Unsupervised Learning

Luke Kehoe
lkehoe6

luke.kehoe@gatech.edu

1 INTRODUCTION

This assignment will be investigating the nature and optimization of clustering and dimensionality reduction algorithms.

The two data sets I have chosen for my analysis were the Cleveland Heart database (CHD) and the Auto MPG data set (AMD), both sourced via UCI. The respective goal of the machine learning techniques being implemented on these were predicting the presence of heart disease and the classification of a cars fuel efficiency.

The CHD is made of a 14 attribute subset of health bio-metrics from hospital patients taken from the original set of 76 along with a binary target attribute to indicate the confirmed presence of heart disease.

The AMD is made up of 8 automobile attributes including a continuous set of values for miles-per-gallon. The mpg is the attribute that set as the target attribute to predict. Fuel efficiency *in praxi* is considered poor, typical or very good. So the continuous nature of this attribute needed to be pre-processed in order to make the problem a classification problem.

2 RESULTS AND ANALYSIS

2.1 Clustering

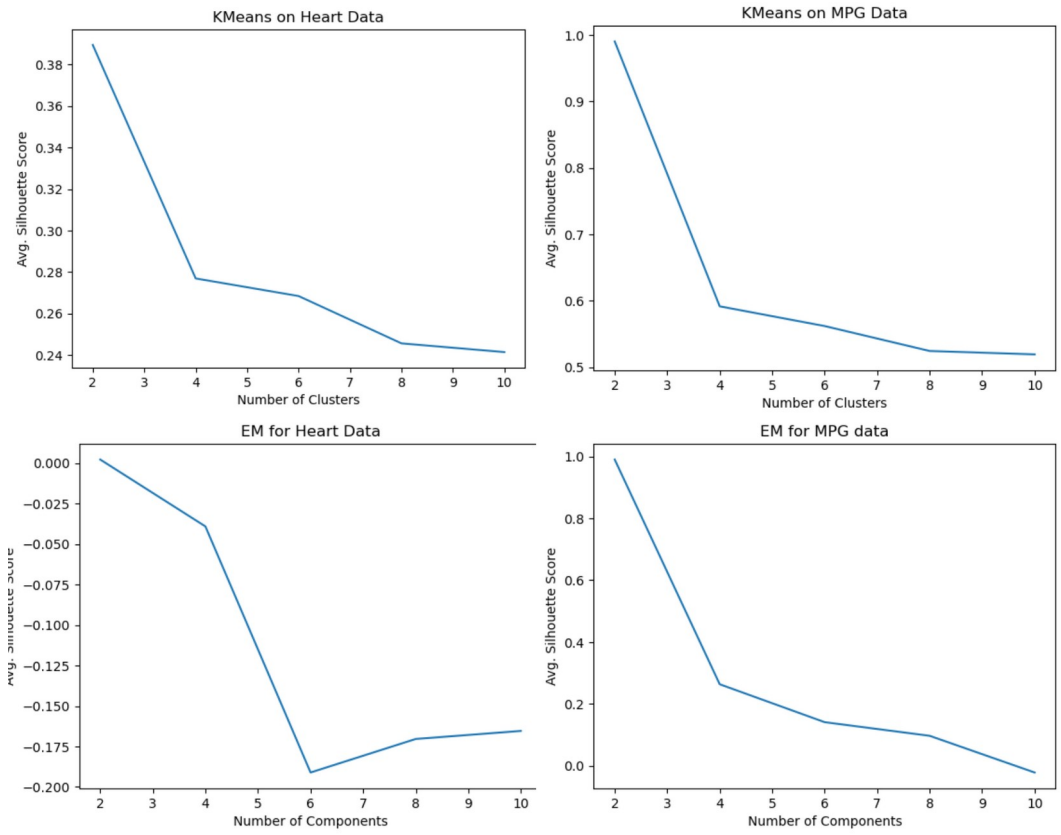

Chart 1

I chose to use silhouette scoring as the metric to measure the effectiveness of clustering using different component amounts. This is a function that scores based on the relationship of distance both between and within clusters. Silhouette score of 1 is the maximum value with -1 being the lowest.

As can be seen in chart 1 there was no increase in positive cluster definition from a starting point of 2 components. This suggests the data in general does not have easily distinguishable relationships.

The AMD data was the only one of the two that was consistently 'good' starting with a maximum score of 1 for both clustering algorithms. While the CHD had much lower scores which only declined with added components.

This behavior suggests that the clustering algorithms are over-fitting by trying to group more specifically than two. The clusters could be too small given the number of samples available in these data-sets increasing the noise to signal ratio of this classification into groups. Given the sample size and feature count this seems to be a reasonable explanation. With the fewer features on AMD it may explain the better score if there are strong relationships between them. It suggests the increased features and maybe more complex relationship between them is leading to a lower score for the same number of components in CHD.

3 DIMENSION REDUCTION

3.1 Introduction

For the dimension reduction analysis I chose to use Explained Variabilty (EV) where possible as the metric to score the algorithms by. Dimension reductions purpose is to increase the efficiency of building potential models, remove noisy data (thereby mitigate over-fitting) and improve interpretability.

EV is a good measure for this as it is a metric for how much of the original variability is explained by the components and how well it's preserved in the lower-dimension sub-space generated by the algorithms, such as PCA.

For t-SNE, due to it's use of probability distributions, I chose KL Divergence which measures the amount of information loss between distributions. A lower score being better.

For ICA it also uses distributions but uniquely is hence measuring how non-Gaussian so for this naturally a measure of kurtosis was used with higher being better.

3.2 Primary Component Analysis

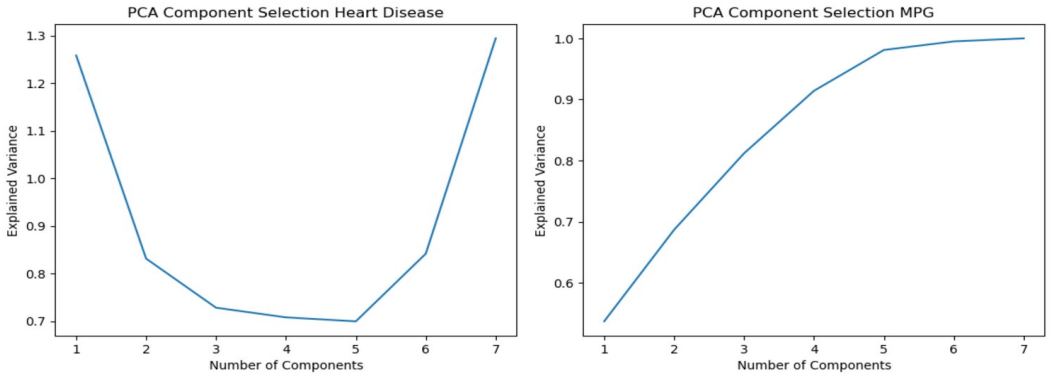

Chart 2

In chart 2 it can be seen that, for CHD, just a couple components capture a lot and for AMD additional components continue to increase a score that starts quite low.

This suggests that the AMD may actually be more complex and require more axis to make it well organized.

An obvious visually striking difference is the shape of the curves, specifically that EV for CHD reverses trend and increases after five components. I'd postulate that there may be strong correlation between among the data that is re-captured after a certain point.

3.3 Random Projection

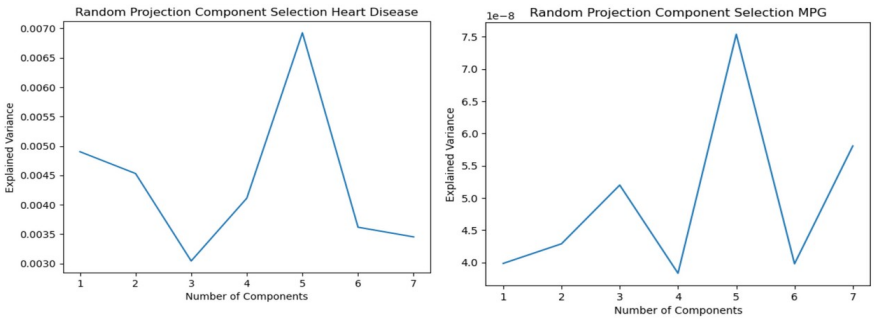

Chart 3

As can be seen in chart 3, the explained variance of the datasets is remarkably similar when using RP. Given the different scores using other measures this could be explained by the the statistical properties of the data sets being similar and their somewhat similar size. RP does not rely on specifics, but as the name implies, relies on the statistical properties of the data.

3.4 T-Distributed Stochastic Neighbor Embedding

T-Distributed Stochastic Neighbor Embedding (t-SNE) is non-linear and gives insight into the arrangement of higher dimensional space. It is used to reduce dimensions via distribution comparisons between higher and lower spaces. It does this at the cost of being computationally more expensive.

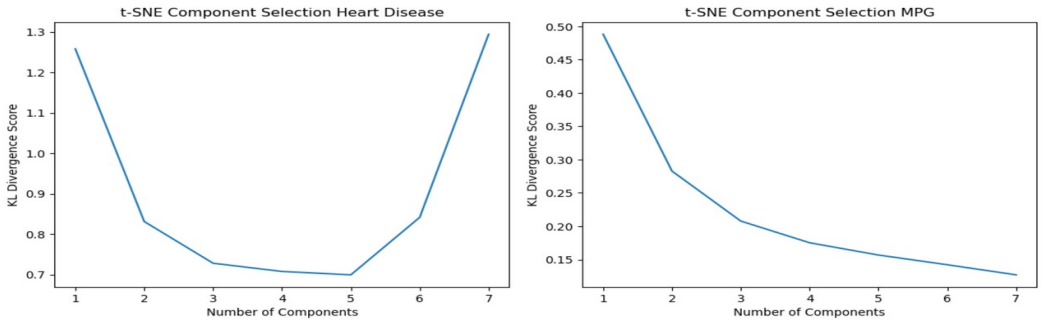

Chart 4

In the AMD, component count continued to improve the KL Divergence score (measuring preserved similarity) as it went up. While CHD worsened after a best measure at five components. This would suggest the AMD has a clearer structure in higher dimensional space and is less noisy. If CHD is noisier and more complex than it will be more difficult to preserve similarity of the lower dimensional spaces with more components. A type of over-fitting.

3.5 Independent Component Analysis

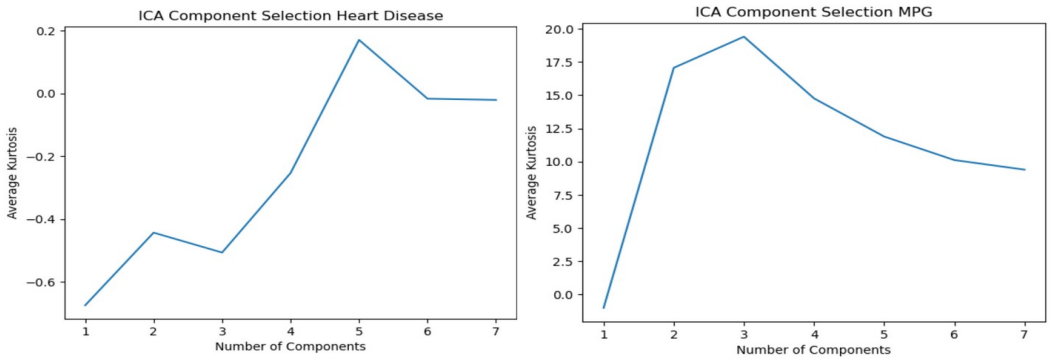

Chart 5

In the above charts the higher the kurtosis the score the better. We are looking to identify independent variables/components. The shape of both curves is expected despite the differences in the underlying datasets. ICA tends to preferebly select the most independent components so as the number of components increases the scores of kurtosis will eventually be expected to decrease. A higher peak for CHD suggests more identifiable independent features present.

4 NEURAL NETWORK ON DIMENSIONALLY REDUCED CHD

4.1 Reduced Data on Neural Network

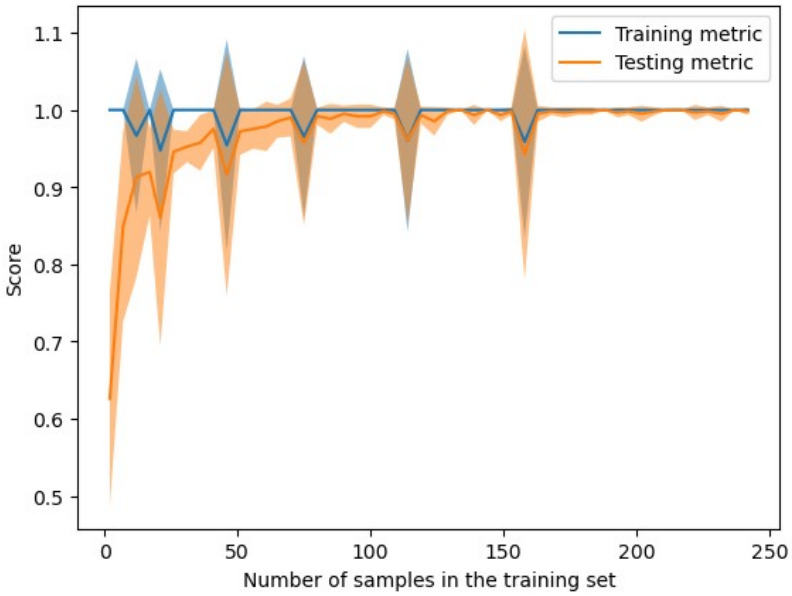

Chart 6

There a couple interesting results from pre-processing to reduce dimensionality. Chart 6 shows the results of the NN for PCA but as can be seen in chart 7 below results were similar for the other algorithms also.

The first interesting outcome is the improvement of the accuracy scores in both training and test data. I would hope to think this is from the prevention of over-fitting that comes from the reduction of noise PCA provides. However, the second standout result is the spikes in deviation over sample sizes. If these appeared somewhat random it could be just that. In this case they are seemingly regular to a degree which leads me to believe there is important information that has been lost and is also over-fitting to noise making it susceptible to it when it shows up.

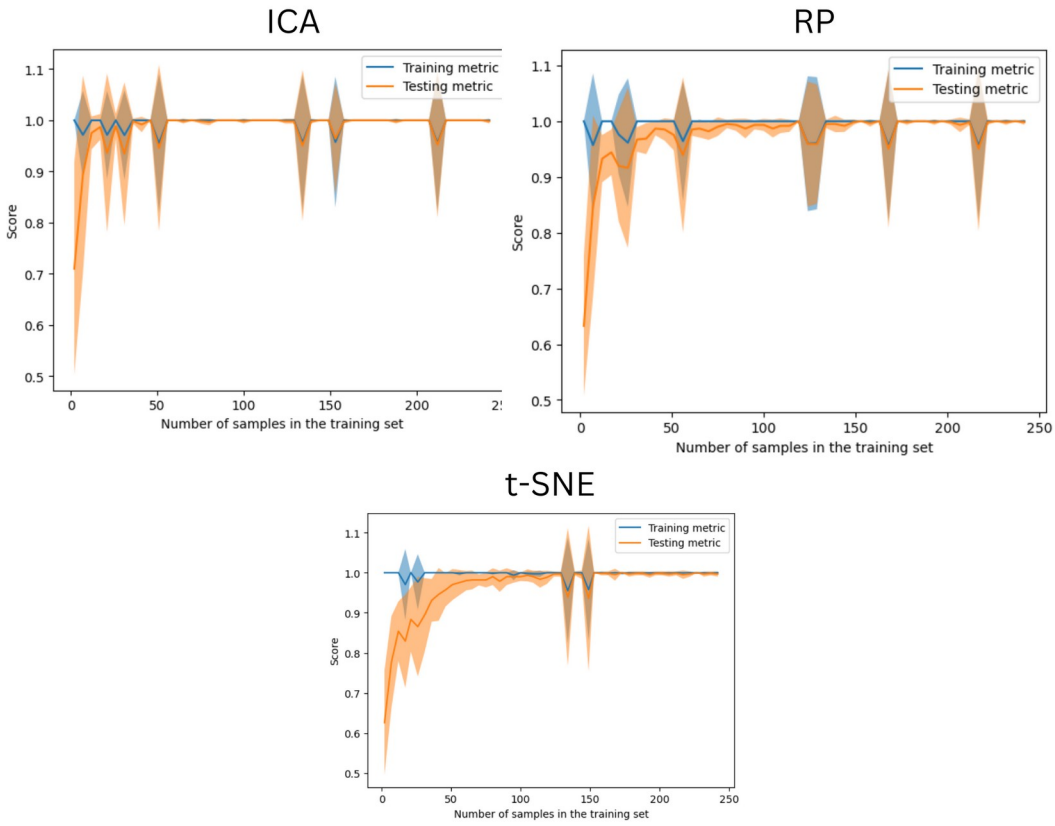
ICA
RP
t-SNE
Training metric
Testing metric
Score
Number of samples in the training set

Chart 7

5 NEURAL NETWORK & REDUCTION WITH CLUSTERING

A final experiment performed was using the clustering algorithms, K-Means and Expectation Maximization, as an additional dimensional reduction or in other words; as new features.

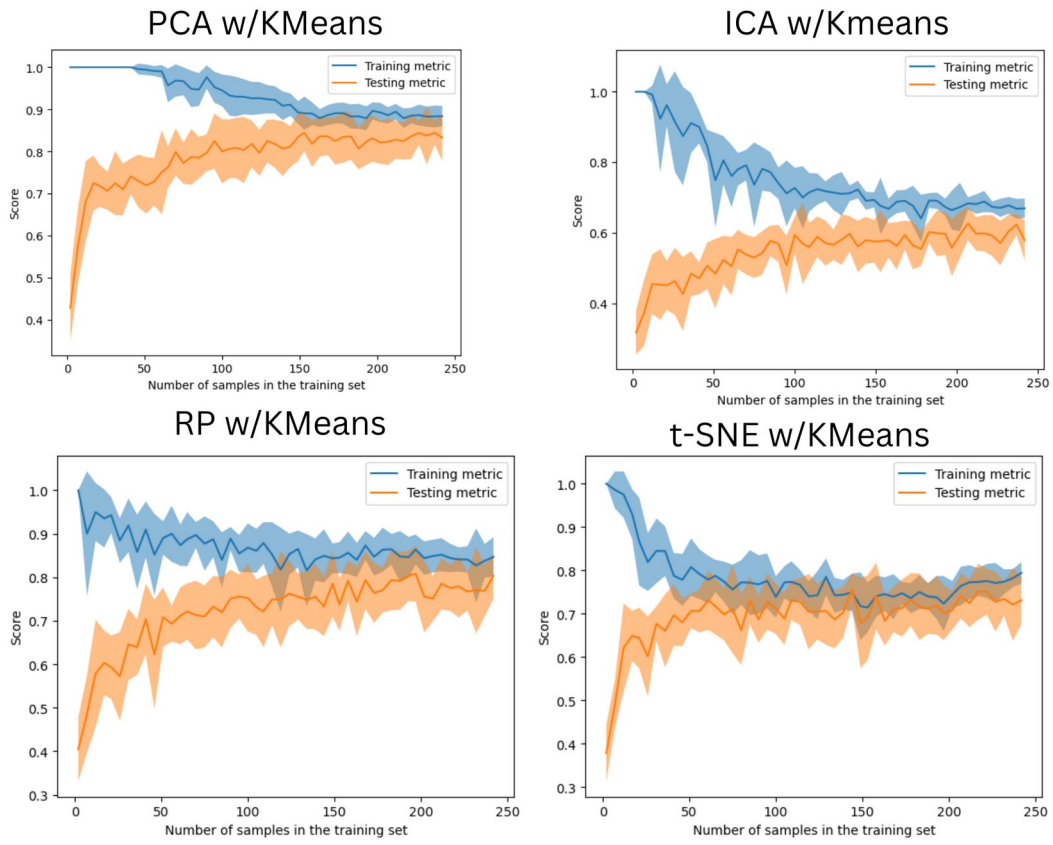

Chart 8

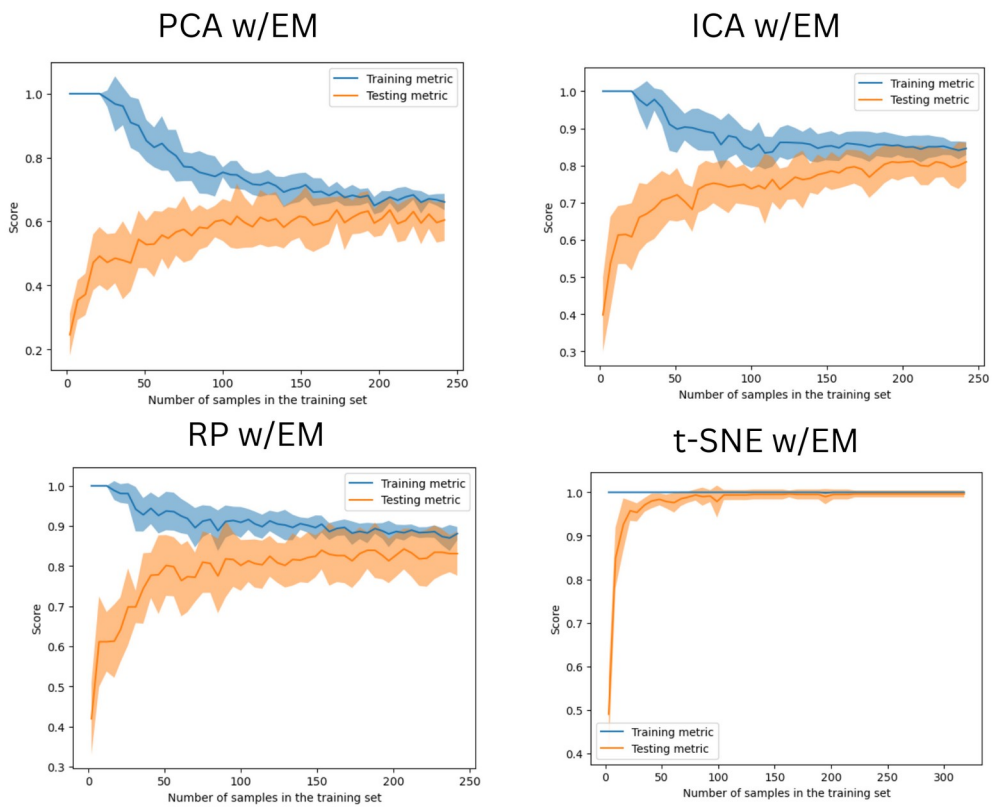

Chart 9

The above charts represent the learning curves of the new data which includes the corresponding component as a feature on each sample. What is clear when comparing to the NN performance to that on the data prior is the decreased performance and increased deviation. Even on the training data.

My interpretation of this outcome would be that it is primarily caused by the loss of information that is being introduced by adding the clustering as a feature. Doing so may have reduced the amount of variation the feature set captures and introduced bias. Particularly so as a compounding but independent problem could be the loss of information. Separately and together these make it more difficult to model.

Finally this technique could introduce non-linear relationships and mutual influence to the data. Again, making it difficult to model relatively.

6 REFERENCES

1. https://scikit-learn.org/stable/
2. https://matplotlib.org/