CS 7641 Assignment 3: Unsupervised Learning and Dimensionality reduction

Nimesh Chudasama

nimesh@gatech.edu

***Abstract*—** This paper investigates the relationship between two clustering algorithms and four dimensionality reduction algorithms. In particular, it highlights how datasets cluster when run with and without dimensionality reduction algorithms. In particular we explore K Means and Expectation Maximization on their own, and we explore these two clustering algorithms after we run PCA, ICA, Randomized Projection, and TSVD on them.

1 DATASETS

The two datasets I selected were distinct in nature. The two datasets differed heavily in terms of number of attributes, classification task, and number of instances. One was a binary classification task and the other contained a multi-class classification task.

Both datasets were native from SKLearn's dataset collection.

1.1 Breast Cancer Dataset [1][2][3]

The Breast Cancer dataset contained 569 instances, 30 numeric attributes and was a binary classification task. The attributes contained information about the benign or malignant tumor. This dataset contained 212 instances of malignant tumors and 357 instances of benign tumors.

A lot of medical diagnosis can be improved through machine learning. Machine learning algorithms today are just as good as diagnosing ailments of the human body than doctors [4]. One can argue even better, as they don't have emotional bias, although they do have inductive bias depending on the algorithm used to obtain accuracy.

1.2 Digits Dataset [5][6][7][8]

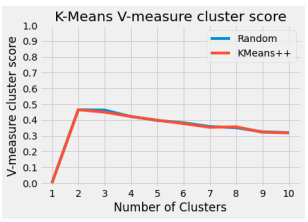
The Digits dataset contained 1797 instances and 64 attributes. The attributes were from the 8x8 image which contained pixels in the range 0-16. Where 0 was pure white, 16 was purely black and everything in between was a shade of gray. There are ten classes in this dataset, namely the digits 0 through 9.

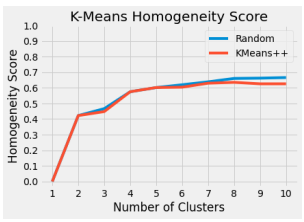
OCR or optical character recognition is an important task that many companies are trying to solve. There are many open source algorithms present for this task, such as TesseractOCR, but nonetheless it would be interesting to compare a binary classification task versus a multiclass classification task.

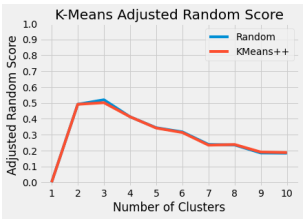
2 CLUSTERING ALGORITHMS

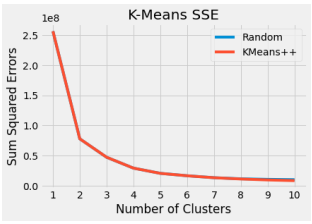
2.1 Breast Cancer Dataset K-Means

The obvious choice for clusters is two with k=2, due to the fact that this is a binary classification task. Nonetheless, generating some graphs would be useful to see what the value should be.





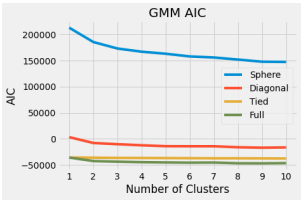


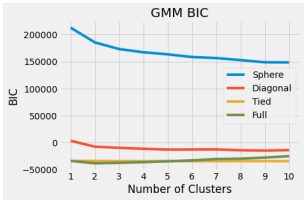


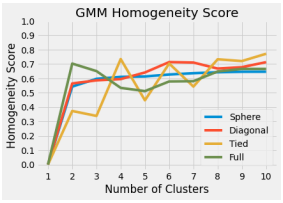
The four graphs generated are V-measure cluster score, Homogeneity score, Adjusted Random Score and Sum of Squared Error. Looking at V-measure cluster score, the max lies at 2 clusters. Interestingly, the homogeneity score is at 10 clusters. This could be that choosing 10 random points as K, yielded a better score. Homogeneity score is when each cluster only contains data points which are members of a single class. Investigating further would help break ties. In Adjusted Random Score, the optimal number of clusters is 3. For the sum of squared errors the elbow point is at 2. I further verified using a library called kneed, which has a kneelocator class which returns where the elbow/knee point is. Although the K-Means adjusted random score had 3 clusters as it's optimal value, 2 is extremely close to that value and two other measures produced a value of k=2. Therefore k=2 is the optimal choice for this dataset which is what we inferred from our intuition.

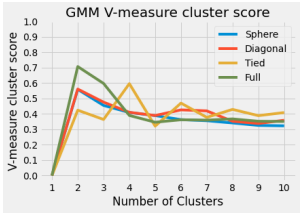
2.2 Breast Cancer Dataset Expectation Maximization

The EM algorithm could result in a different number of clusters than our inferred two clusters due to Gaussian Mixture variances.







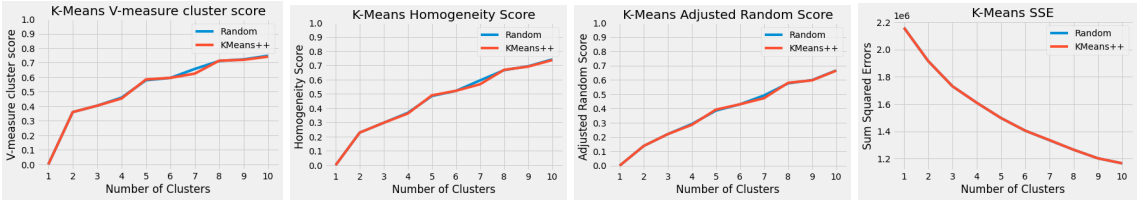


Four different types of Gaussian Mixture algorithm types were run, spherical, diagonal, tied, and full. And all four of these were plotted on AIC, BIC, Homogeneity Score and V-measure cluster score. Sum of Squared Errors was not run due to different covariance matrices. In AIC and BIC the lower the better, in both instances, when the number of clusters is around 2 or 3 it smooths out. On Homogeneity score, if we observe the different algorithms, the max peak occurs at 10 clusters. And for V-measure cluster score the max is at 2 clusters which we expected. Therefore we choose 2 clusters with EM.

2.3 Digits Dataset K-Means

The obvious choice for clusters is 10 due to there being 10 classes. They should all be distinct in a 64

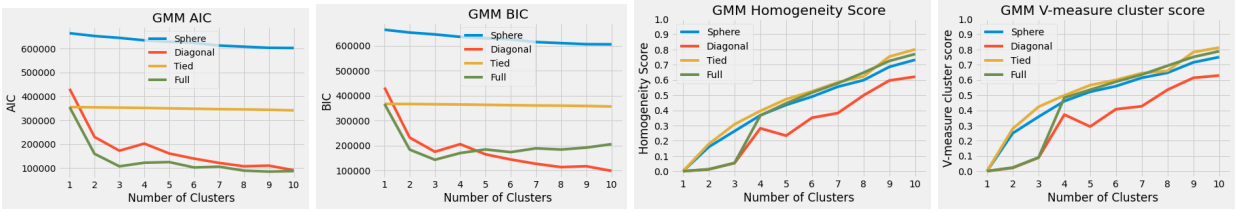
dimension space. We generated four graphs to visualize this.



The four graphs were again V-measure cluster score, homogeneity score, adjusted random score and sum of squared errors. Looking at all these graphs, it's immediate to realize that that 10 clusters is the most obvious choice. In all the graphs, 10 clusters was the optimal choice. Therefore we choose k=10 which matches with our intuition for our dataset.

2.4 Digits Dataset Expectation Maximization

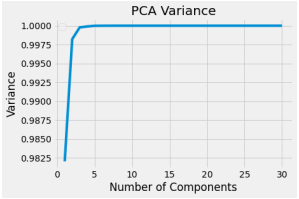
The EM algorithm could result in a different number of clusters than our inferred ten clusters due to Gaussian Mixture variances.



Again, four different Gaussian Mixture algorithm types were run, spherical, diagonal, tied, and full on AIC, BIC, Homogeneity Score and V-measure cluster score. All of these immediately showed that the optimal cluster is 10 clusters. Therefore we choose 10 clusters, which is in line with what we thought was the optimal choice for the ten digits.

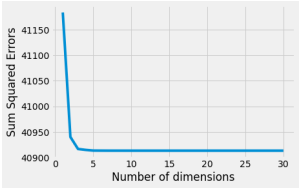
3 DIMENSIONALITY REDUCTION ALGORITHMS

3.1 Breast Cancer Dataset PCA



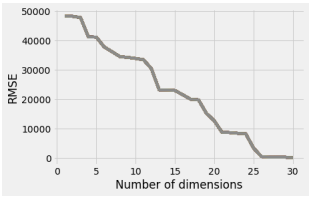
For this plot, the cumulative sum for the variance ratio was used. Surprisingly, we can reduce this algorithm from 30 dimensions to 15 dimensions. and it converges to 1.0 variance. This means that according to PCA, half the dimensions are not useful to explain the variance. To push limits on how much this truly matters, we chose 2 dimensions to explain 99.8% variance of the dataset.

3.2 Breast Cancer Dataset ICA



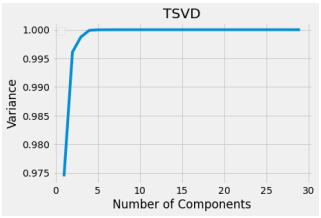
Looking at the SSE of the data reconstruction the most obvious choice of dimensions for the breast cancer dataset is 3 dimensions. This has the lowest SSE and it plateaus from there.

3.3 Breast Cancer Dataset Randomized Projection



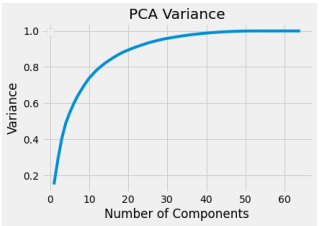
On this plot, I used the RMSE to check the reconstruction error. I ran this algorithm 64 times to see how the data reconstruction varied, but it was similar with the lines pretty much on top of each other. Using the data, the optimal number of dimensions is 20, as it has quite low error and it plateaus from there.

3.4 Breast Cancer Dataset TSVD



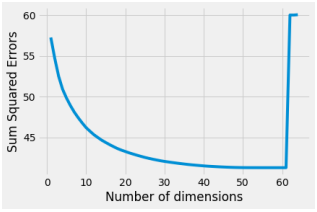
Truncated SVD (LSA) is a bit different. It's similar to PCA but the estimator does not center the data before calculating the SVD. Again It looks like the variance of the algorithm is high just at 2 dimensions. Therefore we chose 2 dimensions after running TSVD which captures 99.6% of the variance.

3.5 Digits Dataset PCA



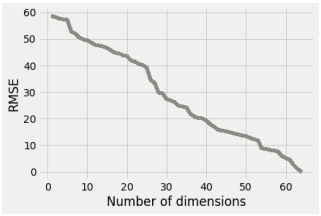
The Digits dataset PCA graph is quite different from the Breast Cancer graph. On this graph, the optimal variance is when there are around 25 components. This means we can reduce the dimensions from 64 to 25 components to capture 93.3% of the variance.

3.6 Digits Dataset ICA



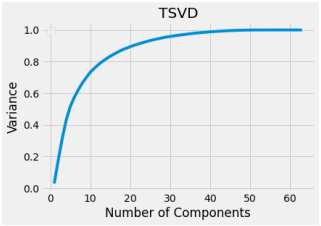
Looking at the SSE of the data reconstruction the most obvious choice of dimensions for the digits dataset is 30 dimensions. The SSE doesn't lower that much, and with the curse of dimensionality, getting around half the dimensions to reconstruct the data helps a lot.

3.7 Digits Dataset Randomized Projection



On this plot, I used the RMSE to check the reconstruction error. I again ran this algorithm 64 times. Using the data, the optimal number of dimensions is quite unclear, nonetheless we can see a sharp decline at around 25 dimensions that continues to around 42 dimensions. Therefore we chose 42 dimensions with Randomized Projection

3.8 Digits Dataset TSVD

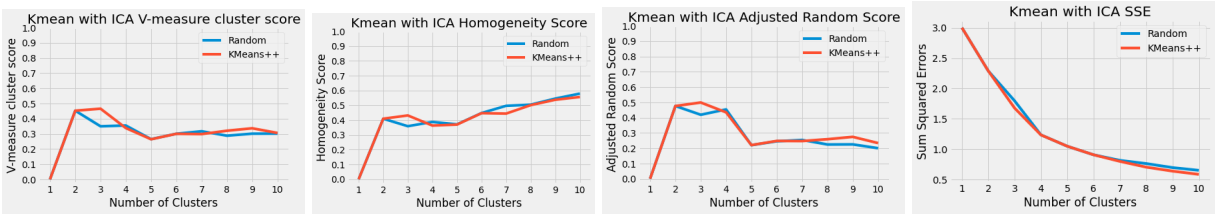


It looks like the variance of the algorithm is high at 30 dimensions. Therefore we chose 30 dimensions after running TSVD which captures 95.9% of the variance.

4 CLUSTERING WITH DIMENSIONALITY REDUCTION

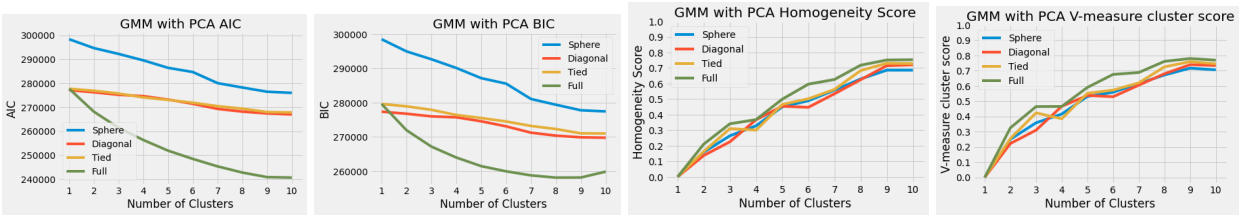
Both K-Means and EM were run after dimensionality reduction for both datasets, which means 16 experiments were run. These were the most interesting ones found.

4.1 Breast Cancer Interesting Finds



Running ICA to reduce dimensions to 3 and then clustering with K-Means to find the number of clusters yields an optimal of 3 clusters. This probably means that data was in fact lost and when recomputing centers, starting with k=3 clusters yields better scores.

4.2 Digits Interesting Finds



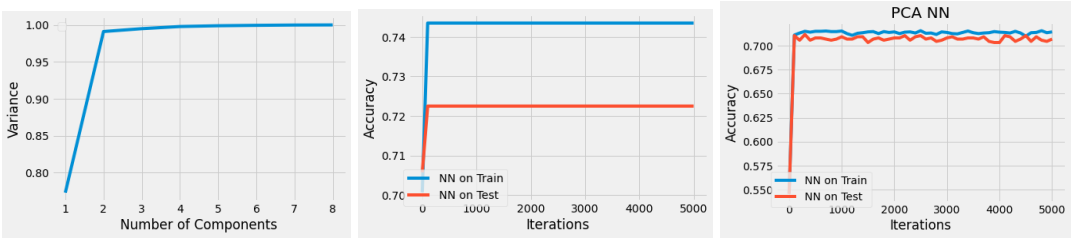
Using PCA, to reduce dimensionality to 25 dimensions on EM, it seems that the optimal clusters changes from 10 to 9. This means that some data is lost and it finds that 9 clusters better represent the data. Surprisingly, using K-Means, the optimal number of clusters is still 10, which means that if you use the mean to compute the center, there are still 10 distinct centers but K-Means also finds that 9 clusters follow closely in terms of the graphs.

5 NEURAL NETWORK WITH DIMENSIONALITY REDUCTION

5.1 Abalone Dataset

The Abalone dataset has 4177 instances, 8 attributes and a Gaussian Distribution with respect to age. I divided the dataset into six buckets of age from 0-5, 5-10, 10-15, 15-20, 20-25, and 25+.

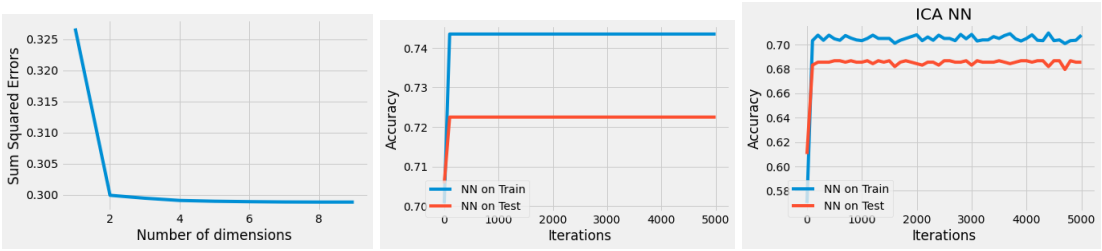
5.2 Abalone Dataset PCA



For all the examples. The left is the respective dimensionality reduction. The middle image is the old NN with it's optimal parameters, with no dimensionality reduction and the right picture is the NN after dimensionality reduction was used.

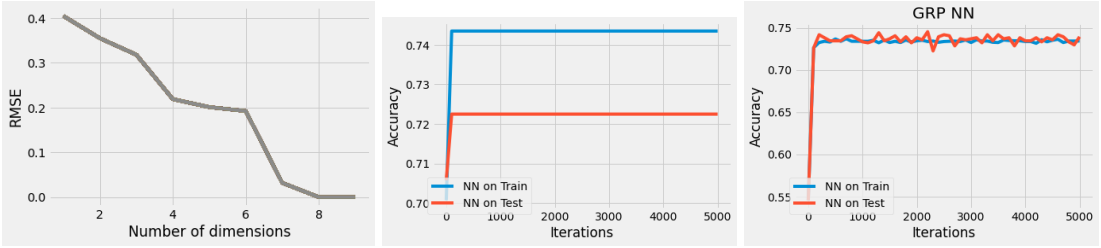
For PCA, the optimal number of components was 2, as that had the largest increase. The performance of the NN suffered through dimensionality reduction by about 3.5% for train and 2% in the test set.

5.3 Abalone Dataset ICA



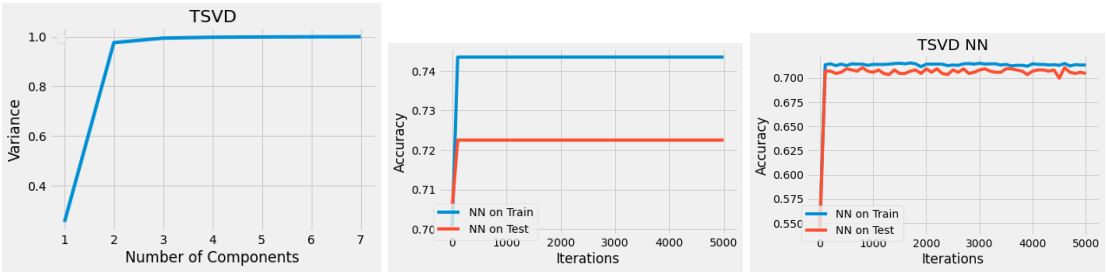
For ICA, the optimal number of components was 2, as that had the lowest Sum Squared Errors. The performance of the NN suffered through dimensionality reduction by about 4% accuracy drop in train and 4% accuracy drop in test.

5.4 Abalone Dataset Randomized Projection



Using Randomized Projection, the optimal number of dimensions was 7, as that had the lowest RMSE. It ran 64 times and the line seemed to be consistent. In this case with 1 dimension reduced it actually did better in the test set by 2% and the same in the training set, maybe the data reconstruction allowed it to generalize better.

5.5 Abalone Dataset TSVD

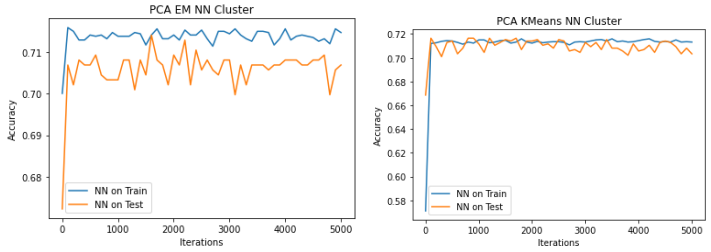


For TSVD the optimal number of dimensions was determined to be 2. In this case, The neural network performed 4% worse in train and approximately 2% worse in the test set.

6 NEURAL NETWORK WITH DIMENSIONALITY REDUCTION AND CLUSTERING

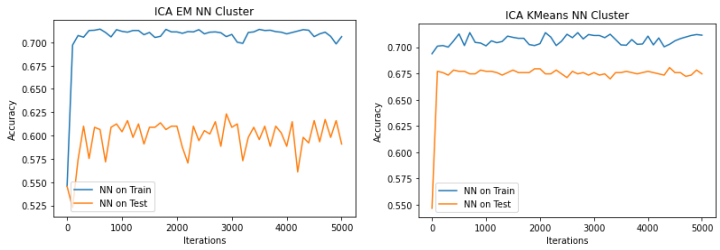
In all the following scenarios, I used 6 clusters as that was the amount of distinct clusters there should be. I was wondering if adding these clusters would improve performance in any way as there would be more attributes for the data.

6.1 Abalone Dataset PCA KMeans/EM



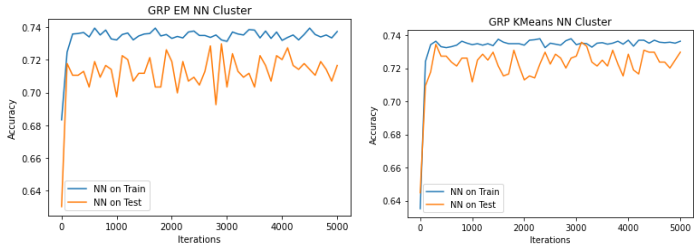
In this scenario, PCA was run with 2 dimensions as that was the optimal dimensions in respect to variance as mentioned previously. In this scenario, adding KMeans clusters as attributes was better than adding EM clusters as attributes to the dataset. KMeans performed approximately 1% better in test and train. KMeans boasts a similar performance to the original attributes on the test set.

6.2 Abalone Dataset ICA KMeans/EM



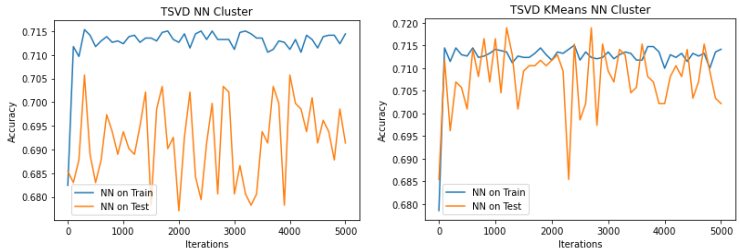
ICA was run with 2 dimensions and both EM and KMeans clusters were appended to the original dataset. KMeans performed significantly better on the test dataset than EM.

6.3 Abalone Dataset Randomized Projection KMeans/EM



Randomized Projections was run with 7 dimensions and again EM and Kmeans clusters were added as attributes to the dataset. KMeans performed better than EM in this scenario.

6.4 Abalone Dataset TSVD KMeans/EM



This was interesting to me. I used TSVD with 2 dimensions and added KMeans and EM clusters as attributes. The accuracy varied wildly. It seems KMeans performed better and has less overall range when comparing the accuracies.

6.5 Neural Network Summary

Reducing the dimensions, greatly reduced training time without affecting accuracy too much as they run on lower space state.

Adding more attributes using KMeans and EM didn't affect accuracy too much, and since there were attributes to add, it actually increased training times.

7 REFERENCES

[1] W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

[2] O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.

[3] W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.

[4] https://www.medicalnewstoday.com/articles/326460

[5] C. Kaynak (1995) Methods of Combining Multiple Classifiers and Their Applications to Handwritten Digit Recognition, MSc Thesis, Institute of Graduate Studies in Science and Engineering, Bogazici University.

[6] Alpaydin, C. Kaynak (1998) Cascading Classifiers, Kybernetika.

[7] Ken Tang and Ponnuthurai N. Suganthan and Xi Yao and A. Kai Qin. Linear dimensionality reduction using relevance weighted LDA. School of Electrical and Electronic Engineering Nanyang Technological University. 2005.

[8] Claudio Gentile. A New Approximate Maximal Margin Classification Algorithm. NIPS. 2000.