CS 6140 – Data Mining
Project Proposal

1/31/2018

Our group will consist of Nick Stephenson, Yulong Liang, and Zhimeng Pan. The data we plan to use will be a dataset related to health and disease. Currently, we are interested in the *Diabetes 130-US hospitals for years 1999-2008 Data Set* from the Machine Learning Repository of the University of California at Irvine. We want to cluster the patients into different groups to discover the similarities within each group and the differences among the groups. This topic will give us an intuition of how data mining can be implemented in biology and medical research.

Neural networks do feature extraction somewhat magically, and thus there is a lack of interpretability and reliability – especially in fields like human health. Features extracted by data mining techniques like clustering and dimensionality reduction may be more interpretable and meaningful. We will use different dimensionality reduction techniques (PCA, t-SNE, and PLMP) and clustering algorithms (hierarchical clustering, k-means clustering, and DBSCAN) to deal with the data and compare the advantages, limitations, and scopes among those techniques.