Dimension Reduction

Brandon Maness

**Introduction**

This assignment was geared toward learning PCA and LDA for dimensional reduction. The dataset is a classic dataset, so the main objective was to reduce the iris dataset to the most relevant components.

**Method**

Two models were used for this task, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). LDA changes the data dimensionality to maximize class separation by shifting the axes of origin around the datasets, this method utilizes the dataset fully. PCA changes the data dimensionality to maximize the variance of the data, that is, find the most influential components to the data. PCA do not fully utilize the dataset, it does not take into account the class of the testing data.

**Experiment**

**Data Description**

The iris dataset is a classic dataset use to visualize statistics and machine learning. We are using a stripped version of this dataset. The data was collected years ago by a botanist, the data was of each iris flower. Our data components are as follows:  
**sepal\_len**: The sepal length, float  
**sepal\_wid**: The sepal width, float  
**petal\_len**: The petal length, float  
**petal\_wid**: The petal width, float   
**class**: The type of iris of the flower,  there are 3 types in our set, Setosa, Versicolor, and Virginica

**Results**

PCA gave us a very high value for variance in the sepal length and decent variance in sepal width, petal length and width were not much help, this showed up early in the eigenvalues from the covariance matrix and did not change much through PCA. LDA gave me a very high eigenvalue for sepal width, so large that the others were not considerate.

**Observations/Analysis**

The larger eigenvalues in the covariance matrix scaled exponentially when passed through LDA, which in a large dataset with many classes would seem to help a lot to separate out non-needed features. We could separate this data with 99% accuracy if we relied on the sepal width.

I attempted to plot the 1d histograms but I was constantly receiving color warnings, so I decided to leave that out. I couldn’t figure out the feature plots either, those will have to wait for another day.

**Conclusion**

I learned a lot about LDA and PCA, I feel very comfortable with using them and I finally have a good understanding of what the eigenvalues and eigenvectors can do for me. I wish I would have had more experience in graphing, I hope to continue working on this until I’m able to graph the feature space properly.