CS7641 Machine Learning (Fall 2020)

Assignment 3 – Unsupervised Learning

Sergiy Palguyev (gtid: spalguyev3)

I. Abstract

This assignment explores clustering and dimensionality reduction algorithms. The goal of these algorithms is to simplify datasets to make it easier for different learners to learn and predict outcomes from them. The data will be treated as "unsupervised" to best evaluate how these algorithms can group and cluster together various data points. The goal of this paper is to analyses the strengths and weaknesses of these algorithms in detail and with an application to Neural Networks.

II. Datasets

Two data sets are chosen to be used for this analysis from the UCI database, same as those from Assignment 1.

A. *Banknote Authentication Dataset*

The Banknote Authentication dataset [1] consists of four continuous values representing image attributes used for evaluation of an authentication procedure for banknotes. The classification is binary, classifying each banknote as authentic, or not.

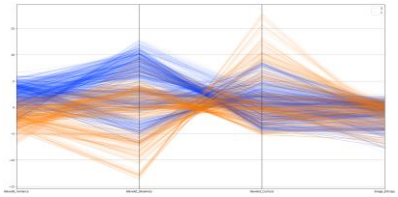


Figure 1: Banknote Parallel Plot

Figure 1 show the relationship between the attributes of the classes contained in the dataset. It is evident that one class has a much lower Wavelet Skewness and Kurtosis value range than the other class. Similarly Wavelet Variance has higher values for one class than the other.

B. *Letter Recognition Dataset*

The Letter Recognition dataset [2] consists of sixteen attributes which define the image representation of a written letter. The goal of this dataset is a multi-class classification of the letter of the English alphabet, or twenty-six letters.

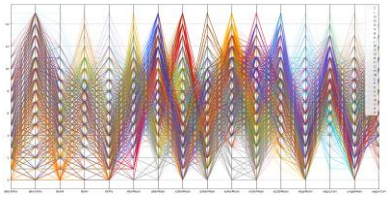


Figure 2: Letter Parallel Plot

Figure 2 shows the relationship between the various attributes of the different classes in the dataset. Not much useful information can be garnered from this view of the data, other than an appreciation and awe for its complexity, and need for dimensionality reduction.

III. Clustering Algorithms

Clustering algorithms seek to learn, from the properties of the data provided, the optimal division and labeling of the different groups of points.

A. *K-Means (CM) Clustering*

K-means searches for clusters in what is an unlabeled multi-dimensional dataset; unlabeled, thus un-supervised. K-means assumes, in generality that the center of each cluster is the mean of all points belonging to that cluster. K-means attempts to group data clusters into spheres and generally should do poorly with non-sphere shaped data.

B. *Expectation Maximization (EM) Clustering*

Expectation Maximization is a special case of k-means algorithm. The algorithm begins with a "guess" of some cluster center, assign data points to the closest cluster and then set the centers to the mean. This process is repeated over and over until convergence is attained.

IV. Dimensionality Reduction Algorithms

A. *Principle Component Analysis (PCA)*

The basic principle of PCA is determining the data components and variance which represent the importance of a feature in describing the distribution of the data. Dimensionality Reduction through PCA consists of zeroing one or more of the smallest principal component vectors. This will result in a lower dimensional projection of the data while preserving variance.

B. *Independent Component Analysis (ICA)*

ICA is a good alternative to PCA as it is better suited for working with data which may have non-Gaussian noise, not necessarily orthogonal in the original feature space. If the data reflects only Gaussian processes, PCA and ICA results should be similar.

C. *Randomized Projections (RP)*

RP projects high dimensional data to lower dimensional space using a random matrix with Gaussian distribution.

D. *Factor Analysis (FA)*

Conversely to PCA, FA attempts to determine the data dependency on factors, rather than features. Reducing the number of dependent factors can also decrease dimensionality of multi-dimensional data sets.

V. Performance Metrics

A. *Sum Squared Errors*

The elbow method applied to Sum Squared Errors graph provides a good assertion of optimal cluster number. Whereby the elbow provides a pivot point beyond which adding more clusters provides diminishing returns.

B. *Silhouette Score*

CS7641 Machine Learning (Fall 2020)

Assignment 3 – Unsupervised Learning
Sergiy Palguyev (gtid: spalguyev3)

This is an intrinsic method of measuring performance with an absence of "ground truth". The Silhouette coefficient, valued between 1 and 0, indicated whether the clusters are compact around their centroid or overlapping with others respectively.

C. AIC and BIC Scores

AIC is an estimate of distance between unknown True and fitted True of the model, whereby a lower AIC score implies the two are close. BIC score computes the likelihood that a model is the "true" model and also considers a lower BIC value as the more likely. BIC penalizes model complexity more heavily than AIC.

D. Homogeneity, Completeness, Mutual Information

Homogeneity (HM) is a measure of the amount of data points in a cluster with all belong to only one class. Completeness (CM) is a measure of all points belonging to the same class, also belonging to the same cluster. Normalized Mutual Information (NMI) measures the agreement of label assignments on the dataset.

E. Accuracy, F-1 and Confusion Matrix

Precision Score evaluates how many True Positives are detected out of the entirety of all positives (True Positive and False Positive). The Recall Score evaluates how good the model can determine True Positives against classes which were mislabeled by the model as negative, but were actually positive. The F-1 Score seeks to combine the Precision Score and the Recall Score into one evaluative metric, or harmonic mean of the two metrics. This metric prefers classifiers which have similar values for both Precision and Recall scores.

VI. Analyses

A. Banknote Dataset - No dimensionality reduction

Since it is difficult to display multidimensional data effectively, a representative scatter graph of the first three features is plotted to give a general representation of data behavior.

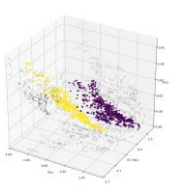


Figure 3: Banknote Dataset

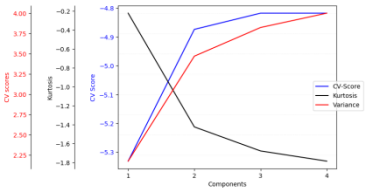


Figure 4: Banknote Component Plot

As seen in Figure 3,the 2 classes of data have narrow-to-wide spread behavior coming from the relationship observed at the YZ plan projection. The XZ and XY relationship projections show a fairly wide spread amongst data relationships. It is presumed that reduction along the data features projected on the YZ plane should lead to a visible dimensionality reduction.

A Component plot in Figure 4, shows how Cross-Validation, Kurtosis and Variance. It is interesting to notice that if the dataset is reduced to three components, the CV value does not decrease, and the Variance remains relatively high. This will be the number of components used for dimensionality reduction algorithms below.

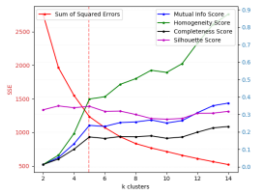


Figure 5: KM Cluster Plot

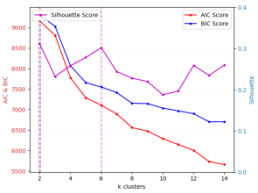


Figure 6: EM Cluster Plot

Performing KM dimensionality reduction, Figure 5, the plot of SSE against Homogeneity, Completeness, Normalized Mutual Information and Silhouette Score shows a clear picture of best k cluster number. The elbow of SEE is computed to be at k=5. This fits very neatly with the other metrics, showing an "elbow" in all others indicating that increasing cluster numbers adds little value.

For EM dimensionality reduction, Figure 6, the SSE plot versus AIC and BIC scores shows a similar result to KM with an optimal number of 6 clusters according to Silhouette Score. AIC and BIC scores do not display as prominent as a pivot as the KM metrics do. As BIC scores rate model correctness, it is interesting that it does not fall as fast as AIC scores with increased cluster size.

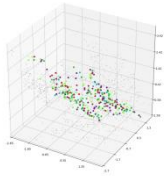
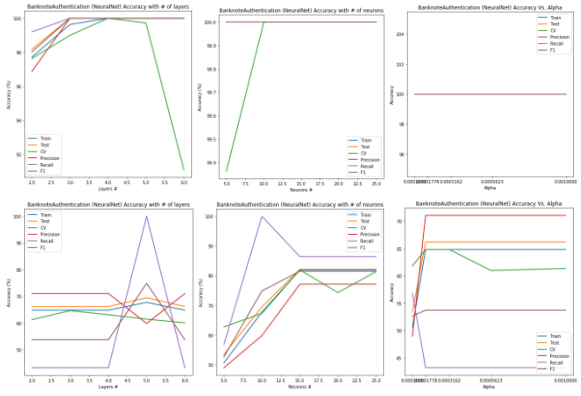


Figure 7: Clustering at k=5

Figure 7 displays one view of clustered data with 5 clusters. This view does not seem to show well clustered data, It is possible other views may show better clustering visually.



CS7641 Machine Learning (Fall 2020)

Assignment 3 – Unsupervised Learning

Sergiy Palguyev (gtid: spalguyev3)

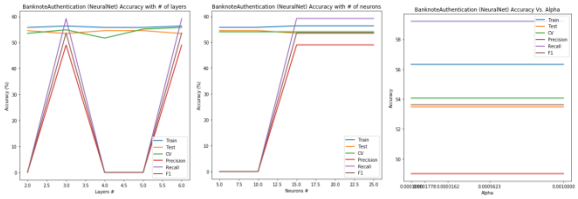


Figure 8: 3x3 Plot Cluster of Neural Network Tuning

Training Neural Networks on the included data shows some very interesting results. Figure 8 is a 3x3 matrix of NN training results. Column 1-3 represents Layer #, Neuron #, and Alpha # hyper-parameter tuning while rows 1-3 display original data, KM output data and EM output data as inputs. The first, immediate observation is that all metric scores drop with KM and EM as input data.

The scores for EM input data are all much lower than for KM. This fact is most likely due to the "guessing" nature of the EM algorithm in attempting to determine cluster centers. Also interesting to notice there are a few dips in training the Neural Nets on the EM dataset where the metric scores for Precision/Recall/F-1 scores drop to 0. This implies, with those values of Layer or Neuron numbers, either True Positives or True Negatives were 0. This is also corroborated with the Train/Test and CV scores also hovering slightly above 50%.

B. Letter Dataset - No dimensionality reduction

The Letter dataset has 16 attributes with complex relationships. The first 3 attributes are plotted as a scatter graph.

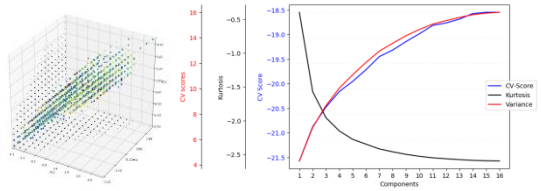


Figure 9: Letter dataset plot

Figure 10: Letter Components Plot

Figure 9 displays the relationship between the first three attributes. It may be observed that there is interesting behavior n variation of data parameters as values increase along the Y and Z axis as seen from the XY and YZ projections.

The CV/Variance/Kurtosis graph, Figure 10, displays very interesting behavior. It is evident from this graph that above 11 components, the difference gained in Variance is quite small in correlation with small gains in CV score. This value is used for components number in further evaluation.

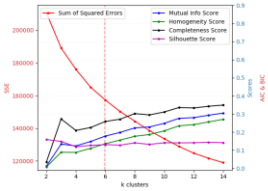


Figure 11: KM Cluster Plot

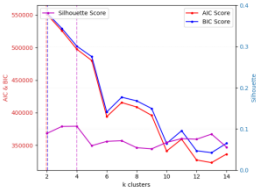


Figure 12: EM Cluster Plot

The KM Custer plot, Figure 11, show very little variability in increasing Homogeneity, Completeness or Mutual Information Score. Similarly, the Silhouette Score stays fairly flat with cluster increases. The SSE elbow method indicated a cluster number of 6 best describes the dataset.

The EM Cluster plot, Figure 12, is a fair bit more dramatic than the KM plot. AIC and BIC values fluctuate in tandem towards 0, showing a deterioration between unknown values and predicted values.

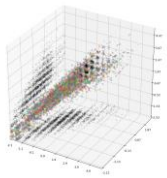


Figure 13: Clustering at k=6

The clustering scatter plot in Figure 13 shows a very interesting pattern. The 3D data does not show it but the 2D projections on the XA and XY planes show prominent data clustering into dedicated groups. This is a very interesting result considering no dimensionality reduction has been performed, and only a small subset of dataset features are being plotted. This result shows that clustering should prove to be a good transformation of data for Machine Learning algorithms.

C. Banknote Dataset - PCA Reduced

Principal Component Analysis evaluation allows projecting the resulting reduction back into original dataset space.

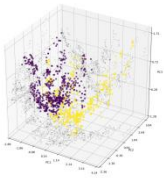


Figure 14: PCA Fit

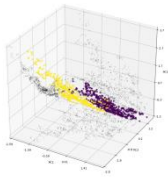


Figure 15: PCA Inverse

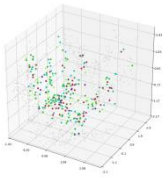


Figure 16: PCA @ k=5

Figure 14 shows the new projection of data into the reduced feature set. Inverse transforming the projected data back to the original dataset, Figure 15 plot shows a very interesting projection along the YZ plane where the variance in data points observed in Figure 3 is now "reduced".

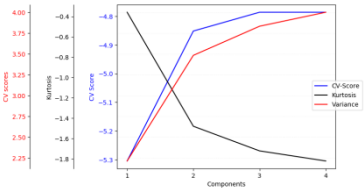


Figure 17: PCA Components Plot

The PCA components plot, Figure 17, still shows that 3 components account for most of the information contained in the dataset, with CV adding nothing and Variance only slightly increasing with the fourth feature.

CS7641 Machine Learning (Fall 2020)

Assignment 3 – Unsupervised Learning

Sergiy Palguyev (gtid: spalguyev3)

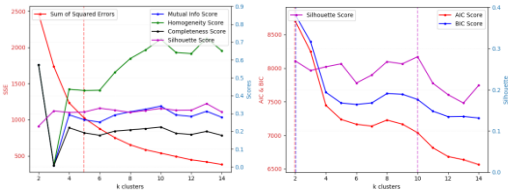


Figure 18: PCA KM Cluster Plot

Figure 19: PCA EM Cluster Plot

Figure 18, similar to Figure 5, shows an SSE with an elbow at 5 clusters. However, the HM, CM and NMI scores are all "stunted" with more clusters. The trends level off almost 10% below those from Figure 5. Figure 19, shows a very similar EM Cluster score plot as that of Figure 6. However, as cluster numbers rise to 14, the AIC score noticeably drops, implying an unknown True and fitted True have little correlation.

Figure 16, clustering plot shows the newly projected, reduced data set. The clusters of data are presented more clearly, but the clustering does not display good grouping.

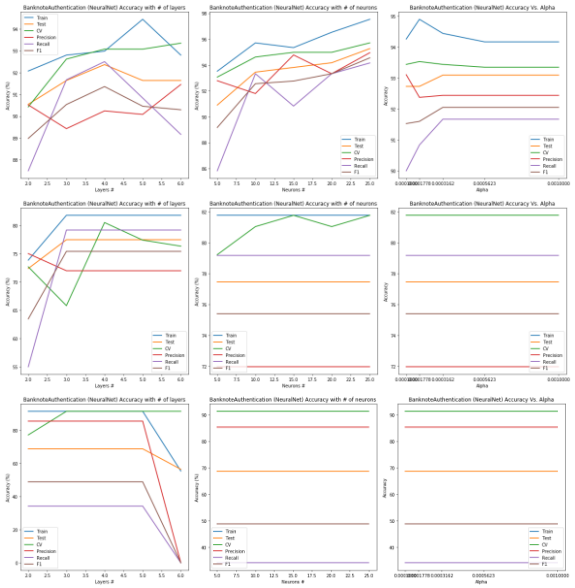


Figure 20:3x3 Plot Cluster of Neural Network Tuning

Figure 19, first row displays Neural Network trained on PCA transformed data only. It is evident form these plots that accuracy stays good throughout tuning process with better accuracy achieved with increase in Layer # up to 4 layers, Neuron # at maximum test values and Alpha # which levels off quickly, around 0.0001. Second row, PCA transformed and passed through KM, shows far poorer performance with accuracy scores between 70-80% or high Layer Number. Neuron and Alpha tuning did not change accuracy scores at all. Finally, the third row, PCA transformed and passed through EM shows plateaued performance across all tuning parameters except Layer #, decreasing with high Layer numbers.

D. Letter Dataset - PCA Reduced

The Letter dataset is much more difficult to display due to the large amount of features. With PCA, the new reduction in features can be inverse transformed back into the original space.

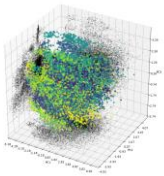


Figure 21: PCA Reduced

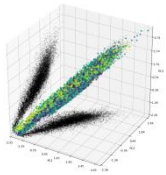


Figure 22: PCA Inverse

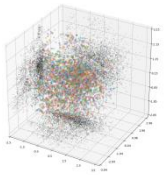


Figure 23: PCA @ k=8

As seem in Figure 23, the plot of fitted point inverse transformed back into original space displays a very interesting contrast for Figure 9, with no reduction. More specifically, it is obvious that the shape of the data is more narrow along the YZ, XY plane projections as well as the 3D plot of the data, showing a more uniform stretched set of data.

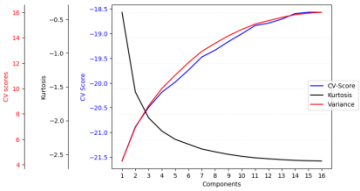


Figure 24: PCA Components Plot

The Components plot, Figure 25, looks very similar for PCA, leading to a value 11 components as a good reduction size for the remainder of tests while retaining good Variance and CV values and minimizing Kurtosis.

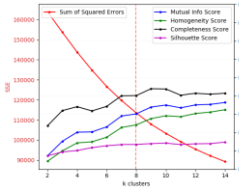


Figure 25: PCA KM Cluster Plot

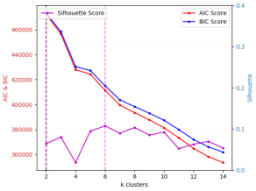


Figure 26: PCA EM Cluster Plot

The PCA KM cluster plot, Figure 26, displays an interesting behavior. As k-cluster number increases, the SSE value reduces, as expected, with an elbow value of 8 cluster. However, the NMI, HM, CM scores increase with an increase in cluster size. This implies that a test for large cluster numbers may lead to better results. The current elbow is an approximation from the executed tests, it is quite possible that with larger numbers, the elbow could be a different number of clusters where the NMI, CM and HM scores begin to plateau as data points become siloed instead of grouped.

The PCA EM plot, Figure 27, shows a clear maximum Silhouette score at 6 clusters, however, comparing the AIC/BIC scores with that of Figure 12 shows a much narrower band of scoring. Where in Figure 12 the trend stayed between 550k-350k, The PCA reduced EM plot AIC and BIC scores trended between 460k and 360k. The lower score on small cluster numbers shows a better correlation cluster

CS7641 Machine Learning (Fall 2020)

Assignment 3 – Unsupervised Learning
Sergiy Palguyev (gtid: spalguyev3)

structures for unknown data. This is in line with the use of dimensionality reduction. Figure 24, shows the cluster plot with k=8.

E. Banknote Dataset - ICA Reduced

As mentioned, ICA is good algorithm for non-Gaussian noisy data, otherwise the results should look the same as PCA.

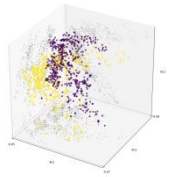
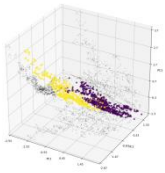
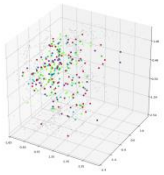




Figure 27: ICA Reduced Figure 28: ICA Inverse Figure 29: ICA @ k=6

The reduced data projection, Figure 27, immediately looks different from PCA reduction in Figure 14. Most datapoints seem to now reside in the lower X axis range while with PCA reduction the new projection seemed to reduce the Y axis range data. Regardless, the inverse transformation in Figure 28, shows a very similar picture to that of PCA inverse transformation in Figure 15, with a narrower variance in the data on the YZ plane projection.

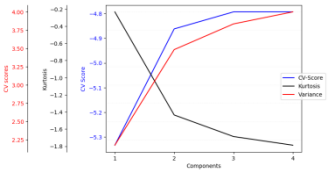
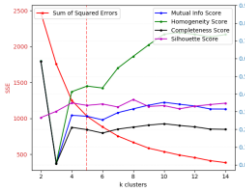


Figure 30: ICA Components Plot

The ICA components plot, Figure 30, retains the same behavior as PCA, demonstrating that only 3 components are needed to retain the variance of the entire dataset without looking CV or Variance too greatly.



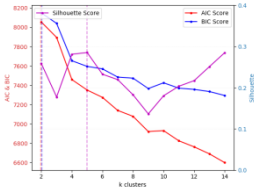


Figure 31: ICA KM Plot Figure 32: ICA EM Plot

Immediately an interesting result can be noticed. Both KM plot, Figure 31, and EM plot, Figure 32 agree that the k value of 5 is optimal. The KM plot at 5 clusters is logical as further increase in clusters leads to a sudden spike in homogeneity. The EM plot, similarly shows almost a "elbow" on the AIC and BIC graphs, indicating a good cluster value without increasing homogeneity in the EM clusters. It is also interesting to notice that the Y range for both plots closely matches that of PCA. Considering this, and the fact that with PCA the k values were 6 with KM and 4 with EM. It may be a good assumption to deduce that the noise in the dataset is Gaussian, considering the very similar results between PCA and ICA. Figure 29 shows the dataset clustering at k=5.

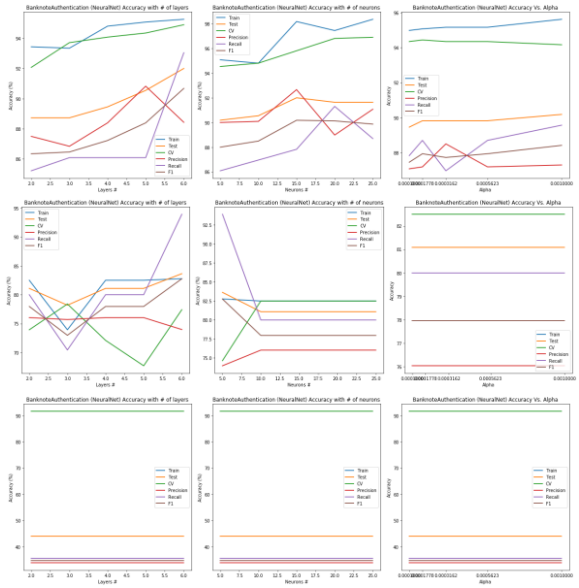


Figure 33: 3x3 Plot Cluster of Neural Network Tuning

The Neural Net results matrix, Figure 33, shows some very interesting results. The EM clustered and ICA reduced dataset does not change its accuracy in any way regardless of any hyper-parameter tuning. This is consistent with the results observed in PCA NN training, Figure 20. However, the reduced and reduced and KM clustered datasets provide slightly better accuracy scores throughout hyper-parameter tuning. The variance is very small though, and thus concluding noise type from these results is difficult, but considering the similarities, it is most likely Gaussian noise.

F. Letter Dataset - ICA Reduced

Again, ICA reduction should provide the same results as PCA if the data has Gaussian noise.

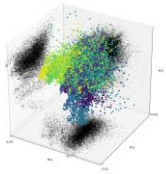
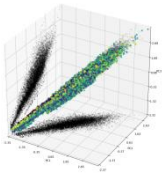
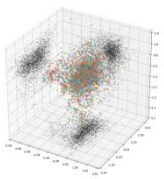




Figure 34: ICA Reduced Figure 35: ICA Inverse Figure 36: ICA @ k=7

Immediately it is noticeable how ICA reduction, Figure 34, transforms data very differently from PCA reduction, Figure 21. The ICA reduced data is densely compacted, and the formation of certain groups can be seen as blue and purple clusters are obviously segregating from the green and yellow ones. The inverse transformed data in Figure 35 looks similar to PCA version in Figure 22 with slender 3D

representation of the data with similarly shaped projections onto the YZ and XY planes.

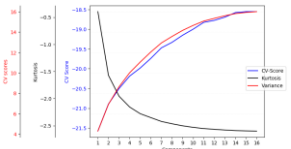


Figure 37: ICA Components Plot

Again, the ICA Components Plot shows 11 components as a good value retaining Variance while minimizing Kurtosis. For ICA reduction, minimizing kurtosis is the important metric and it is evident that beyond 11 components, kurtosis plateaus with diminishing returns as more components are evaluated.

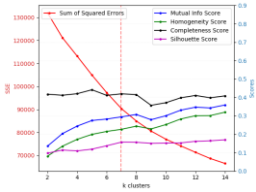


Figure 38: ICA KM Plot

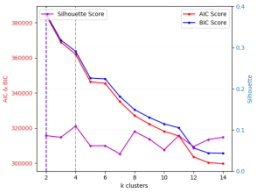


Figure 39: ICA EM Plot

The KM plot in Figure 38, shows relatively flat NMI, HM and CM values with increase in cluster values. The SSE elbow at 7 clusters is the only indicator of a "good" k pick as other indicators seem relatively uninteresting. This result closely resembles that of PCA KM evaluation where a similar behavior was observed along the various metrics and 6 clusters was the best SSE approximation. Figure 36 displays ICA transformed dataset at 7 clusters.

The ICA EM plot in Figure 39 although looks similar to the PCA EM plot in Figure 26 but the value ranges for AIC and BIC range 460k-360kfor PCA but only 380k-300k for ICA. The value range difference implies that ICA with EM clustering is better at determining "True" clusters than PCA with EM clustering.

G. Banknote Dataset - RP Reduced

RP uses Gaussian distribution random projection to project data to lower dimensions.

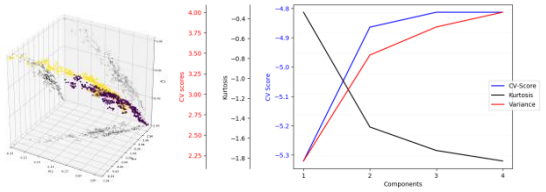


Figure 40: RP Reduced

Figure 41: RP Components Plot

The plot in Figure 40 is in stark contrast to the original dataset visualization. Narrow and closely spaced data groups can be seen with clear separation observed at the XY plane projection. Similar to previous algorithms, 3 components seem to be a sweet-spot for this dataset as viewed in Figure 41. Maintaining Variance and CV, 3 components retain most of the information contained in the original dataset.

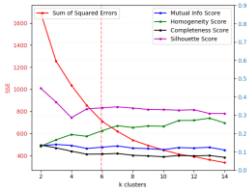


Figure 42: RP KM Plot

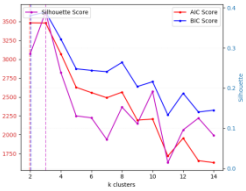


Figure 43: RP EM Plot

The KM Plot, Figure 42, displays very different behavior to that of the ICA plot in Figure 31 and PCA plot in Figure 18. This result shows a distinctive plateau of all metrics irrelevant of cluster number. The elbow point of clusters provides a best bet on the cluster number as other metrics remain flat lined. What this implies is that as cluster number increases, the number of points belonging to one class or another does not change dramatically.

The EM Plot, Figure 43, is very dramatic, with the highest Silhouette value at 3 clusters and both BIC and AIC falling sharply afterwards. This implies that while the data points are best modeled around their respective centroids at 3 clusters, the clusters are not well fitted for unknown True values and are better with more sparse clusters.

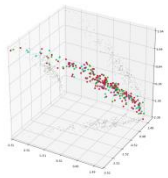
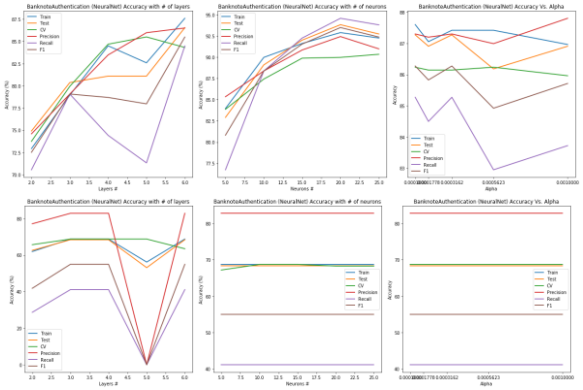


Figure 44: RP @ k=3

Figure 44 shows clustering of RP EM data at 3 clusters. The view which best displays these clusters appears somewhat obstructed, but it does seem as though there are distinct clusters if viewed against the XY plane.



CS7641 Machine Learning (Fall 2020)

Assignment 3 – Unsupervised Learning
Sergiy Palguyev (gtid: spalguyev3)

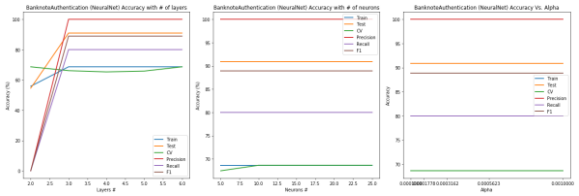


Figure 45: 3x3 Plot Cluster of Neural Network Tuning

Neural Net training with RP reduced and KM/EM projected data has some interesting characteristics when evaluating hyper-tuning the parameters. Figure 45 shows that for simply RP reduced data, accuracy in all metrics increases with the number of Neuron Layers and number of Neurons; the Alpha tuning has little effect on accuracy. NN training with RP reduced and KM clustered data shows neuron numbers play an interesting role, as there is a sharp drop off to 0 for the F1 metric at 5 neurons, implying there are either 0 True Positive or 0 True Negatives in the NN output. Finally, the EM clustered, RP reduced, data performs fairly well with only layer number affecting metric scores as other parameters have no effect and over precision staying at 100% and Test accuracy at above 90%.

H. Letter Dataset - RP Reduced

Again, RP uses Gaussian distribution random projection to project data to lower dimensions.

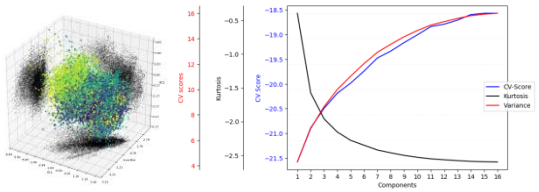


Figure 46: RP Reduced

Figure 47: RP Components Plot

The reduced image of the dataset, Figure 46, does not appear very segregated, and instead is a mix of data points. As evidenced by Figure 47, 11 components continue to do the best job in describing the dataset without loss of Varance.

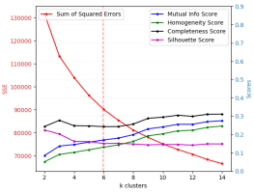


Figure 48: RP KM Plot

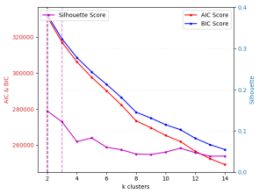


Figure 49: RP EM Plot

Very similar to ICA plot, Figure 38, the KM plot in Figure 48 find the best k value to be 6, with all metrics fairly flat indicating no gain in data belonging to proper classes. Furthermore, the EM plot in Figure 49 ranges the AIC and BIC scores 320k-260k below the ICA EM plot, Figure 39, with range of 380-300. Lower AIC/BIC scores imply closer correlation of predicted to unknown True values, thus a better model for fitting to the dataset.

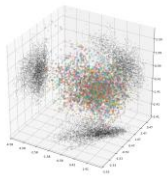


Figure 50: RP @ k=6

Figure 50 displays the transformed dataset with 6 clusters. It is evident, the clustering does not segregate the data well, at least form this view point.

I. Banknote Dataset - FA Reduced

The goal of FA is to reduce the number of factors of a dataset, rather than the number of features. Looking at the problem from a different perspective can assist in reducing multi-dimensional data as well.

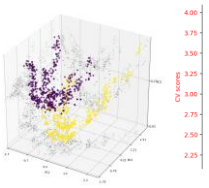


Figure 51: FA Reduced

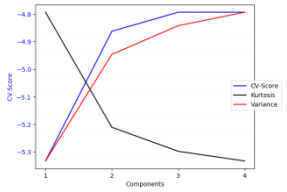


Figure 52: FA Component Plot

The FA reduced data plot, Figure 51, appears to spread the data points further out. This may help in grouping the certain data into clusters. Similar to before, Figure 52 shows that 3 components is sufficient in retaining maximum variance and kurtosis while lowering dimensionality.

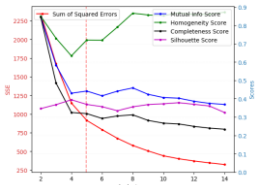


Figure 53: FA KM Plot

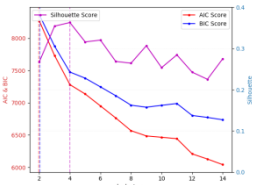


Figure 54: FA EM Plot

An interesting KM graph forms, Figure 53, showing a rapid drop in metrics as more than 2 clusters are evaluated, with the homogeneity score recovering as other plateau at approximately the calculated SSE elbow of 5 clusters. The recovering homogeneity score implies a form of overfitting" as the clusters begin to silo the data points, instead of proper grouping.

Similarly the EM graph, Figure 54, shows a noticeable peak at 4 clusters with the highest Silhouette score correlating with a sort of elbow seen in the AIC/BIC scores on the graph.

CS7641 Machine Learning (Fall 2020)

Assignment 3 – Unsupervised Learning
Sergiy Palguyev (gtid: spalguyev3)

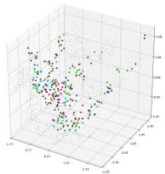


Figure 55: FA @ k=4

Figure 55 displays the 4 clusters evaluated by the FA projection of data.

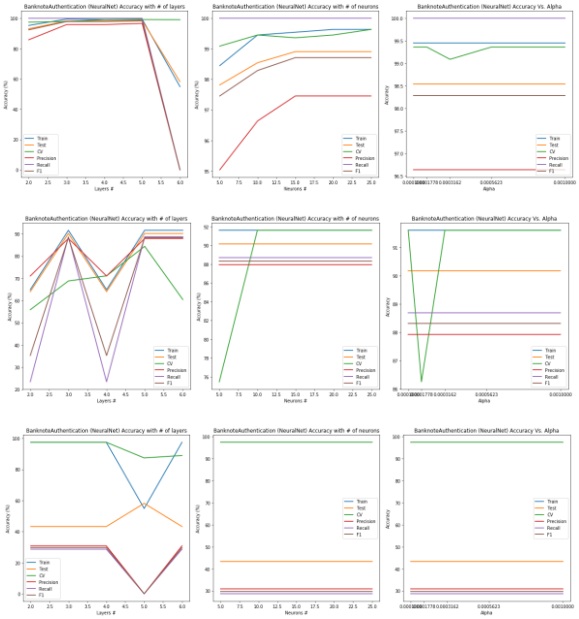


Figure 56: 3x3 Plot Cluster of Neural Network Tuning

The NN hyper-parameter tuning, in Figure 56, looks very interesting with FA reduction. Considering the focus on factors, rather than features, the NN learning on FA reduced data attains very high accuracy across all metrics. The alpha value appears to have no significant role in tuning the NN model, but the number of layers brings accuracy of all metrics nearly to 100%, and number of neurons also adds accuracy with an accuracy plateau above 15 neurons per layer.

J. Letter Dataset - FA Reduced

Again, FA attempts to reduce the number of factors of a dataset, rather than the number of features as the other algorithms.

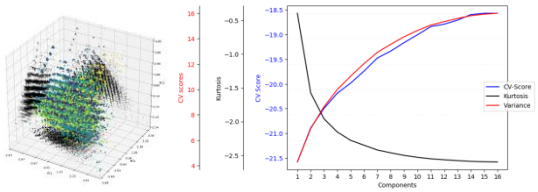


Figure 57: FA Reduced

Figure 58: FA Component Plot

The FA reduction graph, Figure 57, shows a very interesting pattern emerging. Layers in the projected data are emerging. This should enable clustering algorithms to do a much better job at fitting clusters to the data than seen in previous reduction attempts.

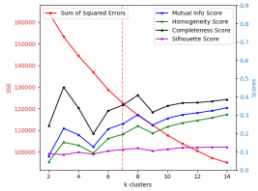


Figure 59: FA KM Plot

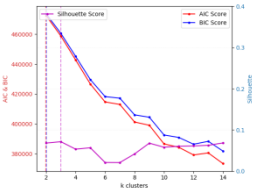


Figure 60: FA EM Plot

Figure 59, FA KM Plot shows an elbow at 7 clusters with SSE ranging much higher than other clustering methods. As an example, Figure 48, RP ranges SSE between 130k-70k while FA KM SSE ranges between 160k-100k. The Completeness score displaying a more dramatic swing implies point assignment to the same class varies greatly with cluster numbers, segregating too many individual points as k grows.

Figure 60, EM plot also shows very low EM values, implying the clusters of data are not well centered on their centroids. Similarly AIC and BIC scores are higher than other metrics, implying correlation to True values requires higher number of clusters.

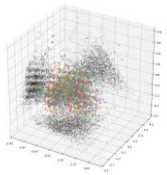


Figure 61: FA @ k=7

Figure 61 shows the clustering of FA projected data with 7 clusters.

VII. Summary and Conclusion

In summary, the best k metric values have been extracted and tabulated per each dataset below. Highlighted in green, the best values are representative of the best Dimensionality Reduction with clustering. The percentage accuracy is a measure of Dimensionality Reduction, without clustering, performance.

CS7641 Machine Learning (Fall 2020)

Assignment 3 – Unsupervised Learning
Sergiy Palguyev (gtid: spalguyev3)

Banknote					
	Original	PCA	ICA	RP	FA
NonClustered	98.50%	90%	86.90%	96.70%	98.10%
SSE	1234.75	1042.19	1029.63	641.65	914.7
NMI	0.24	0.27	0.27	0.07	0.44
HM	0.39	0.43	0.44	0.1	0.72
CM	0.17	0.19	0.2	0.05	0.32
SIL	0.3	0.3	0.29	0.34	0.36
AIC	7105.55	7450.16	7350.62	6412.07	7228.48
BIC	7550.58	7645.17	7595.64	6557.08	7476.49

Figure 62: Summary of Banknote Performance

Banknote dataset was best reduced by the FA Dimensionality Reduction algorithm. With KM clustering of the FA algorithm output, the highest NMI, HM and CM values were attained, implying most data points in a cluster belonged to only one class, points in the same class belonged to the same cluster agreement of label assignment to the dataset was highest with this algorithm combination. FA and its clustering also achieved the highest accuracy with Neural Network training and hyper-parameter tuning exercise.

Interestingly, the RP algorithm did best with EM clustering, achieving the highest Silhouette score while minimizing AIC and BIC. This implies with the RP dimensionality reduction, the EM clusters were most compact around their centroids while attaining the best modeling of "trueness.

Letter					
	Original	PCA	ICA	RP	FA
NonClustered	77.30%	70.70%	44.70%	62.50%	66.80%
SSE	157224	113492.1	90318.15	90154.63	122453.1
NMI	0.18	0.3	0.29	0.17	0.25
HM	0.14	0.24	0.23	0.13	0.2
CM	0.26	0.4	0.42	0.24	0.36
SIL	0.11	0.11	0.11	0.12	0.07
AIC	497345	411412.5	362000.1	317088.4	458830.9
BIC	502038	414999.3	3635014	318877.9	460620.4

Figure 63: Summary of Letter Performance

The Letter dataset performed best with simple PCA dimensionality reduction in both non-clustered output and KM clustering. Surprisingly, the best EM clustering was also achieved with the output of RP dimensionality reduction.

A multitude of questions and possible further avenues arose during this exercise, most of which is summarized in the following section.

VIII. Future Work

1. K-means s best suited for spherical clustering while Expectant Maximization relies on guessing to determine cluster centers. Are there other clustering algorithms better suited for multidimensional data of unordinary shape?
2. ICA failed to converge with high values of components with the Letter dataset, in certain scenarios. There is no indication that the datasets contain non-Gaussian noise, what is happening to ICA convergence?
3. Factor Analysis had some very interesting results. There were 3 algorithms looking at features and only one analyzing factors. Are there other algorithms that can be evaluated that look at data from a different perspective?
4. Neural Net training was a great exercise, but many plots were stagnant and flat. Expanding the range of hyper-parameters for tuning can uncover more interesting behavior. Why do the different hyper parameters sometime have no effect on accuracy?

IX. References

1. Banknote Authentication Dataset – Retrieved from https://archive.ics.uci.edu/ml/datasets/banknote+authentication
2. Letter Recognition Dataset – Retrieved from https://archive.ics.uci.edu/ml/datasets/Letter+Recognition
3. The source code and data can be found at : https://github.gatech.edu/spalguyev3/Fall2020CS7641-A3
4. https://medium.com/@cmukesh8688/k-means-clustering-in-machine-learning-252130c85e23
5. https://medium.com/@cmukesh8688/silhouette-analysis-in-k-means-clustering-cefa9a7ad111
6. https://www.methodology.psu.edu/resources/AIC-vs-BIC/