**CS 4641 - Machine Learning**

**Assignment 3 – Unsupervised Learning**

**Datasets**

For this assignment, I decided to use one of the datasets that I have been using before which is the letters dataset. Briefly, the letters dataset is formed of 20000 data points that each point represents a written letter. 16 attributes describe these letters such as X and Y dimension, the degree of the curves available and many more that describe the physical shape of the written letter. The question is basically doing the OCR (Optical Character Recognition) having all the features and guessing the letter based on the given attributes.

For the second dataset, I decided to use a new one since the old one I was using covered the entire feature space (all possible feature combinations were given) and this lead to very similar cluster that doesn’t allow us to infer anything. Instead, I used the “Wine Quality” dataset, which has 4898 data points that each of them belongs to a certain wine and covers 11 physical features such as acidity, density and the label, which is a score between 1 and 10. The score is basically a subjective score given by human testers and they have not been exposed to the attributes we’ve given; they only tasted the wines to label it with a score. The question here is to understand if any of those chemical or physical features of the wine actually contribute to the quality of the wine.

**Why?**

The reason I am using the letter dataset is the number of labels it has. Since the dataset is trying to capture all the letters in the alphabet, there are total of 26 different labels, and compared to any binary classification problems or low number of labeled datasets, this is much harder to cluster into since a lot of labels will be somehow interesting with each other in a given dimension. Also we know that each attribute physically builds the optical character image, so each attribute is important and should be distinctive in order to distinguish the label. Also, the number of labels is almost-equally distributed throughout the dataset, meaning that there are around 700-800 instances belonging to each of the letters. This will be an important point when it comes to clustering since the algorithms will try to create the clusters in similar sizes.

The wine dataset contrasts with the letter dataset in couple different senses. First of all it has less number of labels (10 labels) and the labels are given in a more arbitrary sense without the features are regarded. So in this paper, I will also be trying to determine if a feature actually is important or not, whereas in the letter dataset we know that each feature is important. In this aspect, dimensionality reduction algorithms will show if a feature is important and always give valuable information or if it can be reduced without really causing any decreases in the final accuracy value. Another difference of wine dataset is the distribution of the labels. There are less number of labels of really poor (ranked 1-3) and excellent wines (8-10), meaning that if we want to cluster each ranking, some clusters will be expected to be larger however the algorithms might not be able to capture those differences.

**Clustering Algorithms**