Thomas Freeman

03/01/22

Bioinformatics for CS

Report for Hw4

**Abstract:**

The goal of this assignment was to implement PCA analysis on the provided data in AA.csv. We should be able to print a copy of the initial data, normalize it, calculate eigenvalues and eigenvectors, and plot this to a scatterplot that looks like the one shown in the article excerpt. Lastly, the process needs to be reconstructed such that we end up with a plot of the original data. These results can be seen at the end of the appendix.

**Method:**

This is all done without using any packages that can do the PCA for us. I also had modified the given CSV file to give each column a name so that all numbers could be used. To start we define our class called PCA, which will be dealing with the calculations of eigen values, eigen vectors, and the covariance matrix. We also print out the shape and matrix. We take the transpose of the eigenvector matrix next, and resort all the values. These eigen vectors are also indexed to put them into the proper order for PCA to work properly. We also have two other methods, init and apply that deal with reproducing the original data. This refers to the number of eigenvalues, which should print as eight. After printing out the data set and dropping the column of letters that aren’t used in the analysis, Standard Scalar is used to normalize the data before running the PCA class. The data is then fit to the PCA showing the eigenvectors normalized, the eigenvectors transposed, and the eigenvectors indexed. The data from PCA is then projected and scaled to the scatter plot, looking similar to the plot in figure 2-10.

**Discussion:**

This was interesting to figure out as I am currently taking machine learning. I thought it was interesting to implement a solution without using very rigorous library structures. The Data in the original article must have been scaled down as without any scaling the datapoints are heavily inflated. I also discovered that depending on how the covariance matrix is calculated, the datapoints can be heavily skewed, more negatively then positively. It did take a bit to implement the transposing and indexing matrices but that’s all apart normalizing the data set to get it ready for PCA. My only trouble with this assignment was figuring out how to reproduce the original data since performing PCA forces you to lose some data, however, if we take the transpose of our eigenvectors again, they are turned back to their original form.

**Results:**

The results are shown with the two Scatterplots in the end of the assignment showing the produced PCA analysis of AA.csv and the reconstruction of the PCA analysis using the transpose of the already transposed data, leading us to the original. These are both labeled as figures 1 and 2. This is shown below along with the given code in the appendix.

**Appendix**

****

****

****

Figure 1

****

Figure 2