Searching for Similarity: Narrative Document

Isabelle Kirby, Bridgette Bryant, Rikita Patangay

Zuhayr Ali

# How kNN and Decision Trees Work for Classification and Regression

# How the 3 Clustering Methods of Step 3 Work

# How PCA and LDA Work, and Why they Might be Useful Techniques for Machine Learning

* Principal Components Analysis is a data reduction technique that helps us reduce the dimensions of our datasets. PCA will manipulate the data and reduces the number of axes in a new coordinate space. In this reduced new coordinate space, each axis will represent a principal component. The first principal component (PC1) will represent the dimension of the most significant variance, and the other principal components represent decreasing variance. Since it is a data reduction technique, we will be losing data and may also lose accuracy in any models. Many times, in machine learning you will come upon high-dimensional datasets that can be hard to explore without reduction. This is when PCA is used in ML. LDA works by seeking to find a linear combination of the predictors that will maximize the separation of the classes while minimizing the within-class standard deviation. LDA is a supervised classification technique. Again, it is basically used to reduce the data so we are able to see the numbers clearer and make predictions with the data.