Data Mining (CSE542)

Homework 02

ID: __ Name: _조원석_ Date:__2023-03-27__

Task-1

Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Normalize the two attributes based on *z-score normalization.*

(b) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.

age	23	23	27	27	39	41	47	49	50
z-age	-1.83	-1.83	-1.51	-1.51	-0.58	-0.42	0.04	0.20	0.28
%fat	9.5	2.65	7.8	17.8	31.4	25.9	27.4	27.2	31.2
z-%fat	-2.09	-0.19	-2.28	-1.16	0.35	-0.26	-0.09	-0.11	0.33
age	52	54	54	56	57	58	58	60	61
z-age	0.43	0.59	0.59	0.74	0.82	0.90	0.90	1.06	1.13
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7
z-%fat	0.71	1.59	0.06	0.58	0.22	-0.46	0.52	1.45	0.83

Ans:

Write your Answer here

(a) normalize the two attributes based on z-score normalization

```
import matplotlib.pyplot as plt
import math
import numpy as np
from scipy import stats

X = [23,23,27,27,39,41,47,49,50,52,54,54,56,57,58,58,60,61]
Y = [9.5, 26.5, 7.8, 17.8, 31.4, 25.9, 27.4, 27.2, 31.2, 34.6, 42.5, 28.8, 33.4, 30.2, 24.1, 32.9, 41.2, 35.7]

x_mean, x_SD = np.mean(X), math.sqrt(np.var(X))
y_mean, y_SD = np.mean(Y), math.sqrt(np.var(Y))

Z_X, Z_Y = [], []

for x in X:
    Z_X.append((x - x_mean) / x_SD)

for y in Y:
    Z_Y.append((y - y_mean) / y_SD)

print(f"age z score : {Z_X}")
print(f"%fat z score : {Z_Y}")
```

```
age z score : [-1.8250109782399915, -1.8250109782399915, -1.51363469759241, -1.51363469759241, -0.5795058556496655, -0.4238177153258747, 0.043246705645497555, 0.19893484596928832, 0.27677891613118366, 0.43246705645497446, 0.5881551967787652, 0.5881551967787652, 0.7438433371025559, 0.8216874072644513, 0.8995314774263468, 0.8995314774263468, 1.0552196177501374, 1.1330636879120328]
%fat z score : [-2.090886533795722, -0.19289994423330453, -2.280685192751964, -1.1642224930093652, 0.3541667786405686, -0.2598877062178606, -0.09241830125647083, -0.11474755525132271, 0.3318375246457167, 0.7114348425582004, 1.5934403753548532, 0.06388647670749321, 0.5774593185890883, 0.22019125467145684, -0.46085099217152803, 0.5216361836019584, 1.4483002243883156, 0.8342457395298865]
```

```
x_z = stats.zscore(X)
y_z = stats.zscore(Y)

print(f"age z  : {x_z}")
print(f"%fat z : {y_z}")
```

```
age z  : [-1.82501098 -1.82501098 -1.5136347  -1.5136347  -0.57950586 -0.42381772
  0.04324671  0.19893485  0.27677892  0.43246706  0.5881552   0.5881552
  0.74384334  0.82168741  0.89953148  0.89953148  1.05521962  1.13306369]
%fat z : [-2.09088653 -0.19289994 -2.28068519 -1.16422249  0.35416678 -0.25988771
 -0.0924183  -0.11474756  0.33183752  0.71143484  1.59344038  0.06388648
  0.57745932  0.22019125 -0.46085099  0.52163618  1.44830022  0.83424574]
```

age z score : [-1.8250109782399915, -1.8250109782399915, -1.51363469759241, -1.51363469759241, -0.5795058556496655, -0.4238177153258747, 0.043246705645497555, 0.19893484596928832, 0.27677891613118366, 0.43246705645497446, 0.5881551967787652, 0.5881551967787652, 0.7438433371025559, 0.8216874072644513, 0.8995314774263468, 0.8995314774263468, 1.0552196177501374, 1.1330636879120328]

%fat z score : [-2.090886533795722, -0.19289994423330453, -2.280685192751964, -1.1642224930093652, 0.3541667786405686, -0.2598877062178606, -0.09241830125647083, -0.11474755525132271, 0.3318375246457167, 0.7114348425582004, 1.5934403753548532, 0.06388647670749321, 0.5774593185890883, 0.22019125467145684, -0.46085099217152803, 0.5216361836019584, 1.4483002243883156, 0.8342457395298865]

(b) Calculate the correlation coefficient (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.

Z_X : age, Z_Y : %fat

```
def pearson(a,b):
    return np.dot((a - np.mean(a)),(b - np.mean(b))) / ((np.linalg.norm(a - np.mean(a))) * (np.linalg.norm(b - np.mean(b))))
print(pearson(Z_X, Z_Y))
```

```
0.7651835434809369
```

```
print(np.corrcoef(Z_X, Z_Y))
```

```
[[1.         0.76518354]
 [0.76518354 1.        ]]
```

```
from scipy.stats import pearsonr as scipy_pearson
print(scipy_pearson(Z_X, Z_Y)[0])
```

```
0.765183543480937
```

Pearson’s product ≒ 0.765183543480937

So these two attributes have positively correlated

Task-2: Values for Data item having three features Height (X), test Score (Y), and Age (Z). **Compute the covariance matrix**

X	Y	Z
Height	Score	Age
64.0	580.0	29.0
66.0	570.0	33.0
68.0	590.0	37.0
69.0	660.0	46.0
73.0	600.0	55.0

	X	Y	Z
X	cov(X,X)	cov(X,Y)	cov(X,Z)
Y	cov(Y,X)	cov(Y,Y)	cov(Y,Z)
Z	cov(Z,X)	cov(Z,Y)	cov(Z,Z)

```
"""x    y    z
Height Score  Age
64.0    580.0   29.0
66.0    570.0   33.0
68.0    590.0   37.0
69.0    660.0   46.0
73.0    600.0   55.0"""

Height = [64.0, 66.0, 68.0, 69.0, 73.0]
Score = [580.0, 570.0, 590.0, 660.0, 600.0]
Age = [29.0, 33.0, 37.0, 46.0, 55.0]

matrix = np.array([Height,Score, Age])
print(np.cov(matrix))
```

```
[[  11.5    50.     34.75]
 [  50.   1250.    205.  ]
 [  34.75  205.    110.  ]]
```

```
cov_mat = np.stack((Height, Score, Age), axis = 0)

print(np.cov(cov_mat))
```

```
[[  11.5    50.     34.75]
 [  50.   1250.    205.  ]
 [  34.75  205.    110.  ]]
```

[[11.5 50.0 34.75]

[50. 0 1250. 0 205.0]

[34.75 205.0 110. 0]]

Task-3: Consider the "mixed" data given in Table 3.11. Here X1 is a numeric attribute and X2 is a categorical one. Assume that the domain of X2 is given as dom(X2) = {a, b}.

Table 3.11.

X_1	X_2
0.3	a
−0.3	b
0.44	a
−0.60	a
0.40	a
1.20	b
−0.12	a
−1.60	b
1.60	b
−1.32	a

Answer the following questions.

(a) What is the mean vector for this dataset?
(b) What is the covariance matrix?

In Table 3.11, assuming that X_1 is discretized into three bins, as follows:

$$c_1 = (-2, -0.5]$$
$$c_2 = (-0.5, 0.5]$$
$$c_3 = (0.5, 2]$$

Answer the following questions:

(a) Construct the contingency table between the discretized X_1 and X_2 attributes. Include the marginal counts.
(b) Compute the χ^2 statistic between them.

x1 = [0.3, -0.3, 0.44, -0.60, 0.40, 1.20, -0.12, -1.60, 1.60, -1.32]

x2 = [a,b,a,a,a,b,a,b,b,a]

```
x1 = [0.3, -0.3, 0.44, -0.60, 0.40, 1.20, -0.12, -1.60, 1.60, -1.32]
x2 = ['a', 'b',  'a',  'a',   'a',  'b',  'a',   'b',   'b', 'a']
x3 = [0,1,0,0,0,1,0,1,1,0]
```

```
print(np.mean(x1))
print(f"{1-np.mean(x3)}a+{np.mean(x3)}b")
```

```
0.0
0.6a+0.4b
```

```
print(np.cov(x1, x3))
```

```
[[1.02337778 0.1       ]
 [0.1        0.26666667]]
```

(a) What is the mean vector for this dataset?

x1 = 0.0

x2 = 0.6a + 0.4b

(b) what is the covariance matrix?

						2.4		2.8
	0.3	0.09		a	0.12	0.16	-0.12	0.36
	-0.3	0.09		b	0.18	0.36	-0.18	0.16
	0.44	0.1936		a	0.176	0.16	-0.176	0.36
	-0.6	0.36		a	-0.24	0.16	0.24	0.36
	0.4	0.16		a	0.16	0.16	-0.16	0.36
	1.2	1.44		b	-0.72	0.36	0.72	0.16
	-0.12	0.0144		a	-0.048	0.16	0.048	0.36
	-1.6	2.56		b	0.96	0.36	-0.96	0.16
	1.6	2.56		b	-0.96	0.36	0.96	0.16
	-1.32	1.7424		a	-0.528	0.16	0.528	0.36
Mean x1:	0	0.92104	Mean x2	0.6	a	0.4	b	
cov_x1_x2	(-0.1a)	0.1b			-0.9	a	0.9	b
var_x1	0.102338							
var_x2	0.266667	a^2	0.311111	b^2				

cov_x1_x2 = sum((x1 - mean_x1) * (x2 - mean_x2)) / (len(x1) – 1)

=> -0.1a+0.1b

var_x1 = sum((x1 - mean_x1) ^{2}) / (len(x1) - 1)

=>0.102338

var_x2 = sum((x2 - mean_x2) ^{2}) / (len(x2) - 1)

=> 0.266667a^2 + 0.311111b^2

[[0.102338 -0.1a+0.1b]

[-0.1a+0.1b 0.266667a^2 + 0.311111b^2]]

(a) Construct the contingency table between the discretized X1 and X2 attributes. Include the marginal counts. (Green table is answer)

0.3	a
-0.3	b
0.44	a
-0.6	a
0.4	a
1.2	b
-0.12	a
-1.6	b
1.6	b
-1.32	a

(-2,0.5]	(-0.5, 0.5]	(0.5, 2]
2	4	0
1	1	2

	x2=a	x2=b	Total
x1=c1	2	1	3
x1=c2	4	1	5
x1=c3	0	2	2
Total	6	4	10

(b) Compute the X^2 statistic between them. $X^2 \fallingdotseq 5.233333$

Expected count = (row total * column total) / grand total

X^2 = sum((observed count - expected count)^2 / expected count)

0.3	a
-0.3	b
0.44	a
-0.6	a
0.4	a
1.2	b
-0.12	a
-1.6	b
1.6	b
-1.32	a

	(-2,0.5]	(-0.5, 0.5]	(0.5, 2]
a	2	4	0
b	1	1	2
	C1	C2	C3

	x2=a	x2=b	Total
x1=c1	2	1	3
x1=c2	4	1	5
x1=c3	0	2	2
Total	6	4	10

0.066667	0.816667	
1.066667	0.816667	
2.4	0.066667	
		5.233333

Expected Count	2.4
X^2	5.233333

Comment

It’s too hard to me.

I will study hard later..

Thank you for giving me a good assignment