Name(s): Michael Belmear

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.

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

see turned in code

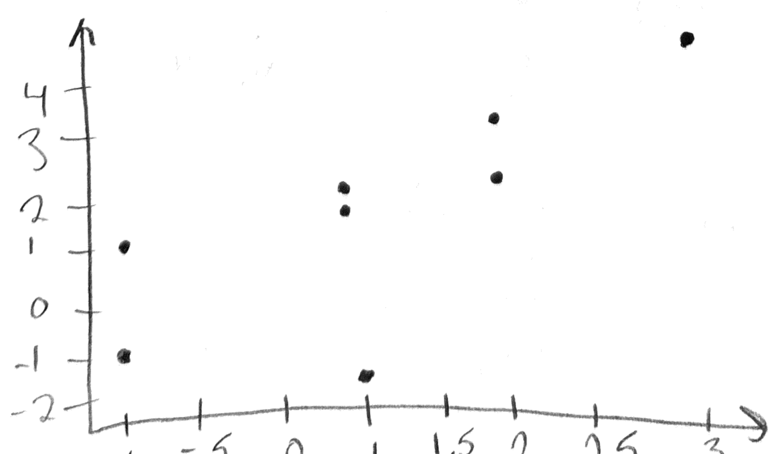
array([-1, 2])

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

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

   1. [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.

see turned in code



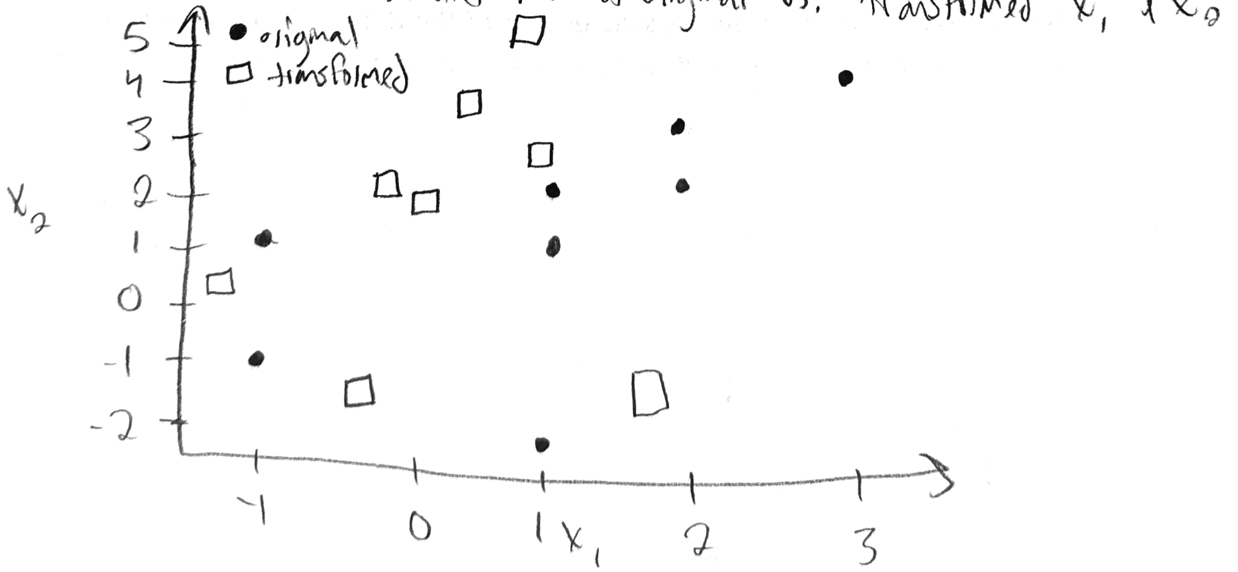
2. [4 points] Treating each row as a 2-dimensional vector, apply the linear transformation A to each row. In other words, find the matrix-vector product $A\,x_i$ for each x_i, where x_i is one row i of D, represented as a vector with two rows and one column. So, for example, $x_2 = \begin{pmatrix} 1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

See additional code

$x_1 \rightarrow [0.11, 1.79]$
$x_2 \rightarrow [-0.14, 2.23]$
$x_3 \rightarrow [0.59, 4.96]$
$x_4 \rightarrow [-0.366, -1.36]$
$x_5 \rightarrow [-1.36, 0.36]$
$x_6 \rightarrow [1.86, -1.23]$
$x_7 \rightarrow [0.73, 2.73]$
$x_8 \rightarrow [0.23, 3.59]$

3. [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 $A\,x_i$, which is a 2-dimensional vector. Each original point in the plot should have the same coordinates as it did in part 2.1.

See code, output:
Scatter plot of original vs. transformed x_1 & x_2



4. [1 point] Write down the multi-variate mean of the data. (Remember that this should be a 2-dimensional vector)

see code

multivariate mean: [1, 1.3125]

5. [2 points] Mean-center the data. Write down the mean-centered data matrix.

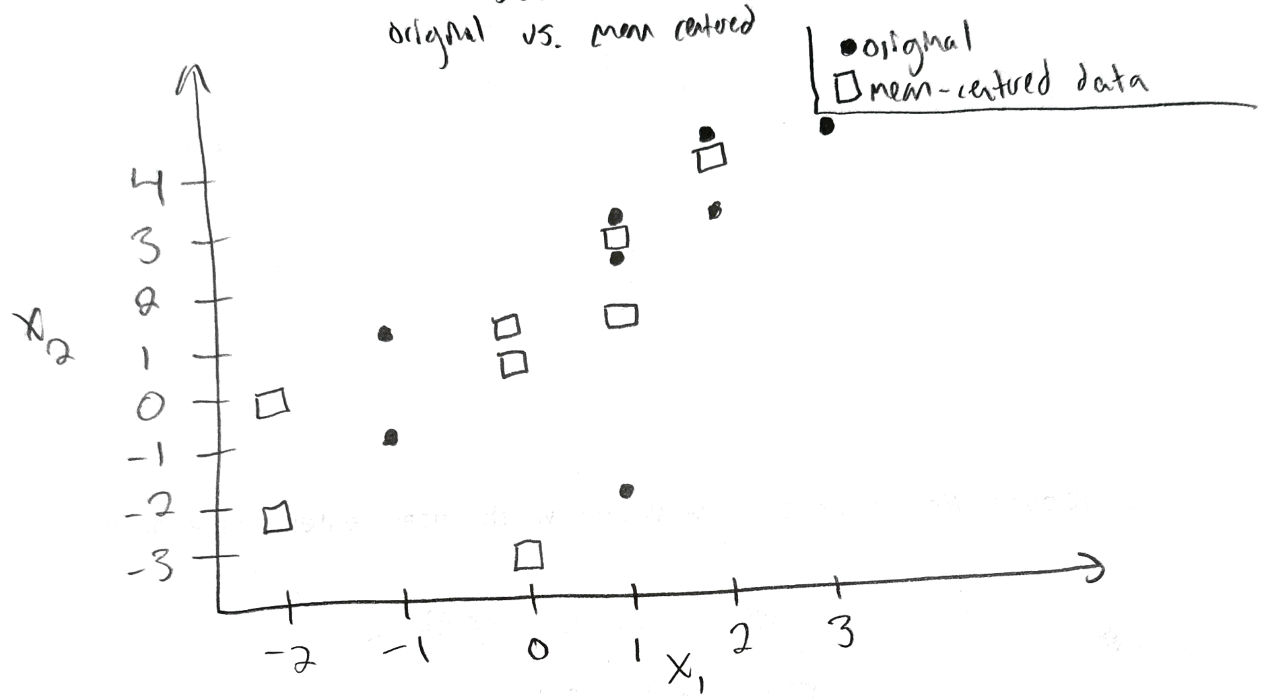
see code

mean-centered data matrix:

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

6. [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

see code



7. [3 points] Write down the covariance matrix of the data matrix D. Use sample covariance.

see code

covariance matrix:

[2, 1.85714],

[1.85714, 3.924]

8. [3 points] Write down the covariance matrix of the centered data matrix Z. Use sample covariance.

see code

Covariance matrix of matrix Z:

[2, 1.85714],

[1.85714, 3.924]

same as matrix D

9. [3 points] Write down the covariance matrix of the data after applying standard normalization.

see code:

covariance matrix after applying normalization:

[1, 0.6629],

[0.6629, 1]

3. EXTRA CREDIT

1. [2 points] Find the eigenvectors and eigenvalues of the matrix C, where C is defined as follows: $C = \dfrac{1}{n-1} Z^T Z$, where Z is the mean-centered data matrix that we used in Problem 2. What is the sum of the eigenvalues? How does it compare to the total variance in the data (smaller, larger, same? How close are the values?)?

2. [2 points] Let u_1 be the 2x1 eigenvector corresponding to the larger eigenvalue. For each row x_i in the data set D, find the dot product $u_1^T x_i$. Let p be the vector obtained by stacking these dot products into a vector:

$$p = \begin{pmatrix} u_1^T x_1 \\ u_1^T x_2 \\ u_1^T x_3 \\ u_1^T x_4 \\ u_1^T x_5 \\ u_1^T x_6 \\ u_1^T x_7 \\ u_1^T x_8 \end{pmatrix}$$

What is the variance of the vector data in vector p? What fraction of the total variance of the data is the variance in p?