Sunny Mei & Miguel Dominguez **Data Science CSC 59970**

Prof. Etemadpour **Date: 10/03/17**

Assigment #3

Comparison between PCA and LDA

This brief report has the purpose of comparing PCA and LDA. To start, we will denote some important points about each method with lists.

PCA:

- Reduce dimensions of the dataset with minimal loss of information
- Project feature space onto a smaller subspace
- Finds directions that maximize the variance in dataset
- Finds axes with maximum variance where data is most spread

LDA:

- Project a dataset into a lower dimensional space
- The lower dimensional space has good class separability
- Used to prevent overfitting due to unnecessary or useless features
- Reduce computational costs
- Used for classification
- Finds axes that maximize the separation between multiple classes
- Maintains class discriminatory information

As we can see, both methods employ techniques for dimensionality reduction but they differ at their specific purpose. In the case for PCA, we find a subspace that is able to tell how spread or how greatly the variance change without losing information about the data. In the other side, LDA is in charge of reducing dimensions while keeping a subspace that can separate multiple classes. Both methods can be used for visualization purposes of the dataset or classification models. It is essential to make sure that information representative of the dataset is not lost along the process. The math behind both method is very similar.

When we applied the PCA to the dataset for cars, we noticed a huge cluster that contains 3 different classes stacked on top of each other. It seems very hard to distinguish which is which. That should be a common thing for PCA because it focuses on finding the max variance within the data and it doesn’t take the labels into account. It treats the dataset as a whole. When using LDA, the cluttering was still present, but wasn’t as bad as PCA. LDA tries to find the most variance between classes while taking labeling into account. Because of this we can somewhat see the separation better compared with PCA.

In addition, the car dataset let us explore the structures and matrices that help us better understand the importance of data features. In this case, eigen values and eigen vectors were essential in determining which features matter the most and which one separates the data into the classes we have. As well, PCA uses similar methods to reach its subspace.

In the end, using LDA gave us a better understanding of how to find the features that are most valuable when analyzing data. We were able to visualize how class can be separated based on features that distinguish the classes. No information was lost during this process, which means that both LDA and PCA offer a smaller subspace that represents the entire data. This is important for machine learning algorithms and data science in general since we are always interested in separating classes or see how much data varies. Finally, a key point to remember both methods is that PCA finds the directions or axes that maximize the data in general while LDA finds the axes that maximize the separability of classes in the dataset.