Name(s): _____

Homework 4: CSCI 347: Data Mining

Show your work. Include any code snippets you used to generate an answer, using comments in the code to clearly indicate which problem corresponds to which code.

1. [2 points] Consider the following matrix *A* and vector *v*. Compute the matrix-vector product *Av*.

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, v = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$Av = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2(-1) + 1(1) \\ 1(-1) + 3(1) \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

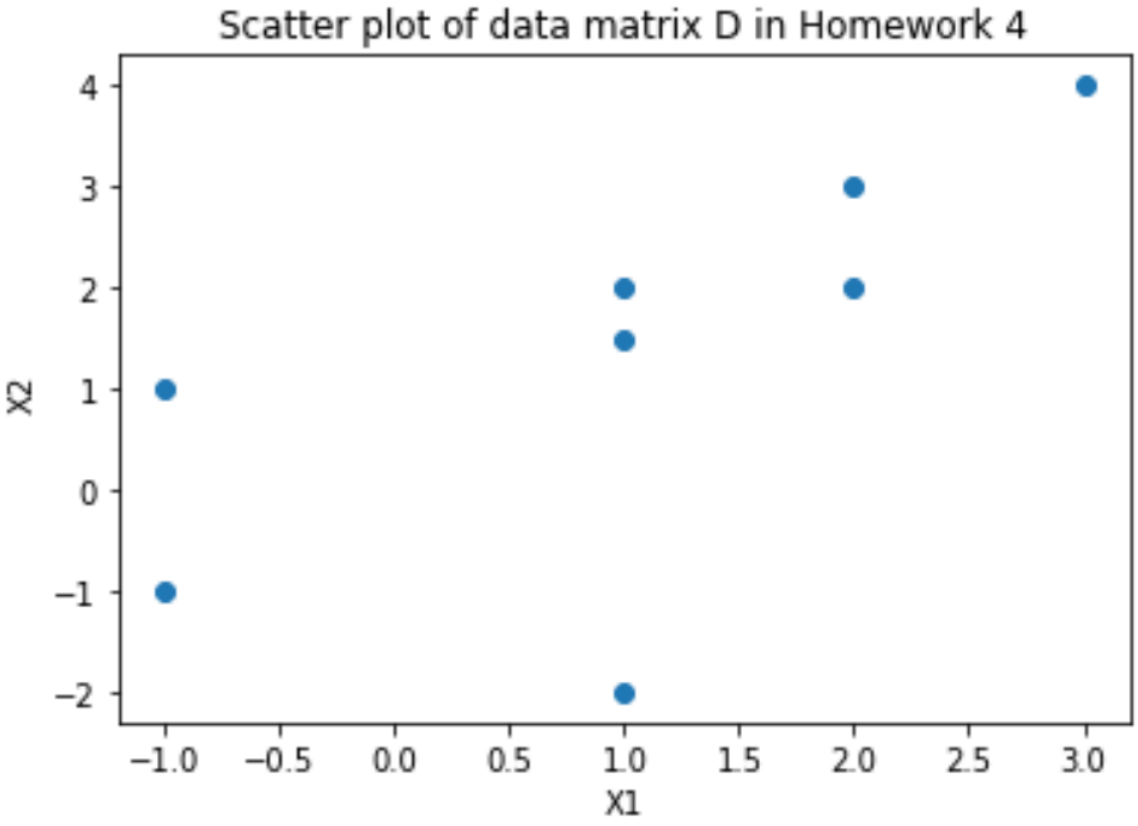
2. Consider the matrix *A* and the data set *D* below:

$$A = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 1.5 \\ 1 & 2 \\ 3 & 4 \\ -1 & -1 \\ -1 & 1 \\ 1 & -2 \\ 2 & 2 \\ 2 & 3 \end{pmatrix}$$

a) [2 points] Use Python to create a scatter plot of the data, where the x-axis is X_1 and the y-axis is X_2, and X_1 and X_2 are the first and second attributes of the data.

The code below can be used to generate the scattered plot

```
import numpy as np
import matplotlib.pyplot as plt
D = np.array([[1, 1, 3, -1, -1, 1, 2, 2], [1.5, 2, 4, -1, 1, -2, 2, 3]])
D = np.transpose(D)
plt.scatter(D[:,0], D[:,1])
plt.xlabel('X1')
plt.ylabel('X2')
plt.title('Scatter plot of data matrix D in Homework 4')
```



b) [4 points] Treating each row as a 2-dimensional vector, apply the linear transformation A to each row. In other words, find the matrix-vector product Ax_i for each x_i, where x_i is one row i of D, represented as a vector with two rows and one column. So, for example, $x_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

Using python, the following code should produce the output:

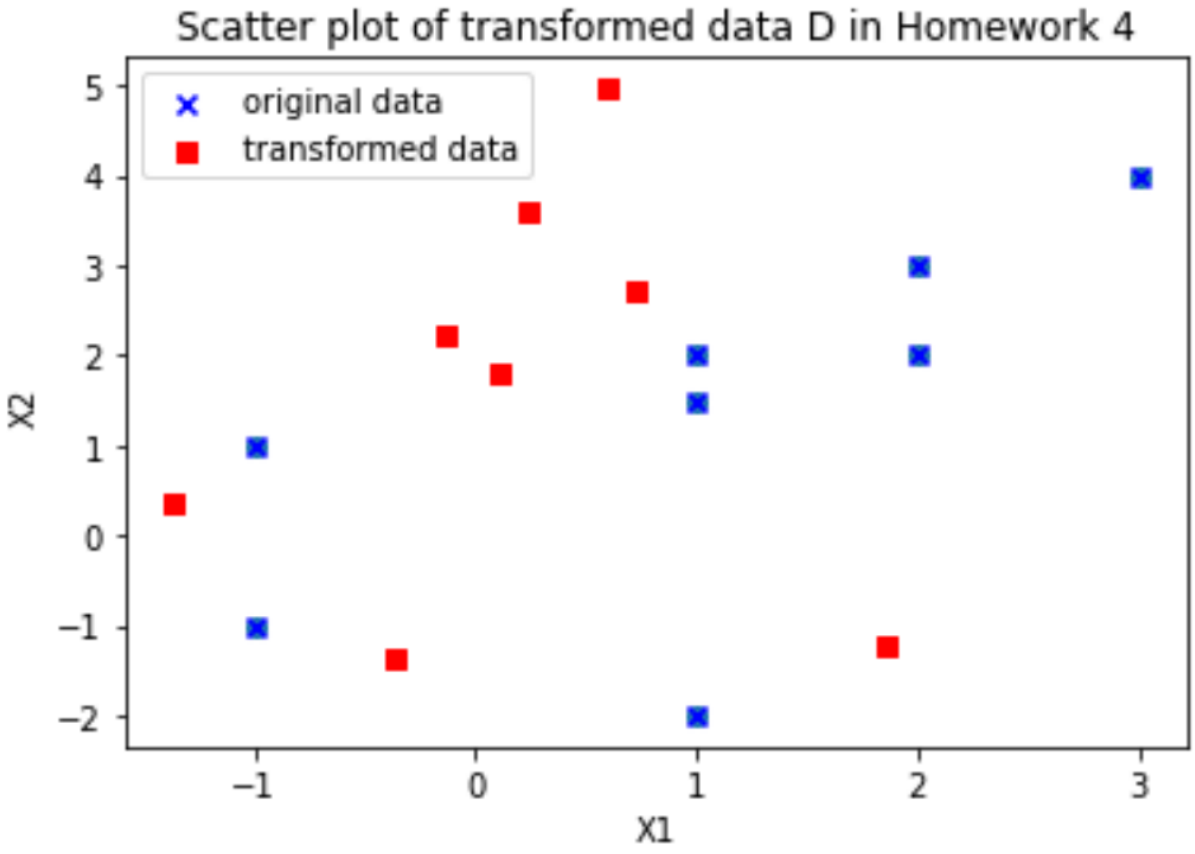
```
A = np.array([[np.sqrt(3)/2, -1/2], [1/2, np.sqrt(3)/2]])
np.transpose(A.dot(np.transpose(D)))
```

Output

```
[[ 0.1160254   1.79903811]
 [-0.1339746   2.23205081]
 [ 0.59807621  4.96410162]
 [-0.3660254  -1.3660254 ]
 [-1.3660254   0.3660254 ]
 [ 1.8660254  -1.23205081]
 [ 0.73205081  2.73205081]
 [ 0.23205081  3.59807621]]
```

c) [3 points] Use Python to create a plot showing both the original data and the transformed data, with the x-axis still corresponding to X_1 and the y-axis corresponding to X_2. Use different colors and markers to differentiate between the original and transformed data. That is, each transformed data point in the plot should be one matrix-vector product Ax_i, which is a 2-dimensional vector. Each original point in the plot should have the same coordinates as it did in part 2.1.

```
AD = np.transpose(A.dot(np.transpose(D)))
plt.scatter(D[:,0], D[:,1], c ='b',marker='x',label = 'original data' )
plt.scatter(AD[:,0], AD[:,1], c ='r', marker = "s", label='transformed data')
plt.legend(loc = 'upper left')
plt.xlabel('X1')
plt.ylabel('X2')
plt.title('Scatter plot of transformed data D in Homework 4')
```



d) [1 point] Write down the multi-variate mean of the data. (Remember that this should be a 2-dimensional vector)

We can obtain this by using the python code below:

```
meanD = np.array([np.mean(D[:,0]), np.mean(D[:,1])])
```

$\mu = (1, 1.3125)$

e) [2 points] Mean-center the data. Write down the mean-centered data matrix.

We can obtain the results by using the following:

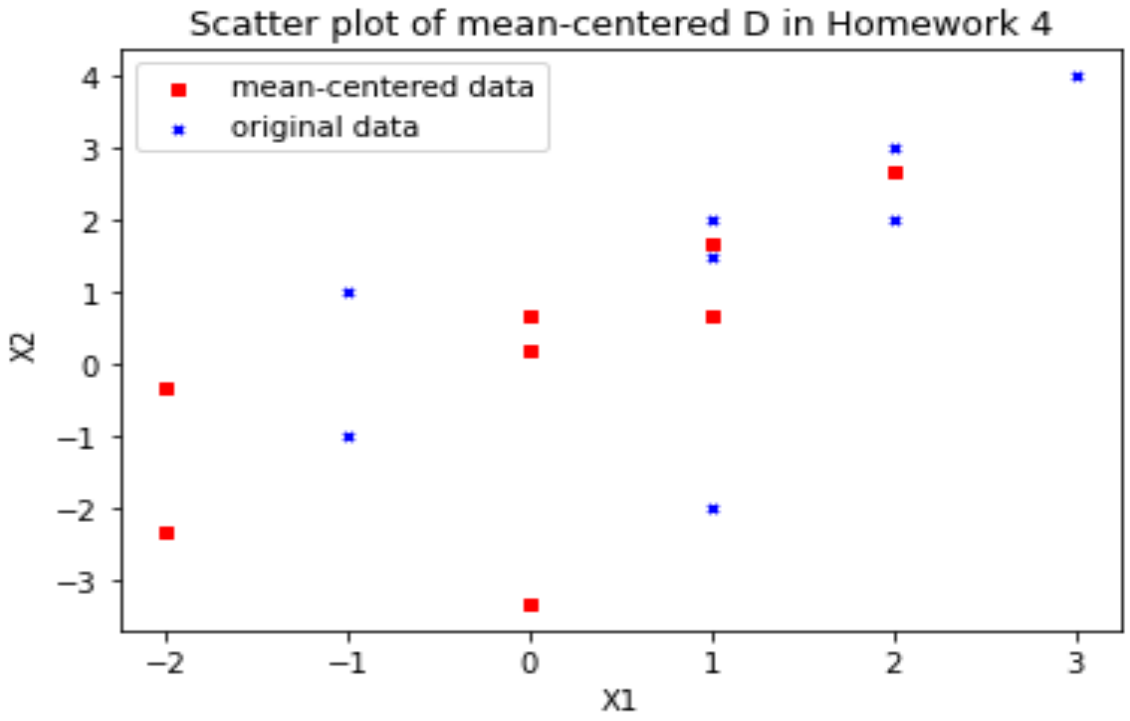
```
meanDs = np.ones(shape=(8,2))*meanD # meanD is defined in Part d
Z = D - meanDs

print(Z)
```

Output:

```
[[ 0.      0.1875]
 [ 0.      0.6875]
 [ 2.      2.6875]
 [-2.     -2.3125]
 [-2.     -0.3125]
 [ 0.     -3.3125]
 [ 1.      0.6875]
 [ 1.      1.6875]]
```

f) [2 points] Use Python to create a scatter plot showing both the original data and the mean-centered data, where the x-axis is X_1 and the y-axis is X_2, and X_1 and X_2 are the first and second attributes of the data. Use different colors and markers to differentiate between the original and mean-centered data.

```
fig = plt.figure()
ax = fig.add_subplot(111)
ax.scatter(Z[:,0], Z[:,1], s=10, c='r', marker="s", label='mean-centered data')
ax.scatter(D[:,0],D[:,1], s=10, c='b', marker="x", label='original data')
plt.legend(loc='upper left')
plt.xlabel('X1')
plt.ylabel('X2')
plt.title('Scatter plot of mean-centered D in Homework 4')
```



g) [3 points] Write down the covariance matrix of the data matrix D. Use sample covariance.

The following python code can be used to obtain the results:

```
Convariance_matrix = np.cov(np.transpose(D))

print(Convariance_matrix)
```

```
[[2.          1.85714286]
 [1.85714286 3.92410714]]
```

h) [3 points] Write down the covariance matrix of the centered data matrix Z. Use sample covariance.

```
COV_Z = np.cov(np.transpose(Z))

print(COV_Z)
```

```
[[2.          1.85714286]
 [1.85714286 3.92410714]]
```

i) [3 points] Write down the covariance matrix of the data after applying standard normalization.

The normalized matrix is given by:

$$S = \begin{pmatrix} \frac{1-1}{\sqrt{2}} & \frac{1.5-1.3125}{\sqrt{3.924}} \\ \frac{1-1}{\sqrt{2}} & \frac{2-1.3125}{\sqrt{3.924}} \\ \frac{3-1}{\sqrt{2}} & \frac{4-1.3125}{\sqrt{3.924}} \\ \frac{-1-1}{\sqrt{2}} & \frac{-1-1.3125}{\sqrt{3.924}} \\ \frac{-1-1}{\sqrt{2}} & \frac{1-1.3125}{\sqrt{3.924}} \\ \frac{1-1}{\sqrt{2}} & \frac{-2-1.3125}{\sqrt{3.924}} \\ \frac{2-1}{\sqrt{2}} & \frac{2-1.3125}{\sqrt{3.924}} \\ \frac{2-1}{\sqrt{2}} & \frac{3-1.3125}{\sqrt{3.924}} \end{pmatrix} = \begin{pmatrix} 0 & 0.095 \\ 0 & 0.347 \\ 1.414 & 1.357 \\ -1.414 & -1.167 \\ -1.414 & -0.158 \\ 0 & -1.672 \\ 0.707 & 0.347 \\ 0.707 & 0.852 \end{pmatrix}$$

So the equations for the variances and covariance are:

$$var(S_1) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{(x_{i1} - \hat{\mu}(X_1))}{\hat{\sigma}(X_1)}\right)^2 = \frac{1}{(n-1)\hat{\sigma}_1^2}\sum_{i=1}^{8} z_{i1}^2 = \frac{1}{\hat{\sigma}_1{}^2}\sigma_1^2 = 1$$

$$var(S_2) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{(x_{i2} - \hat{\mu}(X_2))}{\hat{\sigma}(X_2)}\right)^2 = \frac{1}{(n-1)\hat{\sigma}_2^2}\sum_{i=1}^{8} z_{i2}^2 = \frac{1}{\hat{\sigma}_2{}^2}\sigma_2^2 = 1$$

$$cov(S_1, S_2) = \frac{1}{n-1}\sum_{i=1}^{8}\frac{(x_{i1} - \hat{\mu}(Z_1))}{\hat{\sigma}(Z_1)}\frac{(x_{i2} - \hat{\mu}(Z_2))}{\hat{\sigma}(Z_2)} = \frac{1}{(8-1)\sqrt{\hat{\sigma}_1^2\hat{\sigma}_2^2}}\sum_{i=1}^{8} z_{i1}z_{i2}$$

$$= \frac{\hat{\sigma}_{12}}{\sqrt{\hat{\sigma}_1^2\hat{\sigma}_2^2}} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2} = \rho_{12} = \frac{1.857}{\sqrt{(2)}\sqrt{(3.924)}} = 0.663$$

So, the covariance matrix is:

$$\Sigma = \begin{pmatrix} 1 & 0.663 \\ 0.663 & 1 \end{pmatrix}$$

Acknowledgement: Solutions adopted from Dr. Veronika Strnadova-Neeley