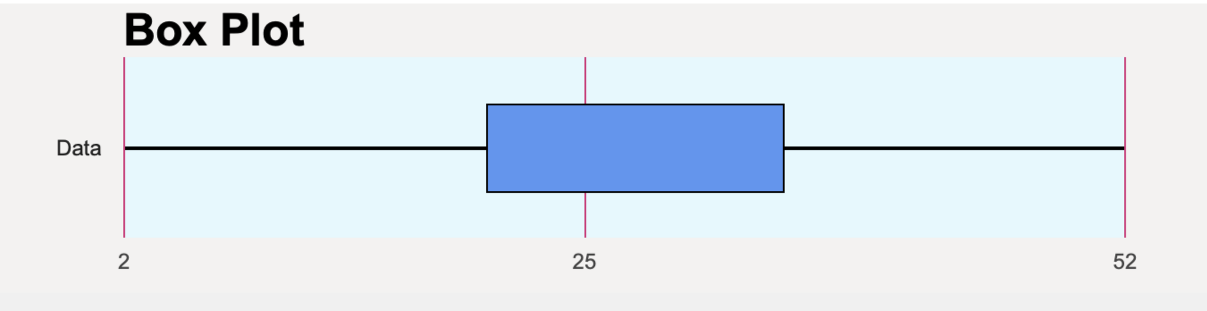
سوال اول

با توجه به اطلاعات اعداد به اطلاعات زیر از داده داده شده میرسیم:

Sample size: 27, Minimum: 2, Q1: 20, Median: 25, Q3: 35, Maximum: 70, Mean: 29.1111, Possible Outliers: 70

نمودار بالا با outliers هست و پایینی بدون outlier. جعبه کشیده شده نیز از Q1 تا Q3 هست(۲۰ تا ۳۵) و خط median نیز در داخل جعبه گذشته است(مقدار ۲۵) که در بالا گزارش شده‌اند.





سوال دوم

داده های نویزی داده های بی معنی هستند. اصطلاح نویزی اغلب به عنوان مترادف برای داده های Corrupted استفاده می‌شود. معنی نویزی شامل هر داده‌ای است که توسط ماشین‌ها به درستی قابل درک و تفسیر نیست. هر داده ای که دریافت، ذخیره یا تغییر داده شده باشد به گونه‌ای که نتواند توسط برنامه‌ای که در ابتدا آن را ایجاد کرده خوانده یا استفاده شود، می تواند به عنوان نویزی توصیف شود. داده های نویزی به طور غیر ضروری میزان فضای ذخیره سازی مورد نیاز را افزایش می‌دهند و می‌توانند بر نتایج تحلیل‌های داده کاوی نیز تأثیر منفی بگذارند.

Outlier یک داده‌ای است که به طور قابل توجهی از بقیه داده‌ها منحرف می‌شود و به شیوه‌ای متفاوت رفتار می‌کند. داده Outlier می تواند ناشی از خطاهای اندازه‌گیری یا ... باشد. این داده‌ها را نمیتوان در یک خوشه یا کلاس معین دسته‌بندی کرد. داده پرت و داده نویزی با یکدیگر متفاوت هستند.

الف)

نویز در ویژگی ها به طور پیش فرض نامطلوب است، زیرا مقادیر ویژگی اصلی را تحریف می کند. داده‌های پرت به طور بالقوه می توانند مقادیر مجاز داده‌ها (یا مقادیر) باشند، به عنوان مثال، شناسایی آنها می تواند هدف اصلی برخی از مسائل داده کاوی باشد. بنابراین، نقاط پرت به طور بالقوه می توانند مطلوب باشند، اما نویز اینطور نیست.

ب)

نویز در مقادیر Attributes می تواند داده ها را تصادفی‌تر یا غیرعادی‌تر به نظر برساند. بنابراین، ممکن است برخی از نمونه ها در داده‌های پر نویزی به صورت Outlier ظاهر شوند.

ج)

داده های نویزدار می توانند به عنوان داده های عادی ظاهر شوند. بنابراین داده‌های نویزی همیشه Outlier نیستند.

سوال سوم

$$Cos(x,y) = \frac{x.y}{\|x\|\|y\|}$$

$$Euclidian(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Jaccard Coefficient

J = number of 11 matches / number of non-zero attributes

$= (f_{11}) / (f_{01} + f_{10} + f_{11})$

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x\ s_y}$$

$$Manhattan\ Distance(x,y) = \sum_{k=1}^{n}|x_k - y_k|$$

الف)

برای این بخش فاصله کسینوسی میان دو بردار برابر با $\frac{2+2+2+2}{\sqrt{16}\sqrt{4}} = 1$ هست. مقدار correlation تعریف نشده است زیرا حاصلش برابر با $\frac{0}{0}$ هست.(چون همه اعداد هر دو بردار یکی هستند کوواریانسشان و انحراف معیارهایشان صفر هست) و فاصله اقلیدسی نیز برابر با $\sqrt{1+1+1+1} = 2$ هست.

ب)

پاسخ این بخش نیز به این صورت است:

$$Cos(x,y) = 0, Correlation(x,y) = -3, Euclidean(x,y) = 2, Jaccard(x,y) = 0$$

فرمول فاصله کسینوسی و اقلیدسی و جکارد در بالا نوشته شده‌اند و تنها با جایگذاری حاصل‌هایشان حساب خواهند شد. در مورد correlation نیز داریم:

$$Mean(x) = 0.5, Mean(y) = 0.5 \rightarrow std(x) = \left[\frac{1}{n-1}\sum_{k=1}^{n}\left(x_i - Mean(x)\right)^2\right]^{0.5} = \frac{1}{\sqrt{3}}$$

$$std(y) = \left[\frac{1}{n-1}\sum_{k=1}^{n}\left(y_i - Mean(y)\right)^2\right]^{0.5} = \frac{1}{\sqrt{3}} \rightarrow$$

$$Cov(x,y) = (-0.5, 0.5, -0.5, 0.5).(0.5, -0.5, 0.5, -0.5) = -1 \rightarrow Correlation(x,y) = \frac{-1}{\frac{1}{\sqrt{3}}\frac{1}{\sqrt{3}}} = -3$$

ج)

در در این بخش با توجه به فرمول نوشته شده نیز فاصله منهتن برابر ۲ هست و به مانند بخش قبل برای حساب correlation عمل میکنیم و داریم:

$$Mean(x) = \frac{2}{3}, Mean(y) = \frac{2}{3} \rightarrow std(x) = \left[\frac{1}{n-1}\sum_{k=1}^{n}\left(x_i - Mean(x)\right)^2\right]^{0.5} = \frac{2}{\sqrt{15}}$$

$$std(y) = \left[\frac{1}{n-1}\sum_{k=1}^{n}\left(y_i - Mean(y)\right)^2\right]^{0.5} = \frac{2}{\sqrt{15}} \rightarrow$$

$$Cov(x,y) = \left(\frac{1}{3},\frac{1}{3},-\frac{2}{3},\frac{1}{3},-\frac{2}{3},\frac{1}{3}\right).\left(\frac{1}{3},\frac{1}{3},\frac{1}{3},-\frac{2}{3},-\frac{2}{3},\frac{1}{3}\right) = \frac{1}{3} \rightarrow Correlation(x,y)$$

$$= \frac{\frac{1}{3}}{\frac{2}{\sqrt{15}}\frac{2}{\sqrt{15}}} = \frac{5}{4} = 1.25$$

د)

در این بخش فاصله کسینوسی میان دو بردار طبق فرمول و پس از جایگذاری برابر با $\cos(x,y) = 0$ خواهد بود. زیرا حاصل ضرب نقطه‌ای دو بردار برابر صفر خواهد بود. برای correlation نیز که برابر صفر هست داریم:

$$Mean(x) = 0, Mean(y) = -\frac{1}{3} \rightarrow std(x) = \left[\frac{1}{n-1}\sum_{k=1}^{n}\left(x_i - Mean(x)\right)^2\right]^{0.5} = \frac{\sqrt{18}}{\sqrt{5}}$$

$$std(y) = \left[\frac{1}{n-1}\sum_{k=1}^{n}\left(y_i - Mean(y)\right)^2\right]^{0.5} = \frac{\sqrt{2}}{\sqrt{3}} \rightarrow$$

$$Cov(x,y) = (2,-1,0,2,0,-3).\left(-\frac{2}{3},\frac{4}{3},-\frac{2}{3},\frac{1}{3},\frac{1}{3},-\frac{2}{3}\right) = 0 \rightarrow Correlation(x,y)$$

$$= \frac{0}{\frac{\sqrt{18}}{\sqrt{5}}\frac{\sqrt{2}}{\sqrt{3}}} = 0$$

سوال چهارم

الف)

مقادیر مجاز فاصله کسینوسی با توجه به اینکه در تعریف برابر مقدار کسینوس زاویه(ابر زاویه) میان دوبردار هست بین یک تا منفی یک میباشد.

ب)

نه لزوما، ممکن است که دو بردار در یک راستا اما با اندازه‌های متفاوت باشند به طور مثال دو بردار:

(1 ,1) و (3 ,3) فاصله کسینوسی ۱ دارند اما یکسان نیستند.

ج)

بله، در صورتی که دو برداری که می‌خواهیم فاصله کسینوسی میانشان را حساب کنیم دارای میانگین صفر باشند فاصله کسینوسی و میزان همبستگی آن‌ها یکسان خواهد بود.

د)

همبستگی صفر نشان می دهد که آماره correlation رابطه‌ای بین دو متغیر را نشان نمی‌دهد. البته این معنی را نمیدهد که اصلاً رابطه ای وجود ندارد. به این معنی است که رابطه خطی وجود ندارد. با این صحبت که این آماره تنها رابطه خطی را بررسی کرده است مشخص است که ممکن است رابطه‌های دیگری که غیرخطی هستند میان دو متغیر وجود داشته باشند و لذا این اطلاعات به تنهایی دو متغیر را مستقل نمیکند.

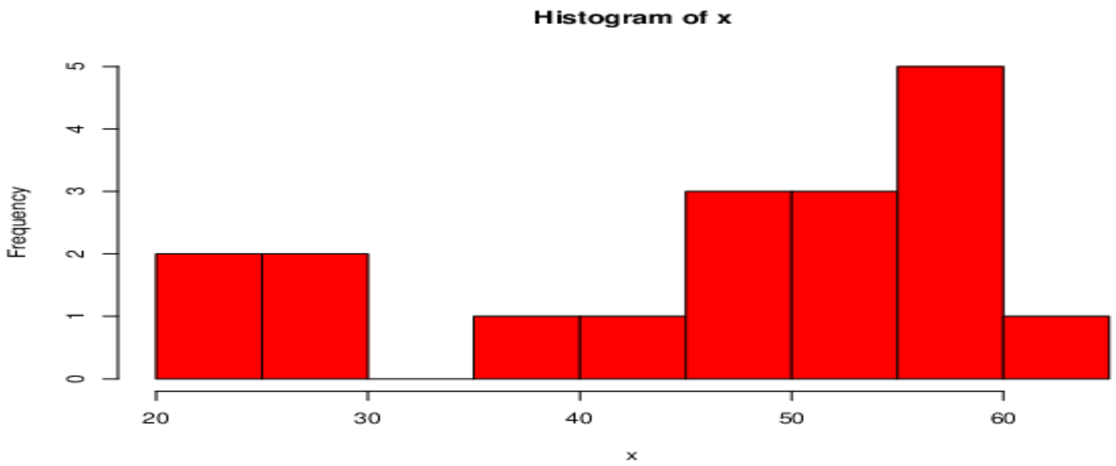
سوال پنجم

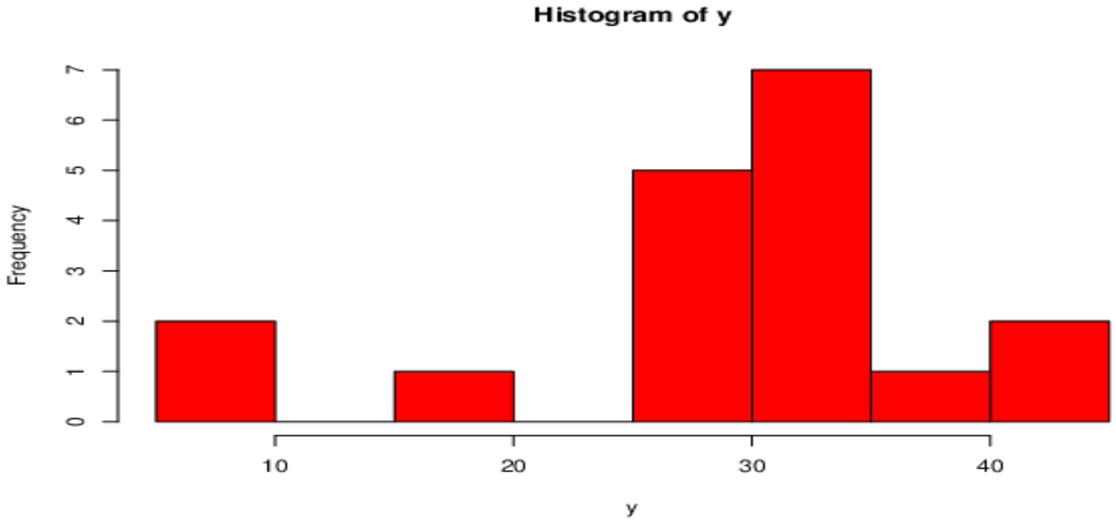
الف)

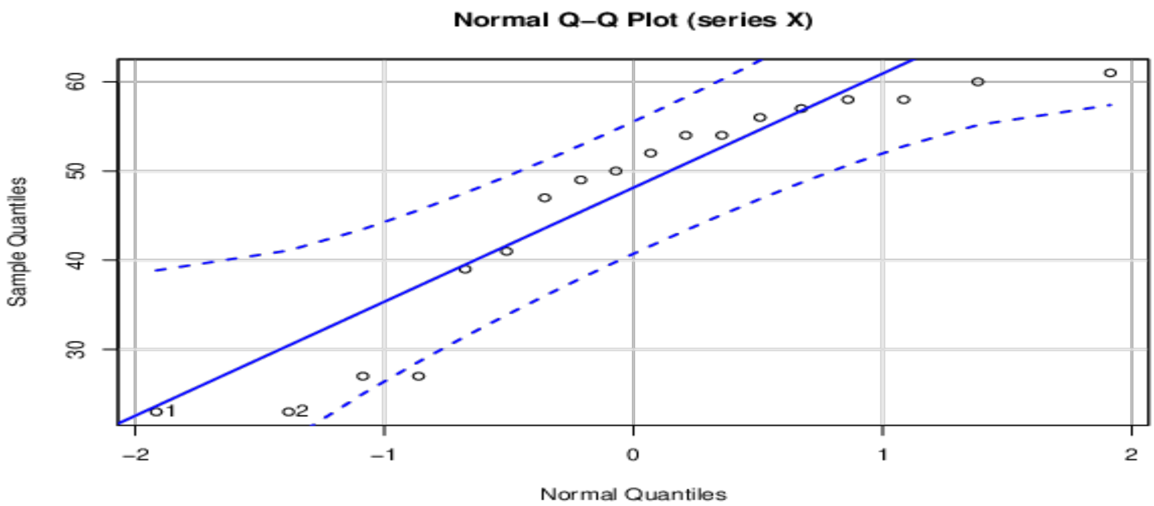
هدف از نمودار quantile-quantile (Q-Q) این است که نشان دهد آیا دو مجموعه داده از یک توزیع آمده اند یا خیر. نمودار با ترسیم Q-Q های مجموعه داده اول در امتداد محور x و ترسیم Q-Qهای مجموعه داده دوم در امتداد محور y ساخته می شود. نمودار Quantile امکان شناسایی هر گونه ویژگی شکل توزیع نمونه مجموعه داده را فراهم می کند، که ممکن است skewed باشد یا انواع دیگری باشند. نمودار quantile توزیع یک مجموعه داده را نمایش می‌دهد درحالیکه نمودار Q-Q توزیع دو مجموعه داده را با یک دیگر مقایسه میکند تا معلوم بکند که از یک توزیع آمده‌اند یا خیر. نمودار Q-Q یک روش گرافیکی است برای اینکه متوجه بشویم که آیا دو نمونه داده از یک توزیع آمده‌اند یا نیامده‌اند. نمودار q-q نموداری از quantile های مجموعه داده اول در برابر quantile های مجموعه داده دوم است.

ب)

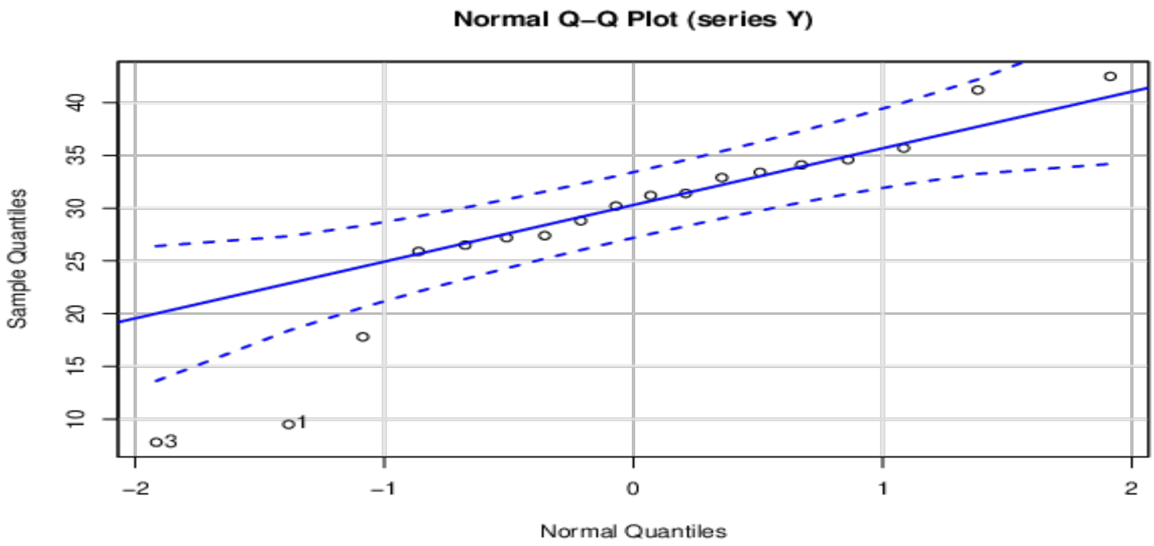
برای این بخش انواع نمودار به مانند نمودار quantile برای هر یک از مقادیر سن و میزان چربی به همراه هیستوگرام‌های هرکدام از این دو مقدار و در نهایت نیز نمودار quantile-quantile مربوط به سن و چربی که همانطور که از نمودار قابل مشاهده است متوجه میشویم که تقریبا رابطه‌ای خطی میان سن و میزان چربی وجود دارد.



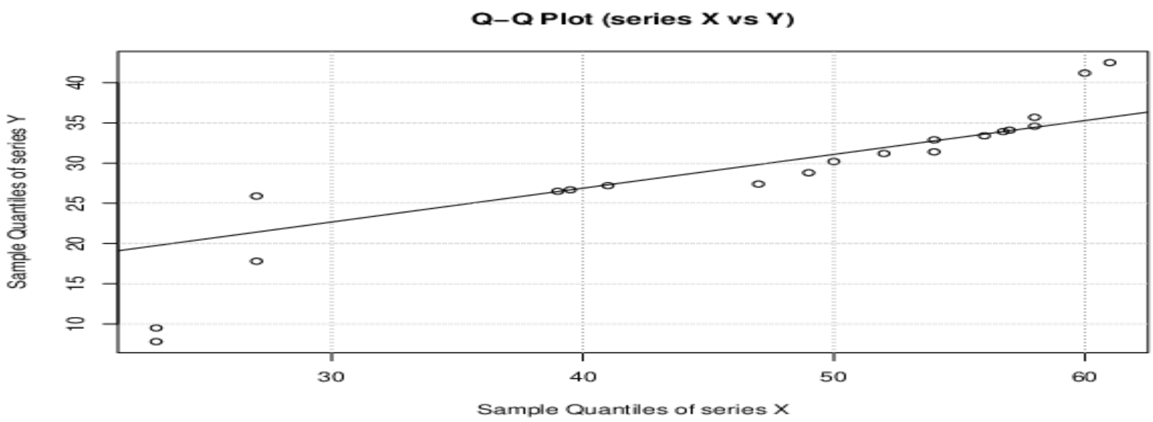




نمودار برای سن(سن = x)



نمودار برای میزان چربی(چربی = y)



نمودار برای سری سن در برابر سری چربی(سن = x و چربی = y)

برای سن ما داریم:

Sample size: 18, Minimum: 23, Q1: 39, Median: 51, Q3: 57, Maximum: 61, IQR: 18, Range: 38

برای چربی نیز ما داریم:

Sample size: 18, Minimum: 7.8, Q1: 26.5, Median: 30.7, Q3: 34.1, Maximum: 42.5, IQR: 7.6, Range: 34.7

سوال ششم

الف)

LOOCV یک مورد یک مورد خاص از اعتبارسنجی Cross-validation است که در آن تعداد فولدها با تعداد نمونه‌های موجود در مجموعه داده برابر است. بنابراین، الگوریتم یادگیری یک بار برای هر نمونه با استفاده از تمام نمونه های دیگر به عنوان یک مجموعه آموزشی و استفاده از نمونه انتخاب شده به عنوان یک مجموعه تست تک موردی اعمال می شود. این فرآیند ارتباط نزدیکی با روش آماری jack-knife estimation دارد.

ب)

در سه مرحله با حذف هر یک از داده‌ها داریم:

X = (1, 3)

Y = (3, 1)

X=

1	1
1	3

Y=

3
1

XB=Y → B=

4
-1

→ x_test, y_test = (1, 1) -> error = 4

X = (1, 3)

Y = (1, 1)

X=

1	1
1	3

Y=

1
1

XB=Y → B=

1
0

→ x_test, y_test = (1, 3) -> error = 4

X = (1, 1)

Y = (1, 3)

X=

1	1
1	1

Y=

1
3

XB=Y → B=

?
?

→ در این دستگاه مقادیر جواب تعریف نشده‌اند و در حاصل جمع قرار نمیگیرند.

LOOCV ->$LOOCV \to MSE = [4 + 4] * \frac{1}{2} = 4$

سوال هفتم

الف)

در رگرسیون خطی، overfitting زمانی رخ می دهد که مدل بیش از حد پیچیده باشد. این معمولا زمانی اتفاق می افتد که پارامترهای زیادی در مقایسه با تعداد مشاهدات وجود داشته باشد. چنین مدلی به خوبی به داده های جدید تعمیم پیدا نمی‌کند. یعنی در داده های آموزشی عملکرد خوبی خواهد داشت اما در داده های تست ضعیف است.

ب)

خیر اینکار ضروری نیست. در واقع حذف موارد پرت وسوسه انگیز است. این کار را بدون دلیل خیلی خوب انجام نباید داد. مدل‌هایی که موارد outlier را نادیده می‌گیرند، اغلب عملکرد ضعیفی دارند. به عنوان مثال، اگر یک شرکت مالی بزرگ‌ترین نوسانات بازار(outliers) را نادیده بگیرد، با انجام سرمایه‌گذاری‌های ضعیف ورشکسته می‌شوند.

روش های تشخیص بیرونی عبارتند از:

تک متغیره -> boxplot که در آن خارج از محدوده ۱.۵ برابری IRQ یک نقطه outlier است.

دو متغیره -> scatterplot with confidence ellipse مثلاً، خارج از confidence ellipse ۹۵ درصد، یک outlier است.

چند متغیره -> فاصله Mahalanobis D2

ج)

$$residual = \epsilon = y - X\beta$$

با توجه به اینکه انحراف باقیمانده رابطه‌اش طبق اسلایدها رابطه بالا می‌باشد. کم بودن این مقدار لزوما مطلوب نمیباشد. زیرا انحرافات میتوانند مقادیر مثبت و منفی داشته باشند در حاصل جمع residualها این اعداد یکدیگر را میتوانند خنثی کنند. با اینحال اما حاصل جمع مربعات residualها درصورتی که کم باشد مطلوب ما می‌باشد.

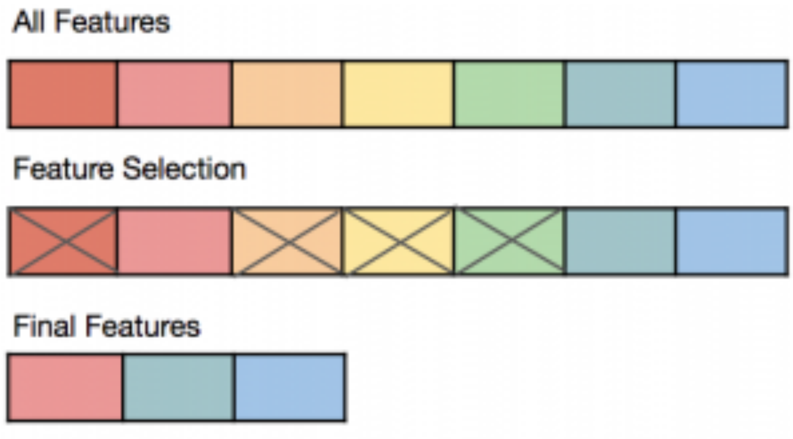
د)

یکی از موارد قابل بحث پیچیدگی محاسباتی بالای روش مستقیم در مقابل روش گرادیان کاهشی میباشد در روش نرمال ما نیازی به تعیین نرخ یادگیری نیستیم و یک روش analytical هست و روش گرادیان یک روش iterative هست. روش گرادیان با دیتاست‌های بزرگ و تعداد فیچر بالا به خوبی کار میکند. در روش نرمال ما feature scaling نیازی نداریم ولی در این روش نیاز به بررسی حالاتی داریم که ماتریسمان معکوس پذیر نباشد.

سوال هشتم

الف)

روش انتخاب ویژگی به مانند تصویر زیر میباشد که در آن برخی ویژگی‌ها حذف می‌شوند



انتخاب ویژگی در مورد انتخاب زیرمجموعه ای از ویژگی ها از ویژگی های اصلی به منظور کاهش پیچیدگی مدل، افزایش کارایی محاسباتی مدل ها و کاهش خطای تعمیم ایجاد شده به دلیل نویز توسط ویژگی های نامربوط است. استخراج ویژگی در مورد استخراج اطلاعات از ویژگی های اصلی تنظیم شده برای ایجاد یک زیرفضای ویژگی جدید است. ایده اصلی پشت استخراج ویژگی فشرده سازی داده ها با هدف حفظ بیشتر اطلاعات مربوطه است. همانند تکنیک‌های انتخاب ویژگی، این تکنیک‌ها نیز برای کاهش تعداد ویژگی‌ها از ویژگی‌های اصلی برای کاهش پیچیدگی مدل، و برازش بیش از حد مدل، افزایش کارایی محاسبات مدل و کاهش خطای تعمیم استفاده می‌شوند.

ب)

در این بخش سه تکنیک را به اختصار شرح میدهیم.

PCA:

ویژگی‌های جدید (مولفه‌های اصلی) را پیدا می‌کند که حداکثر میزان تغییرات را در داده‌ها ثبت می‌کند

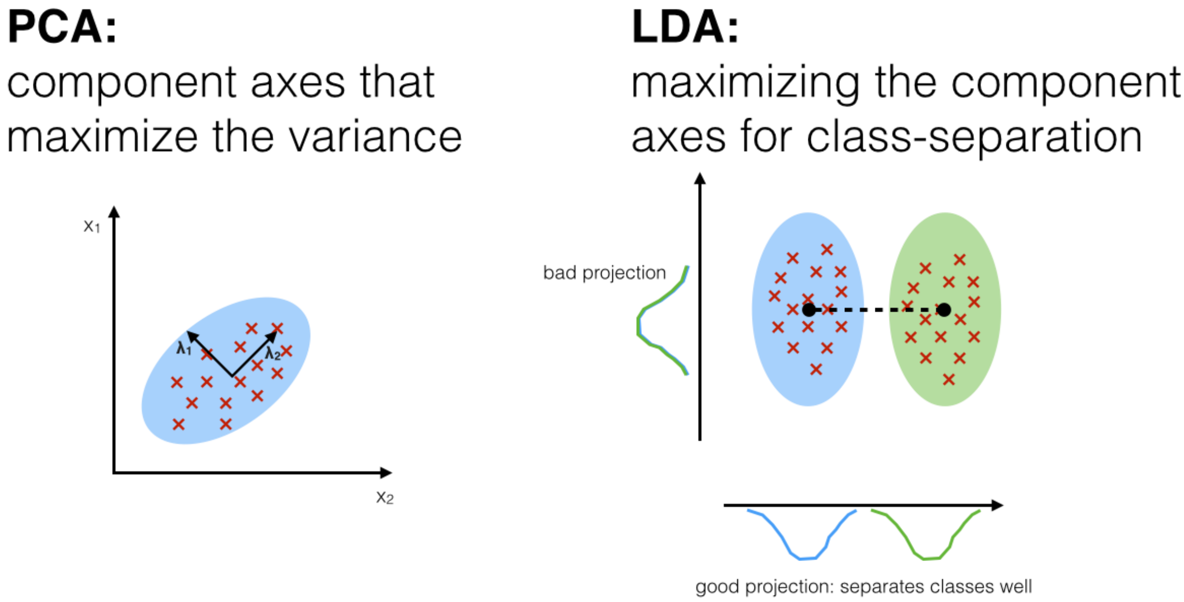
LDA:

LDA به دنبال جداسازی (یا تفکیک) نمونه ها در مجموعه داده آموزشی بر اساس ارزش کلاس آنها است. به طور خاص، این مدل به دنبال یافتن ترکیبی خطی از متغیرهای ورودی است که به حداکثر تفکیک برای نمونه‌ها بین کلاس‌ها (کلاس مرکز یا میانگین) و حداقل جداسازی نمونه‌ها در هر کلاس دست می‌یابد. در زیر یک تصویر برای بیشتر نشان دادن تفاوت‌های این دو روش آورده شده است.

Autoencoder:

autoencoderها نوع خاصی از معماری یادگیری عمیق هستند که برای یادگیری data visualization، معمولاً به منظور کاهش ابعاد استفاده می شوند. این هدف با این روش بدست می آید که در معماری آنها هدف کپی کردن لایه ورودی در لایه خروجی آن خواهد بود. در این معماری در وسط لایه‌های عصبی یک

بازنمایی از داده بدست خواهد آمد که بازنمایی کاهش داده شده از داده خواهد بود و از آن برای تبدیل داده ها به یک فضای کوچک تر استفاده میشود. فرق این روش با روش‌های دیگر در توانایی رمزگذارهای خودکار در انجام تبدیل‌های غیرخطی برروی داده می‌باشد.



سوال نهم

الف)

$$\text{Entropy} = -\sum_{j} p(j|t)\log p(j|t)$$

برای مهارت داریم:

$$ans = -\frac{4}{10} * \log(0.4) - \frac{6}{10} * \log(0.6)$$

برای ژانر نیز خواهیم داشت:

$$ans = -\frac{4}{10} * \log(0.4) - \frac{3}{10} * \log(0.3) - \frac{3}{10} * \log(0.3)$$

ب)

خواهیم داشت:

$$H(Genre) = -0.4 * \log(0.4) - 0.3 * \log(0.3) - 0.3 * \log(0.3)$$

$$H(Ability) = -0.4 * \log(0.4) - 0.6 * \log(0.6)$$

$$H(Genre, Ability) = 0.1\log(0.1) - 0.3\log(0.3) - 0.1\log(0.1) - 0.2\log(0.2) - 0.2\log(0.2) - 0.1\log(0.1)$$

$$Mutual\ Information = H(Genre) + H(Ability) - H(Genre, Ability)$$

سوال دهم

الف)

در این روش منظم‌سازی اگر هایپرپارامتر مربوط به ضریب وزن‌های مدل را زیاد کنیم وزن‌های مدل در بهینه‌سازی کوچکتر و به صفر نزدیک‌تر خواهند شد. با این صحبت در حالتی که مقدار هایپرپارامتر برابر یک هست وزن‌ها کوچکتر از حالتی هستند که هایپرپارامتر برابر صفر هست در نتیجه:

$$\lambda = 0 \rightarrow \theta = \begin{bmatrix} 71.9 \\ 44.42 \end{bmatrix}$$

$$\lambda = 1 \rightarrow \theta = \begin{bmatrix} 1.21 \\ 0.57 \end{bmatrix}$$

ب)

نمودار اول و دوم میتوانند اما نمودار سوم ممکن نیست که مربوط به lasso regulatization باشد.

در واقع، با افزایش لامبدا، برخی از وزن‌های پارامترها که در نقطه‌ای به صفر رسیده‌اند ممکن است (به طور موقت) به مقادیر غیرصفر افزایش یابند که ممکن است که feature interaction باعث این رفتار شود.

همانطور که لامبدا به سمت بی نهایت می رود، همه وزن ها در نهایت صفر خواهند شد.