نام و نام خانودگی: فهیمه چرخاب – صدیقه سادات غنی(40308864)

تکلیف فصل دوم بخش دوم کاهش ابعاد

استاد محترم: جناب آقای دکتر داورپناه جزی

سوال یک : ارائه یک نمونه جدید و اجرای مراحل کامل آن مشابه مثال اسلایدهای 42 و 43

برای بررسی صفات خاصه و کاهش ابعاد با استفاده از تکنیک انتخاب ویژگی، می‌توانیم یک نمونه جدیدی ارائه دهیم که شامل دو ویژگی Xو Y و دو کلاس دسته بندی شده A و Bباشد مراحل کار به شرح زیر است:

1- تعریف داده‌ها

نخست، داده‌های آزمایشی با دو ویژگی Xو Yو برچسبهای کلاس تولید می‌کنیم/ فرض کنید داده‌ها به شکل زیر هستند

X	Y	کلاس
2	3	A
3	4	A
5	6	A
8	7	B
9	8	B
10	9	B

2- محاسبه میانگین و واریانس

برای هر ویژگی، میانگین و واریانس را برای کلاس‌های Aو Bمحاسبه می‌کنیم.

ویژگی :X

میانگین کلاس A :

$$\text{MEAN(A)}=\frac{2+3+5}{3}=\frac{10}{3}\approx 3/33$$

میانگین کلاس :B

$$\text{MEAN(B)}=\frac{8+9+10}{3}=\frac{27}{3}=9$$

واریانس کلاسA

$$\text{VAR(A)}=\frac{(2-3/33)^2+(3-3/33)^2+(5-3/33)^2}{3}\approx 2/33$$

واریانس کلاس B

$$\text{VAR(B)}=\frac{(8-9)^2+(9-9)^2+(10-9)^2}{3}\approx 0/67$$

ویژگی Y

میانگین کلاس A

$$\text{MEAN(A)}=\frac{3+4+6}{3}=\frac{13}{3}\approx 4/33$$

میانگین کلاس B

$$\text{MEAN(B)}=\frac{7+8+9}{3}=\frac{24}{3}=8$$

واریانس کلاس A

$$\text{VAR(A)} = \frac{(3-4/33)^2+(4-4/33)^2+(6-4/33)^2}{3} \approx 2/33$$

واریانس کلاس B

$$\text{VAR(B)} = \frac{(7-8)^2+(8-8)^2+(9-8)^2}{3} \approx 0/68$$

3- محاسبه مقدار انتخاب ویژگی

برای هر ویژگی، مقدار $\frac{|\text{MEAm(A)}-\text{Mean(B)}|}{\sqrt{\frac{\text{var(A)}}{\text{N1}}-\frac{\text{var(B)}}{\text{N2}}}}$ **را محاسبه کرده و آن را با مقدار آستانه مقایسه می‌کنیم، که فرض می‌کنیم مقدار آستانه 1 است.**

که در آن N1و N2 به تعداد نمونه‌ها در کلاس‌های A و B اشاره دارند.

ویژگی :x

N1=3,N2=3

$$\text{GH Value} = \frac{|3/33 - 9|}{\sqrt{\frac{2/33}{3} - \frac{0/67}{3}}} = \frac{5/67}{\sqrt{1}} = 5/67$$

ویژگی :y

$$\text{GH Value} = \frac{|4/33 - 8|}{\sqrt{\frac{2/33}{3} - \frac{0/67}{3}}} = \frac{3/67}{\sqrt{1}} = 3/67$$

4- انتخاب ویژگی‌ها

مقدار محاسبه‌شده برای هر ویژگی بالاتر از آستانه است. بنابراین، هر دو ویژگی X و Y به عنوان ویژگی‌های انتخاب شده باقی می‌مانند.

با استفاده از میانگین، واریانس و معیار انتخاب ویژگی، قادر به شناسایی ویژگی‌های مؤثر برای کاهش ابعاد و دسته‌بندی داده‌ها هستیم. این روش به ما کمک می‌کند تا ویژگی‌هایی که توانایی تفکیک کلاس‌های مختلف را دارند شناسایی کنیم.

سوال دو: ارائه یک نمونه اجرای مراحل کامل آنتروپی اسلایدهای 45 و 46

1-محاسبه ماتریس شباهت S_{IJ} :

روش 1) $e^{-\alpha D_{IJ}} = SIJ$ (با دقت به اینکه D_{IJ} فاصله بین دو ویژگی است)

روش دوم:

$$S_{IJ}=\frac{(\sum_{K=1}^{N}|X_{IK}=X_{JK}|)}{N}$$

2-محاسبه فاصله بین ویژگیD_{IJ}

$$D_{IJ}=\left[\sum_{K=1}^{N}\frac{(X_{IK-}X_{JK})}{MAX_K-MIN_K)^2}\right]\frac{1}{2}$$

$$S_{IJ}=\frac{(\sum_{K=1}^{N}|X_{IK}=X_{JK}|)}{N}$$

3-محاسبه آنتروپی

استفاده از S_{IJ}برای بدست آوردن آنتروپی

$$ENTROPY=-\sum_{I=1}^{N-1}\sum_{J=I+1}^{N}\left[S_{IJ\times}\log SIJ+(1-S_{IJ})\times\log 1-SIJ)\right]$$

1-تعریف ویژگی

ویژگی1	ویژگی 2
1	2

3	2
4	3

2- محاسبه فاصله D_{IJ}

محاسبه MAX و MIN برای هر ویژگی

برای ویژگی1 MIN = 1 ، MAX = 3

برای ویژگی2 MIN = 2 ، MAX = 4

3-محاسبه ماتریس فاصله D_{IJ}

$$D_{12}=\sqrt{\frac{(1-2)^2}{3-1}+\frac{(2-3)^2}{4-2}}=\sqrt{0/5+0/5}=\sqrt{1}$$

$$D_{13}=\sqrt{\frac{(1-2)^2}{3-1}+\frac{(2-3)^2}{4-2}}=\sqrt{1+1}=\sqrt{2}$$

$$D_{23}=\sqrt{\frac{(2-3)^2}{3-1}+\frac{(3-4)^2}{4-2}}=\sqrt{0/5+0/5}=\sqrt{1}$$

4-محاسبه ماتریس شباهتS_{IJ}

استفاده از $\propto = 1$:

$$S_{12}=e^{-D_{12}}=e^{-1}\approx 0/3679$$

$S_{13} = e^{-D_{13}} = e^{-\sqrt{2}} \approx 0/2430$

$S_{23} = e^{-D_{23}} = e^{-1} \approx 0/3679$

5- محاسبه آنتروپی

ENTROPY=−(S12×log(S12)+(1−S12)×log(1−S12)+S13×log(S13)+(1−S13)×log(1−S13)+S23×log(S23)+(1−S23)×log(1−S23))

log(S12)=log(0/3679)≈−1/0003

log(1−S12)=log(1−0/3679)=log(0/6321)≈−0/4605

S12×log(S12)+(1−S12)×log(1−S12)≈0/3679× (−1/0003)+0/6321× (−0/4605)≈−0/3683−0/2911≈−0/6594

برایS13:

log(S13)=log(0/2430)≈−1/3924

log(1−S13)=log(1−0/2430)=log(0/7570)≈−0/2845

S13·log(S13)+(1−S13)×log(1−S13)≈0/2430× (−1/3924)+0/7570× (−0/2845)≈−0/3384−0/2152≈−0/5536

برایS_{23}

log (S23)=log (0/3679)≈−1/0003

log(1−S23)=log(1−0/3679)=log (0/6321)≈−0/4605

S23×log(S23)+(1−S23)×log(1−S23)≈0/3679× (−1/0003)+0/6321× (−0/4605)≈−0/3683−0/2911≈−0/6594

ENTROPY=−(−0/6594−0/5536−0/6594) ENTROPY≈−(−1/8724)≈1/8724

سوال سوم: ارائه یک نمونه اجرای مراحل کامل PCA اسلایدهای 47 و 48

$$C \times E_{error=} \lambda \times E_{ERROR}$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$\lambda 1 > \lambda 2 > \lambda 3 \dots \dots > \lambda N - 1 > \lambda N$$

$$\frac{\sum_{I=1}^{K} \lambda I}{\sum_{I=1}^{K} \lambda I} \geq 0.95$$

$$DATA_{NEW} = VECTOR^{T} \times DATA_{ORGINAL}^{T}$$

فرض کنید ما یک ماتریس داده Xداریم که شامل mنمونه و nویژگی است.

مراحل اجرایPCA

1-داده‌های اصلی

ابتدا یک ماتریس داده‌های اصلی Xرا در نظر می‌گیریم (به عنوان نمونه):

$$X=\begin{vmatrix}2 & 3\\2 & 1\\3 & 2\end{vmatrix}$$

اینجا فرض می‌کنیم که هر ردیف یک نمونه و هر ستون یک ویژگی است.

2-مرکز کردن داده‌ها:

برای مرکز کردن داده‌ها، میانگین هر ویژگی را محاسبه و آن را از داده‌ها کم می‌کنیم

$$\text{میانگین}=\begin{bmatrix}\text{اول ستونی میانگین}\\\text{دوم ستونی میانگین}\end{bmatrix}=\begin{bmatrix}2/33\\2\end{bmatrix}$$

بعد از مرکز کردن، ماتریس جدید به دست می‌آید

$$X_{centered}=X-\text{میانگین}=\begin{bmatrix}2-2/33 & 3-2\\2-2/33 & 1-2\\3-2/33 & 2-2\end{bmatrix}=\begin{bmatrix}-0/33 & 1\\-0/33 & -1\\0/67 & 0\end{bmatrix}$$

3-محاسبه ماتریس کوواریانس

ماتریس کوواریانس را با استفاده از داده‌های مرکز شده محاسبه می‌کنیم

$$C=\frac{1}{m-1}\left(X_{CENTRED}^{T}\times X_{CENTRED}\right)$$

$$C=\frac{1}{2}\begin{bmatrix}(-0/33) & (-0/33) & (0/67)\\ 1 & -1 & 0\end{bmatrix}\times\begin{bmatrix}-0/33 & 1\\ -0/33 & -1\\ 0/67 & 0\end{bmatrix}=\begin{bmatrix}0/3333 & 0\\ 0 & 1\end{bmatrix}$$

4-محاسبه مقادیر ویژه و بردارهای ویژه:

برای محاسبه مقادیر و بردارهای ویژه، ما باید معادله زیر را حل کنیم

$$\det(\mathbf{C}-\lambda\mathbf{I})=0$$

که به ما مقادیر ویژه $\lambda1,\lambda2$را می‌دهد. فرض کنید که مقادیر ویژه به شکل زیر است:

$$\lambda1=1/5,\ \ \lambda2=0/5$$

5-رتبه بندی مقادیر ویژه:

طبق خواصPCA

$$\lambda1>\lambda2$$

6-نسبت واریانس تجمعی:

حال می‌خواهیم نسبت واریانس تجمعی را محاسبه کنیم

$$\frac{\sum_{I=1}^{K}\lambda I}{\sum_{I=1}^{K}\lambda I}\geq 0.95$$

مجموع مقادیر ویژه=λ1+λ2=1/5+0/5=2

k=1 **نسبت واریانس تجمعی برای**$=\frac{1/5}{2}=0/75$(less than 0.95)

K=2 **نسبت واریانس تجمعی برای**$=\frac{2}{2}=1$ (greater than 0.95)

7-انتخاب مؤلفه‌ها:

از آنجا که برای k=2 **با نسبت واریانس تجمعی بیش از** 0.95 **رضایت داریم، باید هر دو مؤلفه اصلی را نگه داریم.**

8-داده‌های جدید:

حال می‌توانیم داده‌های جدید را با استفاده از فرمول زیر محاسبه کنیم

$$\times \mathrm{DATA}_{\mathrm{ORGINAL}}^{\mathrm{T}}\mathrm{DATA}_{\mathrm{NEW}} = \mathrm{VECTOR}^{\mathrm{T}}$$

فرض کنیم ماتریس بردارهای ویژه به شکل زیر باشد:

$$\mathrm{VECTOR}^{\mathrm{T}}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

$$=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \times \begin{bmatrix}2 & 3\\ 2 & 1\\ 3 & 2\end{bmatrix}^{\mathrm{T}} \mathrm{DATA}_{\mathrm{NEW}}$$

$$\times \mathrm{DATA}_{\mathrm{ORGINAL}}^{\mathrm{T}}\mathrm{DATA}_{\mathrm{NEW}} = \mathrm{VECTOR}^{\mathrm{T}}$$

ترانهاده ماتریس VECTOR

$$\mathrm{VECTOR}^{\mathrm{T}}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{T} =\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

ترانهاده ماتریس $DATA_{ORGINAL}$

$$DATA^{T}_{ORGINAL}=\begin{bmatrix}2 & 3\\ 2 & 1\\ 3 & 2\end{bmatrix}^{T}=\begin{bmatrix}2 & 2 & 3\\ 3 & 1 & 2\end{bmatrix}$$

محاسبه $DATA_{NEW}$

$$DATA_{NEW}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\times\begin{bmatrix}2 & 2 & 3\\ 3 & 1 & 2\end{bmatrix}^{T}$$

$[1\ \ 0]\times[2\ \ 3]=1\times2+0\times3=2$
$[1\ 0]\times[2\ 1]=1\times2+0\times1=2$
$[1\ \ 0]\times[3\ \ 2]=1\times3+0\times2=3$
$[0\ 1]\times[2\ 3]=0\times2+1\times3=3$
$[0\ 1]\times[2\ 1]=0\times2+1\times1=1$
$[0\ 1]\times[3\ 2]=0\times3+1\times2=2$

تشکیل ماتریس $DATA_{NEW}$

$$DATA_{NEW}=\begin{bmatrix}2 & 2 & 3\\ 3 & 1 & 2\end{bmatrix}$$

که به ما داده‌های جدید را می‌دهد.از طریق مراحل بالا تحلیل PCA روی داده‌ها انجام شد. این مراحل شامل مرکز کردن داده‌ها، محاسبه ماتریس کوواریانس، محاسبه مقادیر ویژه و بردارهای ویژه، و نهایتاً محاسبه داده‌های جدید است. از این روش می‌توان برای کاهش ابعاد داده‌ها استفاده کرد و ویژگی‌های کلیدی آن‌ها را استخراج نمود.

سوال چهارم :انجام کلیه مراحل برای مثال کای اسکویر اسلاید 51 تا رسیدن به نتیجه اعلام شده

$$x^2 = \sum_{I=1}^{2} \sum_{J=1}^{K} (A_{IJ} - E_{IJ})^2 / E_{IJ}$$

$$= (R_I \times C_J) / NE_{IJ}$$

1-تعریف داده ها

2- مرتب‌سازی داده‌ها

داده‌ها را بر اساس مقادیر مرتب می‌کنیم. در اینجا، داده‌ها از 1 تا 8 مرتب است.

کلاس	مقدار(Att)
A	1
A	1
A	2
A	3
B	3
B	4
B	4

B	5
B	5
A	6
A	7
B	8

3-تعریف بازه‌های اولیه

برای این مثال، فرض می‌کنیم که ما دو بازه (Interval) داریم:

بازه 1: [1,4]

بازه 2:[8 ,5]

4- محاسبه فراوانی در بازه‌ها: تعداد نمونه‌ها در هر بازه و کلاس را محاسبه می‌کنیم.

در بازه اول [1,4]

کلاس A: 5 **مقادیر** 1, 1, 2, 3

کلاسB: 3 **مقادیر** 3, 4, 4

در بازه دوم :[5,8]

کلاس 3 :A **مقادیر** 6, 7

کلاس 3 :B **مقادیر** 5, 5, 8

5–جدول فراوانی(Observed Frequency)

بازه	کلاس(A_ij) A	کلاس(B_ij) B
[1, 4]	5	3
[5, 8]	3	3

6–محاسبه فراوانی‌های مورد انتظار E_{ij}

محاسبه جمع کل مشاهدات

جمع کل مشاهدات = 11 (5 + 3 + 3 + 3)

محاسبه جمع مشاهدات در هر کلاس:

جمع کلاس(3 + 5) 8 = A

جمع کلاس(3 + 3) 3 = B

7-مراحل محاسبه فراوانی مورد انتظار(Expected Frequency)

$=(R_I \times C_J)/NE_{IJ}$

R_I**مجموع تعداد نمونه‌ها در بازه** **i**

C_J**تعداد نمونه‌ها در کلاس** **j**

N تعداد کل نمونه‌ها

برای بازه 1 **و کلاس** A

$$E_{11} = \frac{8 \times 8}{11} \approx 4/57$$

برای بازه 1 **و کلاس** B

$$E_{12} = \frac{8 \times 3}{11} \approx 3/43$$

برای بازه 2 و کلاسA

$$\mathbf{E_{21} = \frac{3 \times 8}{11} \approx 2/18}$$

برای بازه 2 و کلاسB

$$\mathbf{E_{22} = \frac{3 \times 3}{11} \approx 0/82}$$

8-تشکیل جدول نهایی

جدول نهایی را تشکیل می‌دهیم که شامل فراوانی مشاهداتی و مورد انتظار است:

بازه	کلاسA (Aij)	کلاسB (Bij)	E11	E12	E21	E22
[1, 4]	5	3	4.57	3.43		
[5, 8]	3	3	3.43	2.57	**2/18**	**0/82**

9-محاسبه آماره کای اسکویر

$$x^2 = \sum_{I=1}^{2} \sum_{J=1}^{K} (A_{IJ} - E_{IJ})^2 / E_{IJ}$$

1-برای $[1,4]$ و کلاس A

$$\frac{(5-4/57)^2}{4/57}=0/042$$

2-برای $[4, 1]$ و کلاس B

$$\frac{(3-3/43)^2}{3/43}=0/054$$

برای $[8, 5]$ و کلاس A

$$\frac{(3-3/43)^2}{3/43}=0/054$$

برای $[8, 5]$ و کلاس B

$$\frac{(3-2/57)^2}{2/57}=0/078$$

$\chi 2 \approx 0/042+0/054+0/054+0/078 \approx 0/228$

آمار کای اسکویر (Chi-Square) ما تقریباً برابر با 0.228**است. برای اینکه بفهمیم آیا این مقدار به صورت معنی‌داری متفاوت از آنچه در یک توزیع تصادفی انتظار می‌رود، هست یا خیر، نیاز به مقایسه این مقدار با مقادیر بحرانی موجود در جدول کای اسکویر داریم.**

مراحل بررسی معنی‌داری:

درجات آزادی:برای محاسبه درجات آزادی (Degrees of Freedom) در یک تحلیل کاای اسکویر از فرمول زیر استفاده می‌کنیم:

$$\text{درجات آزادی} = (r-1) \times (c-1)$$

که در آن rrتعداد ردیف‌ها و ccتعداد ستون‌ها در جدول است. در این مثال، چون ما دو ردیف و دو ستون داریم:

$$(r-1) \times (c-1) = (2-1)(2-1) = 1/1 = 1$$

بنابراین، درجات آزادی ما برابر با 1است.

استفاده از جدول کای اسکویر:

حال با توجه به مقدار آمار کاای اسکویر (0.228) و درجات آزادی (1)، می‌توانیم به جدول کاای اسکویر مراجعه کنیم تا مقدار بحرانی را برای سطح معنی‌داری مورد نظر (معمولاً 0.05 یا 0.01) پیدا کنیم.

برای مثال، برای سطح معنی‌داری 0/05، مقدار بحرانی برای 1 درجه آزادی تقریباً 3/841است.

اگر مقدار محاسبه شده (0.228) کمتر از مقدار بحرانی (3.841) باشد، می‌توانیم نتیجه‌گیری کنیم که نتایج ما به طور معنی‌داری متفاوت نیستند و فرضیه صفر را نمی‌توان رد کرد.

در این مورد، چون 3/841 > 0/228، می‌توانیم بگوییم که هیچ شواهد کافی برای رد فرضیه صفر وجود ندارد، و بنابراین می‌توانیم نتیجه بگیریم که بین متغیرهای مورد مطالعه هیچ ارتباط معنی‌داری وجود ندارد.

تمرین 5:

داده‌های ویژگی سن طبق جدول زیر میباشند:

```
Age_data = [3, 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35,
36, 40, 45, 46, 52, 70]
```

الف) برای محاسبه میانگین، مد و میانه به روش زیر عمل میکنیم:

$$\text{Mean} = \frac{\sum x_i}{n} = \frac{3+13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+35+35+35+35+36+40+45+46+52+70}{28}$$

$$\text{Mean} = \frac{812}{28} = \underline{\mathbf{29}}$$

Median = **25**

Mode = **25** and **35**

ب) محاسبه چارک‌ها و انحراف چارکی

چارک‌ها مقادیری از داده هستند که آن‌ها ره به چهار قسمت مساوی تقسیم میکنند. این مقادیر به سه دسته تقسیم می‌شوند:

1. **چارک اول (Q1) : مقدار داده‌ای است که 25٪ داده‌ها کمتر از آن هستند.**
2. **چارک دوم (Q2) : همان میانه است و مقدار داده‌ای است که 50٪ داده‌ها کمتر از آن هستند.**
3. **چارک سوم (Q3) : مقدار داده‌ای است که 75٪ داده‌ها کمتر از آن هستند.**

$$Q1 = \frac{N+1}{4} = \frac{28+1}{4} = 7.25$$

چون موقعیت 7.25 بین مکان‌های 7 و 8 است، Q1 میانگین داده‌های مکان 7 و 8 خواهد بود:

$$Q1 = \frac{Age[7]+Age[8]}{2} = \frac{20+20}{2} = \underline{\mathbf{20}}$$

میانه (50٪ از داده‌ها) در موقعیت زیر قرار دارد:

$$Q2 = \frac{N+1}{2} = \frac{28+1}{2} = \frac{29}{2} = 14.5$$

چون موقعیت 14.5 بین مکان‌های 14 و 15 قرار دارد، Q2 میانگین داده‌های مکان‌های 14 و 15 میباشد:

$$Q2 = \frac{Age[14]+Age[15]}{2} = \frac{25+25}{2} = \underline{\mathbf{25}}$$

چارک سوم (75٪ از داده‌ها) در موقعیت زیر قرار دارند:

$$Q3 = \frac{(N+1)\times 3}{4} = \frac{(28+1)\times 3}{4} = \frac{29\times 3}{4} = \frac{87}{4} = 21.75$$

چون موقعیت 21.75 بین مکان‌های 21 و 22 قرار دارد، Q3 میانگین داده‌های مکان‌های 22 و 21 میباشد:

$$Q3 = \frac{Age[21]+Age[23]}{2} = \frac{35+35}{2} = \underline{\mathbf{35}}$$

انحراف چارکی برابر است با اختلاف میان چارک اول و سوم:

$$IQR = Q3 - Q1 = 35 - 20 = \underline{\mathbf{15}}$$

پ) پیدا کردن داده‌های خارج از محدوده

برای تشخیص داده‌های خارج از محدوده، از انحراف چارکی استفاده میکنیم. داده‌هایی که خارج از محدوده مشخصی باشند، به عنوان داده‌های پرت در نظر گرفته میشوند. این محدوده معمولا به صورت زیر تعریف میشود:

$$Lower\ limit = Mean - 2 \times Standard\ Deviation$$

$$Upper\ limit = Mean + 2 \times Standard\ Deviation$$

Standard Deviation **که همان انحراف استاندارد میباشد، از فرمول زیر محاسبه میشود:**

$$St_{Dev} = \sqrt{Var}$$

$$Var^2 = (\frac{1}{N}) \sum_{i=1}^{N} (x_i - \mu_A)^2$$

$$\begin{aligned} Var^2 = \left(\frac{1}{28}\right)[&(3-29)^2 + (13-29)^2 + (15-29)^2 + (16-29)^2 + (16-29)^2 + (19-29)^2 \\ &+ (20-29)^2 + (20-29)^2 + (21-29)^2 + (22-29)^2 + (22-29)^2 + (25-29)^2 \\ &+ (25-29)^2 + (25-29)^2 + (25-29)^2 + (30-29)^2 + (33-29)^2 + (33-29)^2 \\ &+ (35-29)^2 + (35-29)^2 + (35-29)^2 + (35-29)^2 + (36-29)^2 + (40-29)^2 \\ &+ (45-29)^2 + (46-29)^2 + (52-29)^2 + (70-29)^2] \end{aligned}$$

$$\begin{aligned} Var^2 = (\frac{1}{28})[&(-26)^2 + (-16)^2 + (-14)^2 + (-13)^2 + (-13)^2 + (-10)^2 + (-9)^2 + (-9)^2 + (-8)^2 \\ &+ (-7)^2 + (-7)^2 + (-4)^2 + (-4)^2 + (-4)^2 + (-4)^2 + (1)^2 + (4)^2 + (4)^2 + (6)^2 + (6)^2 \\ &+ (6)^2 + (6)^2 + (7)^2 + (11)^2 + (16)^2 + (23)^2 + (17)^2 + (41)^2] \end{aligned}$$

$$\begin{aligned} Var^2 = (\frac{1}{28})[&676 + 256 + 196 + 169 + 169 + 100 + 81 + 81 + 64 + 49 + 49 + 16 + 16 + 16 + 16 + 1 \\ &+ 16 + 16 + 36 + 36 + 36 + 36 + 49 + 121 + 256 + 529 + 289 + 1681] \end{aligned}$$

$$Var^2 = \frac{5056}{28} = 180.5$$

$$St\_Dev = \sqrt{Var^2} = \sqrt{180.5} = 13.4$$

محاسبه حد پایین و بالای داده‌ها:

$$Lower\ limit = Mean - 2 \times St\_Dev = 29 - 2 \times 13.4 = 29 - 26.8 = 2.2$$

$$Upper\ limit = Mean + 2 \times St\_Dev = 29 + 2 \times 13.4 = 29 + 26.8 = 55.8$$

حد پایین و بالا نشان‌دهنده این هستند که هیچ داده‌ایی کمتر از 2.2 نخواهد بود و داده‌های که بیشتر ای 55.8 باشند، به عنوان داده‌های پرت در نظر گرفته میشوند.

در نتیجه، با توجه به اینکه حد بالای این مجموعه داده 55.8 است، داده 70 به عنوان داده پرت یا داده خارج از محدوده شناسایی میشود.

ت) اجرای دو روش دسته بندی

روش اول – نمایش هر دسته با یک شاخص مرکزی (میانگین):

طبق میانگین بدست آمده از مجموعه اعداد (29)، داده‌ها را میتوان به سه دسته محدود کرد:

- داده‌های کمتر از میانگین

- داده‌های محدوده میانگین

- داده‌های بیشتر از میانگین

حال طبق آن، میتوان مجموعه داده‌ها به شکل زیر دسته بندی کرد:

{3, 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25}

{30}

{33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}

محاسبه ارور این دسته بندی با استفاده از مد به صورت زیر است:

Mode_1 = 22

Mode_2 = 30

Mode_3 = 35

$$Error = [|3-22| + |13-22| + |15-22| + |16-22| + |16-22| + |20-22| + |20-22| + |21-22| + |22-22| + |22-22| + |25-22| + |25-22| + |25-22| + |25-22|] + [|30-30|] + [|33-35| + |33-35| + |35-35| + |35-35| + |35-35| + |35-35| + |36-35| + |40-35| + |45-35| + |46-35| + |52-35| + |70-35|]$$

$$Error = [19+9+7+6+6+2+2+1+0+0+3+3+3+3] + [0] + [2+2+0+0+0+0+1+5+10+11+17+35] = 64+0+83 = 174$$

در روش دوم میتوانیم از فرمول زیر برای دسته بندی و یافتن محدوده دسته‌ها استفاده کنیم.

ابتدا برای پیدا کردن تعداد دسته‌ها از فرمول زیر استفاده میکنیم.

$$K = 1 + 3.3 \times log(N)$$

$$K = 1 + 3.3 \times log(28) = 1 + 3.3 \times (1.44) = 1 + 4.75 = 5.75 \sim 6$$

طبق عدد بدست آمده، تعداد دسته‌های این مجموعه 6 تا میباشد. برای محاسبه بازه دسته‌ها میتوان از رابطه زیر استفاده کرد:

$$D = \frac{R}{K}$$

$$R = max(Age\_data) - min(Age\_data) = 70 - 3 = 67$$

$$D = \frac{67}{6} = 11.16 \sim 12$$

طبق این اعداد، میتوان 6 تعداد دسته با فاصله‌ی 12 ایجاد کرد:

{3}

{13, 15, 16, 16, 19}

{20, 21, 21, 22, 22, 25, 25, 25, 25}

{30, 33, 33, 35, 35, 35, 35, 36}

{40, 45, 46, 52}

{70}

محاسبه ارور این دسته بندی با استفاده از میانگین به صورت زیر میباشد:

Mean _1 = 3

Mean_2 = 15.8

Mean _3 = 22.8

Mean _4 = 34

Mean _5 = 45.75

Mean_6 = 70

$$Error = [|3-3|] + [|13-15.8| + |15-15.8| + |16-15.8| + |16-15.8| + |19-15.8|] + [|20-22.8| + |21-22.8| + |21-22.8| + |22-22.8| + |22-22.8| + |25-22.8| + |25-22.8| + |25-22.8| + |25-22.8|] + [|30-34| + |33-34| + |33-34| + |35-34| + |35-34| + |35-34| + |35-34| + |36-34|] + [|40-45.75| + |45-45.75| + |46-45.75| + |52-45.75|] + [|70-70|]$$

$$Error = [0] + [2.8 + 0.8 + 0.2 + 0.2 + 3.2] + [2.8 + 1.8 + 1.8 + 0.8 + 0.8 + 2.2 + 2.2 + 2.2 + 2.2] + [4 + 1 + 1 + 1 + 1 + 1 + 1 + 2] + [5.75 + 0.75 + 0.25 + 6.25] + [0]$$

$$Error = 0 + 7.2 + 16.8 + 12 + 13 + 0 = 49$$

طبق محاسبات انجام شده، ارور دسته بندی که کمتر باشد (دسته بندی دوم با ارور 49) دسته بندی مناسب‌تری میباشد.

تمرین 6:

داده‌های سن افراد و درصد چربی خون آن‌ها به صورت زیر میباشد:

Age	Fat	Age	Fat
23	9.5	52	34.6
23	26.5	54	42.5
27	7.8	54	28.8
27	17.8	56	33.4
39	31.4	57	30.2

41	25.9	58	34.1
47	27.4	58	32.9
47	27.2	60	41.2
50	31.2	61	35.7

الف) انجام روش‌های نرمال سازی روی هر دو مجموعه

1. **روش حرکت نقطه اعشار:**

در این روش، مقادیر مجموعه را بر توانی از 10 متناظر با حداکثر قدرمطلق مقادیر موجود، تقسیم میکنیم.

مقادیر موجود برای مجموعه سن بین اعداد 23 تا 61 میباشند و عدد دو رقمی 61 بزرگترین قدرمطلق در این مجموعه است. پس، نمونه‌ها را باید بر عدد 100 تقسیم کرد.

مقادیر موجود برای مجموعه درصد چربی نیز بین اعداد 7.8 و 42.5 میباشند و عدد دو رقمی 42.5 بزرگترین قدرمطلق در این مجموعه است. پس همانند مجموعه قبل، این مجموعه را نیز بر عدد 100 تقسیم میکنیم.

Age	Age	Fat	Fat
23/100	52/100	7.8/100	31.2/100
23/100	54/100	9.5/100	31.4/100

27/100	54/100	17.8/100	32.9/100
27/100	56/100	25.9/100	33.4/100
39/100	57/100	26.5/100	34.1/100
41/100	58/100	27.2/100	34.6/100
47/100	58/100	27.4/100	35.7/100
47/100	60/100	28.8/100	41.2/100
50/100	61/100	30.2/100	42.5/100

داده‌های جدید نرمال‌سازی شده به صورت زیر میباشند:

Age	Age	Fat	Fat
0.23	0.52	0.078	0.312
0.23	0.54	0.095	0.314
0.27	0.54	0.170	0.328
0.27	0.56	0.259	0.333
0.39	0.57	0.265	0.341

0.41	0.58	0.272	0.346
0.47	0.58	0.273	0.357
0.49	0.6	0.288	0.412
0.5	0.61	0.302	0.425

همانطور که در جدول بالا مشخص است، محدوده جدید داده‌های سن بین 0.23 تا 0.61 و محدوده جدید بین داده‌های درصد چربی بین محدوده 0.078 تا 0.425 تعریف شده‌اند.

2. روش تبدیل محدوده:

در این روش، با استفاده از رابطه زیر، برای نرمال‌سازی محدوده مقادیر به هر محدوده جدید دلخواه تبدیل میکنیم.

$$v = \frac{v - Min}{Max - Min}(NewMax - NewMin) + NewMin$$

محدوده جدید برای داده‌های سن را میتوان بین 25 تا 60 و برای داده‌های درصد چربی بین 15 تا 40 در نظر گرفت.

Min_Age = 23 Min_New_Age = 25

Max_Age = 61 Max_New_Age = 60

Min_Fat = 7.8 Min_New_Fat = 15

Max_Fat = 42.5 Max_New_Fat = 40

$$v_{Age} = \frac{v - 23}{61 - 23}(60 - 25) + 25 = \frac{v - 23}{38}(35) + 25$$

$$v_{Fat} = \frac{v - 7.8}{42.5 - 7.8}(40 - 15) + 15 = \frac{v - 7.8}{34.7}(25) + 15$$

Age	Age	Fat	Fat
$\frac{23-23}{38}(35)+25$	$\frac{52-23}{38}(35)+25$	$\frac{7.8-7.8}{34.7}(25)+15$	$\frac{31.2-7.8}{34.7}(25)+15$
$\frac{23-23}{38}(35)+25$	$\frac{54-23}{38}(35)+25$	$\frac{9.5-7.8}{34.7}(25)+15$	$\frac{31.4-7.8}{34.7}(25)+15$
$\frac{27-23}{38}(35)+25$	$\frac{54-23}{38}(35)+25$	$\frac{17.8-7.8}{34.7}(25)+15$	$\frac{32.9-7.8}{34.7}(25)+15$
$\frac{27-23}{38}(35)+25$	$\frac{56-23}{38}(35)+25$	$\frac{25.9-7.8}{34.7}(25)+15$	$\frac{33.4-7.8}{34.7}(25)+15$
$\frac{39-23}{38}(35)+25$	$\frac{57-23}{38}(35)+25$	$\frac{26.5-7.8}{34.7}(25)+15$	$\frac{34.1-7.8}{34.7}(25)+15$
$\frac{41-23}{38}(35)+25$	$\frac{58-23}{38}(35)+25$	$\frac{27.2-7.8}{34.7}(25)+15$	$\frac{34.6-7.8}{34.7}(25)+15$

$\frac{47-23}{38}(35)+25$	$\frac{58-23}{38}(35)+25$	$\frac{27.4-7.8}{34.7}(25)+15$	$\frac{35.7-7.8}{34.7}(25)+15$
$\frac{47-23}{38}(35)+25$	$\frac{60-23}{38}(35)+25$	$\frac{28.8-7.8}{34.7}(25)+15$	$\frac{41.2-7.8}{34.7}(25)+15$
$\frac{50-23}{38}(35)+25$	$\frac{61-23}{38}(35)+25$	$\frac{30.2-7.8}{34.7}(25)+15$	$\frac{42.5-7.8}{34.7}(25)+15$

جواب‌های نهایی مجموعه داده‌های سن و درصد چربی به صورت زیر میباشند:

Age	Age	Fat	Fat
25	51.7	15	31.8
25	53.5	16.2	32
28.6	535	22.2	33
28.6	55.3	28	33.4
39.7	56.3	28.4	33.9
41.5	57.2	28.9	32.3
47.1	57.2	29.1	35.1
48.9	59	30.1	39
49.8	60	31.1	40

3. روش میانگین و انحراف استاندارد:

برای نرمال سازی داده‌ها با استفاده از میانگین و انحراف معیار، از فرمول زیر میتوان استفاده کرد:

$$V\_new = \frac{V\_old - Mean}{St\_Dev}$$

Age:

$$\text{Mean} = \frac{\sum Age\_i}{N} = \frac{23+23+27+27+39+41+47+47+50+52+54+54+56+57+58+58+60+61}{18} = 836/18 = \mathbf{44}$$

$$St_{Dev} = \sqrt{Var}$$

$$Var^2 = (\frac{1}{N}) \sum_{i=1}^{N} (x_i - \mu_A)^2$$

$$Var^2 = \left(\frac{1}{18}\right)[(23-44)^2 + (23-44)^2 + (27-44)^2 + (27-44)^2 + (39-44)^2 + (41-44)^2 + (47-44)^2 + (47-44)^2 + (50-44)^2 + (52-44)^2 + (54-44)^2 + (54-44)^2 + (56-44)^2 + (57-44)^2 + (58-44)^2 + (58-44)^2 + (60-44)^2 + (61-44)^2]$$

$$Var^2 = (\frac{1}{18})[(-21)^2 + (-21)^2 + (-17)^2 + (-17)^2 + (-5)^2 + (-3)^2 + (3)^2 + (3)^2 + (6)^2 + (8)^2 + (10)^2 + (10)^2 + (12)^2 + (13)^2 + (14)^2 + (14)^2 + (16)^2 + (17)^2]$$

$$Var^2 = \frac{441 + 441 + 289 + 289 + 25 + 9 + 9 + 9 + 36 + 64 + 100 + 100 + 144 + 169 + 196 + 196 + 256 + 289}{18}$$

$$Var^2 = \frac{3062}{18} = 170.1$$

$$St\_Dev = \sqrt{170.1} = \mathbf{2.23}$$

$$V\_new = \frac{V\_old - 44}{2.23}$$

Fat:

Mean = $\frac{\sum Fat_i}{N} = \frac{7.8+9.5+17.8+25.9+26.5+27.2+27.4+28.8+30.2+31.2+31.4+32.9+33.4+34.1+34.6+35.7+41.2+42.5}{18} = 518.1/18 = \mathbf{28.7}$

$$St_{Dev} = \sqrt{Var}$$

$$Var^2 = (\frac{1}{N}) \sum_{i=1}^{N} (x_i - \mu_A)^2$$

$$Var^2 = \left(\frac{1}{18}\right)[(7.8 - 28.7)^2 + (9.5 - 28.7)^2 + (17.8 - 28.7)^2 + (25.9 - 28.7)^2 + (26.5 - 28.7)^2 + (27.2 - 28.7)^2 + (27.4 - 28.7)^2 + (28.8 - 28.7)^2 + (30.2 - 28.7)^2 + (31.2 - 28.7)^2 + (31.4 - 28.7)^2 + (32.9 - 28.7)^2 + (33.4 - 28.7)^2 + (34.1 - 28.7)^2 + (34.6 - 28.7)^2 + (35.7 - 28.7)^2 + (41.2 - 28.7)^2 + (42.5 - 28.7)^2]$$

$$Var^2 = (\frac{1}{18})[(-20.9)^2 + (-19.2)^2 + (-10.9)^2 + (-2.8)^2 + (-2.2)^2 + (-1.5)^2 + (-1.3)^2 + (0.1)^2 + (1.5)^2 + (2.5)^2 + (4.2)^2 + (4.7)^2 + (5.4)^2 + (5.9)^2 + (7)^2 + (12.5)^2 + (13.8)^2]$$

$$Var^2 = \frac{436.91 + 386.64 + 118.81 + 7.84 + 4.84 + 2.25 + 1.69 + 0.01 + 2.25 + 2.25 + 17.64 + 22.09 + 29.16 + 34.81 + 49 + 156.25 + 190.44}{18}$$

$$Var^2 = \frac{1455.945}{18} = 85.64$$

$$St\_Dev = \sqrt{85.64} = \mathbf{9.25}$$

$$V\_new = \frac{V\_old - 28.7}{9.25}$$

Age	Age	Fat	Fat
$\frac{23 - 44}{2.23}$	$\frac{52 - 44}{2.23}$	$\frac{7.8 - 28.7}{9.25}$	31.2
$\frac{23 - 44}{2.23}$	$\frac{54 - 44}{2.23}$	$\frac{9.5 - 28.7}{9.25}$	$\frac{31.4 - 28.7}{9.25}$
$\frac{27 - 44}{2.23}$	$\frac{54 - 44}{2.23}$	$\frac{17.8 - 28.7}{9.25}$	$\frac{32.9 - 28.7}{9.25}$
$\frac{27 - 44}{2.23}$	$\frac{56 - 44}{2.23}$	$\frac{25.9 - 28.7}{9.25}$	$\frac{33.4 - 28.7}{9.25}$

$\frac{39 - 44}{2.23}$	$\frac{57 - 44}{2.23}$	$\frac{26.5 - 28.7}{9.25}$	$\frac{34.1 - 28.7}{9.25}$
$\frac{41 - 44}{2.23}$	$\frac{58 - 44}{2.23}$	$\frac{27.2 - 28.7}{9.25}$	$\frac{34.6 - 28.7}{9.25}$
$\frac{47 - 44}{2.23}$	$\frac{58 - 44}{2.23}$	$\frac{27.4 - 28.7}{9.25}$	$\frac{35.7 - 28.7}{9.25}$
$\frac{47 - 44}{2.23}$	$\frac{60 - 44}{2.23}$	$\frac{28.8 - 28.7}{9.25}$	$\frac{41.2 - 28.7}{9.25}$
$\frac{50 - 44}{2.23}$	$\frac{61 - 44}{2.23}$	$\frac{30.2 - 28.7}{9.25}$	$\frac{42.5 - 28.7}{9.25}$

جواب‌های نهایی برای نرمال سازی مجموعه دسته‌ها با روش میانگین و انحراف استاندارد به شرح زیر میباشند:

Age	Age	Fat	Fat
-1.77	0.42	-2.26	0.26
-1.77	0.57	-2.08	0.28
-1.47	0.57	-1.18	0.44
-1.47	0.72	-0.31	0.49
-0.56	0.79	-0.24	0.57
-0.41	0.87	-0.17	0.62

0.04	0.87	-0.14	0.74
0.19	1.02	0.002	1.34
0.26	1.10	0.15	1.48

ب) محاسبه ضریب همبستگی و تعیین نوع همبستگی

برای محاسبه ضریب همبستگی از فرمول زیر میتوان استفاده کرد:

$$Correlation(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x).Var(y)}}$$

$$Correlation(Age,Fat) = \frac{Cov(Age,Fat)}{\sqrt{Var(Age).Var(Fat)}} = \frac{Cov(Age,Fat)}{\sqrt{170.1 \times 85.64}}$$

از آنجایی که واریانس داده‌های سن و درصد چربی محاسبه شده‌اند، حال تنها نیاز به محاسبه کوواریانس بین این دو مجموعه داده است:

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$\mu_x = Mean\_Age = 44$

$\mu_y = Mean\_Fat = 28.7$

$Cov(Age, Fat)$

$$= (\frac{1}{18})[\ (23-44)(9.5-28.7) + (23-44)(26.5-28.7) + (27-44)(7.8-28.7) + (27-44)(17.8-28.7) + (39-44)(31.4-28.7) + (41-44)(25.9-28.7) + (47-44)(27.4-28.7) + (47-44)(27.2-28.7) + (50-44)(31.2-28.7) + (52-44)(34.6-28.7) + (54-44)(42.5-28.7) + (54-44)(28.8-28.7) + (56-44)(33.4-28.7) + (57-44)(30.2-28.7) + (58-44)(34.1-28.7) + (58-44)(32.9-28.7) + (60-44)(41.2-28.7) + (61-44)(35.7-28.7)\ ]$$

$Cov(Age, Fat)$

$$= (\frac{1}{18})[\ (-21 \times -19.2) + (-21 \times -2.2) + (-17 \times -20.9) + (-17 \times -10.2) + (-5 \times 2.7) + (-3 \times -2.8) + (3 \times -1.3) + (3 \times -1.5) + (6 \times 2.5) + (8 \times 5.9) + (10 \times 13.8) + (10 \times -0.1) + (12 \times 4.7) + (13 \times 1.5) + (14 \times 5.4) + (14 \times 4.2) + (16 \times 12.5) + (17 \times 7)\ ]$$

$Cov(Age, Fat)$

$$= \left(\frac{1}{18}\right)[403.2 + 46.2 + 355.3 + 185.3 + (-13.5) + 8.4 + (-3.9) + (-4.5) + 15 + 47.2 + 138 + 1 + 56.4 + 19.5 + 75.6 + 58.8 + 200 + 199]$$

Cov(Age, Fat) = 1787/18 = 99.27

$$Correlation(Age, Fat) = \frac{99.27}{\sqrt{170.1 \times 85.64}} = \frac{99.27}{120.69} = 0.82 > 0$$

طبق محاسبات انجام شده، ضریب همبستگی بین دو مجموعه عددی مثبت میباشد و این نمایانگر همبستگی کامل مثبت برای دو ویژگی سن و درصد چربی میباشد.