Discussion 15A

1. Using PCA to Detect Fraudulent Transactions (Spring 2023 Final)

PCA has many different uses when applied to real-world data. One potential application is making classification of data much easier.

Suppose we are given some data, where each datapoint represents a transaction. Each one is labeled either normal or fraudulent. We will utilize PCA to develop a useful classifier.

We plot the data in two dimensions, where each dimension is some unspecified feature that will aid us in classifying the points:

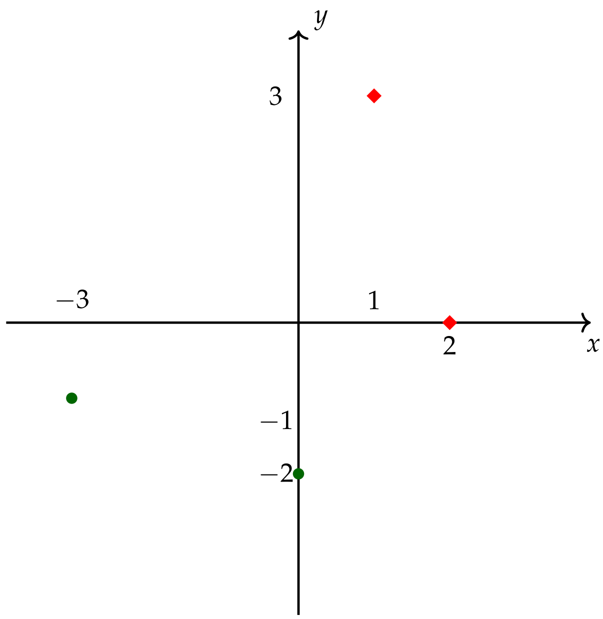

Figure 1: Plot of Transactions in 2-D

Thus, we have a total of 4 transactions, $\begin{bmatrix} -3 \\ -1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ -2 \end{bmatrix}$ are normal, while $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ are fraudulent.

(a) Suppose we now construct a data matrix, where the *data points are columns*.

$$X = \begin{bmatrix} -3 & 0 & 2 & 1 \\ -1 & -2 & 0 & 3 \end{bmatrix} \tag{1}$$

Using this data matrix, **calculate its first principal component $\vec{u}_1$.**

(HINT:

i. You may also make use of the fact that XX^T is given by:

$$XX^\top = \begin{bmatrix} -3 & 0 & 2 & 1 \\ -1 & -2 & 0 & 3 \end{bmatrix} \begin{bmatrix} -3 & 0 & 2 & 1 \\ -1 & -2 & 0 & 3 \end{bmatrix}^\top \tag{2}$$

$$= \begin{bmatrix} 14 & 6 \\ 6 & 14 \end{bmatrix} \tag{3}$$

ii. *You may also make use of the characteristic polynomial of XX^T:*

$$\lambda^2 - 28\lambda + 160 = (\lambda - 20)(\lambda - 8) = 0 \tag{4}$$

)

(HINT: Remember that your principal component should be of unit norm.)

Solution:

Our eigenvalues are $\lambda_1 = 20$ and $\lambda_2 = 8$. Thus, the eigenvector corresponding to our largest eigenvalue $\lambda_1 = 20$ is $\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ from inspection.

Thus, our principal component vector is given as the normalized eigenvector: i.e. $\vec{u}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$.

(b) It's difficult to come up with a useful classifier in two dimensions. Let's use PCA dimensionality reduction to help.

Using your answer in part (a), **project your two-dimensional data points onto one dimension. Express your answer as the vector $\vec{z} \in \mathbb{R}^{1\times 4}$.**

Solution: To project our 2-d data into 1-d, we project our data matrix onto the first principal component as follows:

$$\vec{z} = \vec{u}_1^\top X = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} -3 & 0 & 2 & 1 \\ -1 & -2 & 0 & 3 \end{bmatrix} = \begin{bmatrix} -\frac{4}{\sqrt{2}} & -\frac{2}{\sqrt{2}} & \frac{2}{\sqrt{2}} & \frac{4}{\sqrt{2}} \end{bmatrix} \tag{5}$$

(c) **Now, plot each of these points, on the line below. Indicate the value and label (normal as circle and fraudulent as diamond) for each point.**

Note: your plot doesn't have to be to scale.

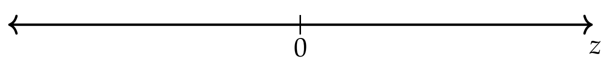

Figure 2: Plot of Transactions in 1-D

Solution:

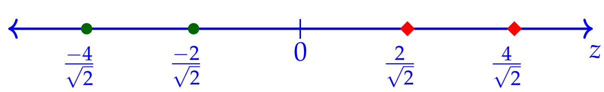

Figure 3: Plot of Transactions in 1-D

(d) Suppose you are given some transaction datapoint, $\vec{x}_i$ and you project it to be one-dimensional, i.e. z_i. **Based on your plot from part (c), come up with an inequality in terms of z_i to identify if that transaction is fraudulent.**

Note: There can be more than one answer, but please only give one.

Solution: $\vec{z}_i > 0$ is a possible answer. In fact any answer in between $\vec{z}_i \geq -\frac{2}{\sqrt{2}}$ and $\vec{z}_i \geq \frac{2}{\sqrt{2}}$ is valid.

(e) It can often be informative to compare our original data with its PCA reconstructions. **Using PCA, reconstruct $\vec{z}$ back into 2-D as the matrix $\widetilde{X} \in \mathbb{R}^{2\times 4}$.**

Solution: Using the reconstruction principle of PCA, we lift the 1D data back into 2 dimensions by doing the following:

$$\widetilde{X} = \vec{u}_1 \vec{u}_1^\top X = \begin{bmatrix} -2 & -1 & 1 & 2 \\ -2 & -1 & 1 & 2 \end{bmatrix} \tag{6}$$

(f) Finally, we visualize our PCA reconstruction. **On the graph below, draw your first principal component direction (extend it as a solid line in both directions). Draw your reconstructioned points from $\widetilde{X}$, with dotted lines connecting them with the corresponding point in X.** You may draw on the plot on the next page.

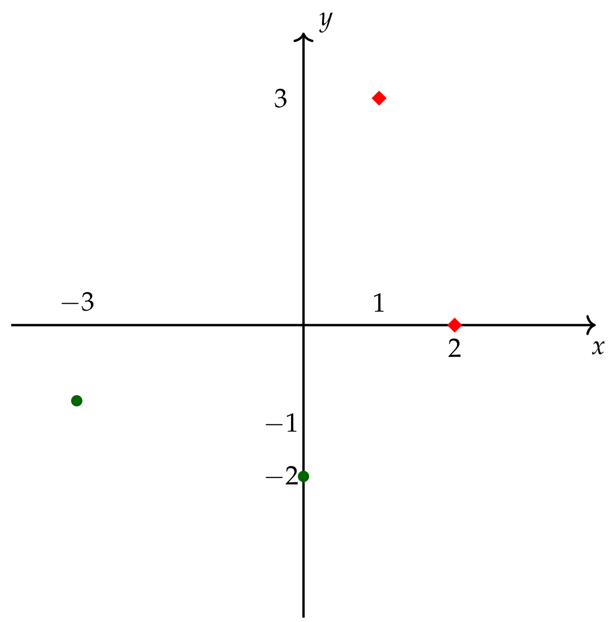

Figure 4: Plot of Transactions in 2-D

Solution:

Using the answer from part (d), students should end up with the following graph:

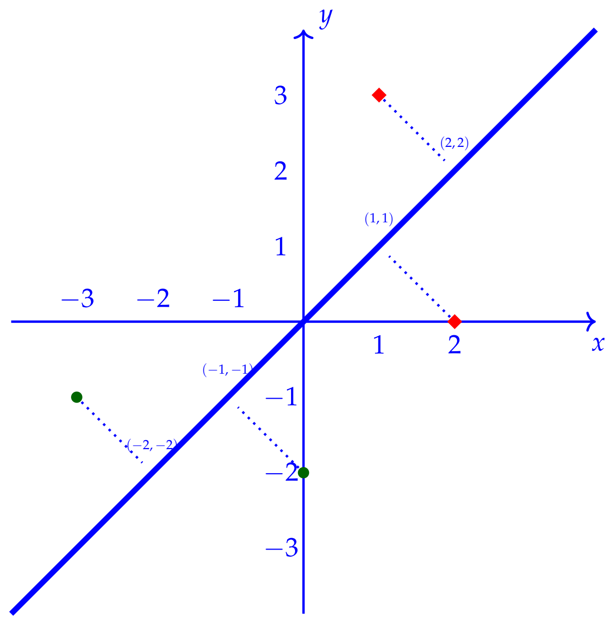

Figure 5: Plot of Transactions in 2-D