1. Discuss dataset and why interesting
2. Method explanations (how did I choose k)
3. Run clustering on both datasets, describe (purity of clusters, compare alg, can I improve performance? How much performance is due to the problem set?)
   1. Wine, k means
   2. Pet, k means
   3. Wine, EM
   4. Pet, EM
4. Run dim reduction on both datasets, describe (how does data look after reduction… more to come)
   1. Wine, PCA
   2. Pet, PCA
   3. Wine, ICA
   4. Pet, ICA
   5. Wine, Rand Proj
   6. Pet, Rand Proj
   7. Wine, other feature selection alg (maybe SVM L1 norm. feature agglomeration available in sklearn)
   8. Pet, other feature selection alg
5. Run clustering on both datasets after dim reduction, describe (amount of reduction, purity of clusters, compare alg, can I improve performance? How much performance is due to the problem set?)
   1. Wine, k means
   2. Pet, k means
   3. Wine, EM
   4. Pet, EM
6. For ONE dataset, train NN on dim reduced data features (wine)
7. For the SAME dataset, dim red -> cluster -> use clusters as features to train NN

(10 page limit)

**Introduction**

Unlike supervised learning, unsupervised learning involves data that does not include labels, and so does not seek to approximate a function that maps data to their labels. It seeks to describe the data, and commonly includes methods for clustering and dimensionality reduction. Clustering seeks to group like data points together, while dimensionality reduction seeks to describe the data in a more compact form. We will use the same datasets from the previous paper on supervised learning, and explore the use of unsupervised learning (clustering and dimensionality reduction) on those datasets.

**Datasets**

*Wine Quality(red):*

This dataset contains 1599 instances of red wine reviews and its physiochemical composition [1]. The red wines sampled were all variants of the Portuguese “Vinho Verde” wine. The dataset is interesting because there are many facets of the wine to consider when reviewing it, including its grape variety, where it was grown, how long it had been aged, and so on. The features given here are not subjective, which can be more easily controlled for during the manufacturing process. There are 11 features in the dataset, including acidity, amount of chlorides, amount of sulphates, and alcohol content. All are continuous variables, and were normalized to zero mean and 1 standard deviation. The reviews were then aggregated into two classes based on the median of the labels.

*Pet Adoption:*

This dataset contains information and outcomes about 25,000 dogs and cats at the Austin Animal Center [2]. This dataset is interesting because a good predictive model can help animal shelter workers focus their efforts on the animals that are less likely to be adopted or returned to their owner. While the wine dataset had features that were entirely continuous, the pet adoption dataset had mostly categorical data, which required different processing and may affect the applicability of the models.

One-hot encoding was attempted. However, that resulted in large number of features (300-1000). Due to time and computational constraints, the features that led to an explosion in the number of features after one-hot encoding were disregarded. What remained were 3 boolean variables and 1 continuous variable. The continuous feature was normalized to zero mean and 1 standard deviation. The classes were then aggregated into 2 outcomes. “Adopted” and “returned to owner” were treated as good outcomes, while “died”, “euthanized”, and “transferred” were treated as bad outcomes. The data was only slightly skewed, with about 10,000 good outcomes, and 15,000 bad outcomes.

**Methods**

The algorithms were implemented in Python, with the help of the Scikit-Learn, Matplotlib, and Pandas library. We explored 2 clustering methods: k-means and Gaussian mixture models. Models were trained over a range of k, and evaluated with a variety of metrics. Since k-means minimizes the Euclidean distance error (investigate this)

**Clustering**

**Dimension Reduction**

**Clustering and Dimension Reduction**

**Neural Net Training**

**Conclusion**

**Sources**

[1] Wine quality dataset source: https://archive.ics.uci.edu/ml/datasets/wine

[2] Pet adoption dataset source: https://www.kaggle.com/c/shelter-animal-outcomes