CS-7641 Machine Learning: Assignment 3

Unsupervised Learning and Dimensionality Reduction

Chenzi Wang
cwang493@gatech.edu

November 7, 2015

1. Unsupervised Learning through Cluster Analysis

The purpose of this assignment is to become familiar with unsupervised learning algorithms (in particular, those associated with cluster analysis) to understand their usefulness. The two clustering algorithms used are k-means clustering and Expectation Maximization (EM) clustering. Cluster analysis deals with grouping data into groups of similar sets of data. When analyzing high-dimensional data, these algorithms can become computationally demanding and less efficient. Dimensionality reduction algorithms work to reduce the dimensions of the data points, which can make the cluster analysis more computationally lean and efficient. The dimensionality reduction algorithms used in the current assignment include PCA, ICA, Randomized Projections, and naïve algorithms (one using solely the mean value of data and variance and the other using a bilinear downsample by factor of four). All coding was performed in MATLAB.

1.1 Clustering Algorithms

k-means clustering: The idea behind k-means clustering is to separate n number of data points into a user defined k number of clusters. There are k number of centers, one for each cluster. Where these clusters are placed changes results; in general, the way to get best results is to place the cluster centers as far away from each other as possible to cover the most result space. Data points are then associated with the center of its nearest cluster. If a data point is not closer to any one cluster center, it is not assigned for that iteration. The centers of the clusters are then repositioned to the center of the data points that were assigned to it. Once again, data points are assigned to the nearest cluster center. This is iterated until no data points are assigned to different cluster centers when the cluster centers are repositioned. This algorithm is fairly straightforward and computationally lean. However, there are some disadvantages. The number of clusters must be specified before starting the algorithm. If there are two overlapping regions of data but only one cluster center near it, those data points will be assigned to the one cluster instead of a more appropriate number of clusters. This algorithm does not work for categorical data, does not work well with noisy data and outliers, and does not work well with non-linear data.[1]

Expectation Maximization (EM) clustering: This form of clustering uses as iterative method for determining the best grouping method for data. The algorithm uses a data model which depends on unobserved variables, whether they be missing data or imaginary. The algorithm iterates through two steps: expectation and maximization. The first time into the expectation generates, random parameters are used as the estimated parameters. The expectation step generates a function for the expectation of the log-likelihood calculated using the estimated parameters. This function is then sent to the maximization step, which re-evaluates the estimated parameters to maximize the function of expected log-likelihood found in the previous expectation step. These parameters are then sent to the expectation step for the next iteration. The two steps iterate until the change in the log-likelihood function from one maximization step to its previous maximization step iteration is below a preset threshold (typically very small). A major difference between EM clustering and k-means clustering is EM clustering permits soft clusters of data points whereas k-means does not. Soft clusters permit data points to belong to more than one cluster with degree of membership to each cluster determined by its probability of belonging to that cluster. EM clustering also permits missing data. However, EM is somewhat susceptible to falling into local maxima when going through its iterations depending on the randomly generated first parameter estimates.[2]

1.2 Dimensionality reduction Algorithms

These are used to reduce the number of variables in clustering.

---

[1] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[2] https://engineering.purdue.edu/kak/Tutorials/ExpectationMaximization.pdf

Principle Component Analysis (PCA): PCA, in statistics, analyzes the linear relations between variables. It uses an orthogonal transformation to convert a set of variables that are possibly correlated into a set of linearly uncorrelated variables, termed principle components. It is performed on a covariance matrix, which is a matrix where each place, [i, j], in the matrix is the covariance between points, i and j, in a vector of random variables. For the covariance matrix, eigenvectors correspond to the principle components and eigenvalues, λ_i, correspond to the explained variance between these principle components. In this analysis, the largest eigenvalues correspond to the principle components that are most responsible for covariability among the data analyzed. The number of principle components is less than or equal to the number of variables, with the first principle component account for the most variance, followed by the second, third, and so on. A fault in PCA is that it is sensitive to outliers and its components depend on finding linear correlations between all input variables.[3,4,5] For this assignment, the algorithm code used is developed in MATLAB.

Independent Component Analysis (ICA): ICA functions by attempting to separate independent components from linearly mixed sources. In theory, this works only for data sources whose sources are combined linearly. Good results are based on two assumptions: the source signals are independent of each other and the data in each source signal has non-gaussian distributions. Before processing, the data goes through centering, whitening, and dimensionality reduction to reduce the complexity of the problem. During centering, the mean is subtracted to make a zero mean signal. Whitened means any correlations in the data are removed. This creates a covariance matrix where all diagonal values are equal. The ICA algorithm then maximizes the statistical independence of the components through maximization of non-gaussianity using kurtosis and negentropy.[6,7] For this assignment, the ICA algorithm FastICA is used in MATLAB.

Randomized Projections (RP): In a randomized projection algorithm, high dimensional data is projected onto a lower dimensional subspace using a random matrix. A data matrix, having d dimensions and n data points [d x n], is crossed with a random matrix with random [d' x d] matrix to produce the lower dimensional [d' x n] matrix. RP does not use an criteria for determining the importance/contribution of a component. It is computationally lean, such that it may surpass the efficiency of algorithms such as PCA and ICA. Projecting high dimensional data onto a low-dimensional random matrix has been shown to make the data more Gaussian and make the shape of odd-shaped clusters more spherical, showing promise for use in clustering. However, because it is projected onto a random matrix, each random projection may be quite different and produce very different clustering results.[8],[9]

2 Datasets

2.1 Cancer Diagnosis

As discussed in my previous assignment, I decided I needed to change to a different problem from the ones I chose in the first assignment. The problems I had chosen for the first assignment involved stock data prediction. Upon reevaluation, it is obvious this choice did not meet the requirements of this assignment. With the better understanding I now have, it is obvious to me that my previous problem was made up of continuous values and was a problem of regression rather than of discreet optimization necessary for this assignment. Rather than continue with a previous problem which would only produce results on which I cannot make meaningful analysis, I have decided to choose a different problem more aligned to the purpose of the assignment. This new problem and neural network was constructed using MATLAB.

---

[3] http://www.originlab.com/doc/Origin-Help/PCA-Algorithm
[4] http://www.mathworks.com/help/stats/pca.html
[5] http://www.mathworks.com/matlabcentral/fileexchange/26523-the-inface-toolbox-v2-0-for-illumination-invariant-face-recognition/content/INface_tool/auxilary/pca.m
[6] http://sccn.ucsd.edu/~arno/indexica.html
[7] http://research.ics.aalto.fi/ica/book/intro.pdf
[8] http://web.engr.oregonstate.edu/~xfern/rpm_icml03.pdf
[9] http://users.ics.aalto.fi/ella/publications/

The neural network analyzed here is one in which the algorithm attempts to recognize cancer cells from characteristics observed in microscopic images. This problem attempts to accurately determine whether a small number of cells removed from a tumor through a minimally invasive surgical process can be used to determine whether the tumor is malignant or benign. The data used contains 10 measured characteristics as well as the mean, standard deviation, and mean of the largest of the 10 characteristics (giving a total of 30 attributes) along with the diagnosis (malignant/benign). If algorithms using the principles of machine learning can be used effectively, a diagnosis can then be reached without the need of risky invasive surgery or the use of x-rays, which are not very accurate and expose a patient to radiation. If the principles of machine learning can be used effectively for such a problem, it is likely they can be used in other areas of medicine to achieve a diagnosis from measureable characteristics (even such things as body temperature and symptoms) without performing risky or costly investigative procedures.[10] Such a problem is primarily one of pattern recognition. The data was normalized to 0 such that they have a standard deviation of 1.

2.2 Handwritten Number Identification (ID)

In many fields of industry, camera recognition of text, numbers, and symbols is increasingly valuable due to the capabilities it can provide. Such a function can permit handwritten forms such as orders and shipments to be processed and recorded directly without being input manually. It can be used to translate different written languages, irrespective of different fonts or being handwritten, without being manually input by someone who may not be familiar with the characters used in that language (which can lead to errors in input data). Such a capability can also make it possible to handwrite information rather than type it into a computer, which is increasingly desirable with smaller devices where a standard keyboard is not feasible.

Handwritten characters, such as text and numbers, can vary significantly from one person to another as well as different areas of the world. Consequently, dependable recognition is very difficult and a singular reference sample of any character is not possible. The data used for this problem comes from the MNIST database, where the characters were collected from Census Bureau employees and high school students. The images are 28 x 28 pixels with 256 intensity level options for each pixel. The characters are centered in their images such that there are an equal number of darkened pixels across any intersecting line through the center. For processing, each image is turned into a 784 dimensional vector; this is achieved by concatenating each row of pixels.[11] These vectors are then normalized to 0 such that they have a standard deviation of 1. As described in the description of the cancer diagnosis problem, the problems that were chosen for the first assignment do not meet the requirements for this assignment. Consequently, this has been chosen as a replacement.

2.3 Measure of Similarity/Evaluation Method

For the handwritten number identification problem, determining whether the algorithms are successful is fairly straightforward since it is possible output images of the dimensionally reduced data. There is also the ability to output images which show the cluster centers and lower dimensional subspaces on which the data is projected. By doing so, the results can be better understood through seeing where they came from. The cancer diagnosis problem, on the other hand, does not possess this feature. Instead, numerical accuracy is solely used to test the results of the algorithms.

3 Results

3.1 Clustering Analysis – Both Problems

I chose a cluster size of k = 2 for the cancer diagnosis problem since there are two options for diagnosis: benign or malignant. I chose a cluster sizes of k = 10 for the handwritten data, one for each of the ten digits (0, …, 9). I tried a value of k = 5 to see what would happen; as expected, the reconstructed digits were very blurry. Because both clustering algorithms start with randomly generated parameters (the k-means algorithm chooses it original cluster centers randomly and EM clustering chooses random parameters for the first set of estimated parameters), the algorithms are run 5 times with the best results reported. For the k-means algorithm, the results are judged

[10] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/

[11] http://yann.lecun.com/exdb/mnist/

through the calculation of the mean distances of data to the cluster centers they are assigned to. The variance of these mean distances was then taken and the variance calculated. The lower the variance, in this case, the more equally sized the clusters are. Consequently, the result with the lowest variance is used. See Figure 1 for a graphical example; the more even the bars, the less variance. For the EM clustering algorithm, the result which produced parameter that gave the maximum log likelihood was used.

Figure 2 and Figure 3 show reconstructed digits using the k-means and EM algorithm respectively. Visibly, it appears that EM is superior, producing a more distinct, clearer picture than produced by k-means. For further evaluation, another property termed cleanness was identified and calculated. Calculating cleanness is a way of determining whether the clusters were made in such a way as each cluster representing a digit (which would be ideal since there are 10 clusters and 10 digits). To the this end, cleanness is calculated as 1/(# of digits assigned to that cluster), with the ideal value being 1. As can be seen from the graphical representations of this identity in Figure 4, neither the k-means nor EM algorithm have any clusters with values near the ideal. In both, in fact, the maximum cleanness any cluster reaches in < 0.45. This is a testament to the fact that different handwritten samples of the same digit are so different.

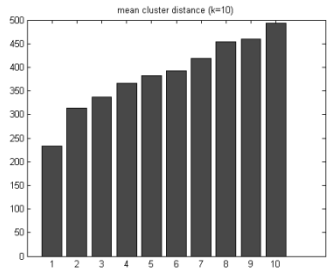
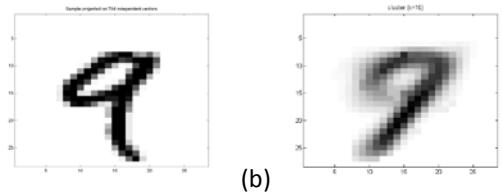
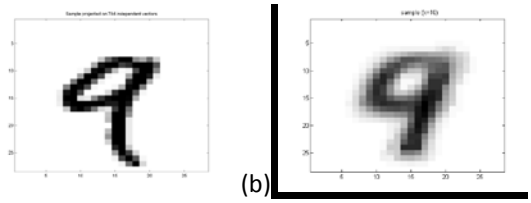


Figure 1: Sample graph of mean distances from data points to their assigned cluster.



(a) (b)

Figure 2: (a) Original handwritten digit and (b) digit reconstructed through k-means, k = 10



(a) (b)

Figure 3: (a) Original handwritten digit and (b) digit reconstructed through EM clustering, k = 10

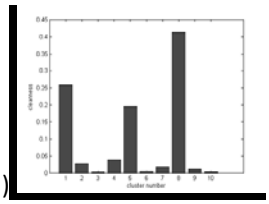
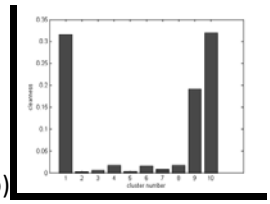
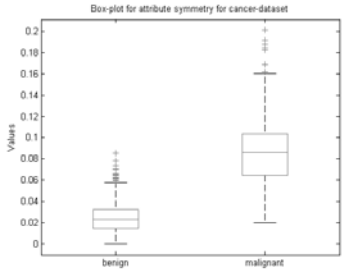
(a)  (b) 

Figure 4: Cleanness of the clusters in the (a) k-means and (b) EM algorithms, k = 10

For the cancer diagnosis problem, the code was designed to output descriptive parameters of the data in the clusters from each of the clustering algorithms including: the radius, compactness, and symmetry of the clusters. These attributes can be found in Table 1. It can be seen that the attributes for both the k-means and EM algorithm are identical for each cluster. This means that, though EM analysis permits soft clusters, both have the same hard clusters. The data encapsulated in the first cluster is around 99% made up of data with a malignant diagnosis whereas around 19% of data in the second cluster has a malignant diagnosis. This shows a relatively clean division of the data between the two clusters. Figure 5 shows graphically the symmetry of the cancer cells with benign and malignant diagnosis as well as their radius. The malignant cancer cells have statistically both a higher degree of symmetry and a larger radius than the benign cells. This may indicate that a model based on fewer cell characteristics may also give fairly accurate results.

Table 1: Clustering results for the cancer diagnosis problem

Algorithm	Cluster	Radius	Compact	Concavity	Symmetry
k-means	1	12.6	496	0.091	0.033
EM	1	12.6	496	0.091	0.033
k-means	2	19.4	1186	0.148	0.1
EM	2	19.4	1186	0.148	0.1

(a) 

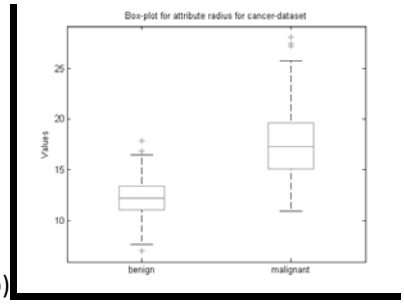
(b) 

Figure 5: (a) Box plots of the symmetry of the benign (on the left) and malignant (on the right) data sets and (b) box plots of their radius

3.2 Dimensional Reduction – Handwritten Number Identification

Because PCA determines a number of principal components whereas ICA and randomized projections use a subspace with user defined dimensions, PCA is run first and the number of principle components was used to setup the dimensions of the subspaces in the other two algorithms to better make a direct comparison. PCA was setup on a subspace such that it encapsulates (100, 99.9, 99, 97.5, 95, 90, 75, 33) percent of the total variance in the data. The eigenvector distribution for the handwritten number data is shown in Figure 6 along with the number of principle components for each of the chosen percentages of the total variance; these come out to be (784, 472, 329, 228, 154, 91, 37, 7). It can be seen from this figure that the number of principle components necessary drops rapidly with decreasing percent total variance. Also note that to obtain 100 percent of the total variance requires 784 principle components, which is the same number of dimensions in the data vectors;

whereas, only 472 principle components are needed to cover 99.9 percent of the total variance (that is around 60 percent less dimensions with a loss of only 0.1 percent of the total variance). This is likely connected to the method of converting these images into highly dimensional vectors being overly exhaustive.

Taking the data image of the number nine for comparison purposes, Figure 7 shows it after dimensional reduction through PCA with enough principle components to capture 33, 75, 95, and 99.9 percent of total variance. One does see that the image does become less blurry and more distinct with increased principle components, for instance 33 percent is notably lacking and not easily identified. However, at 95 percent little detail is lost from the original image and there is little notable difference between itself and the image at 99.9 percent. It is interesting to observe the images of each of the principle components, also called eigen-images, which can be found in Figure 8. These images show that the first principle component, which according to PCA should account for the largest amount of variance, has to do with a vertical stroke through the center of the image. This makes sense since this component of writing numbers may be said to be present in handwriting 5 out of 10 of the digits (1, 4, 6, 7, 9). The 10th component seems to capture some of the looping motions seen in such numbers as (2, 3, 5, 8) and is consequently important to identification. The 500th component, on the other hand, seems to be made up solely of noise in a circle around where most, if not all, of the numbers are located in the image files. This shows that this component can be gotten rid of likely without losing any identification power.

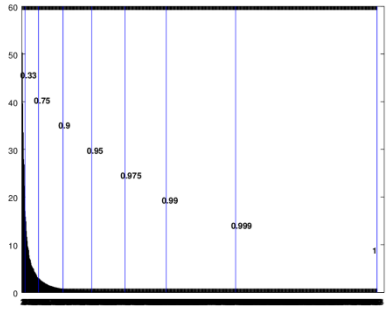


Figure 6: Eigenvector distribution of the covariance matrix of the handwritten number data along with the number of principle components (determined through PCA) for the specified percent total variances

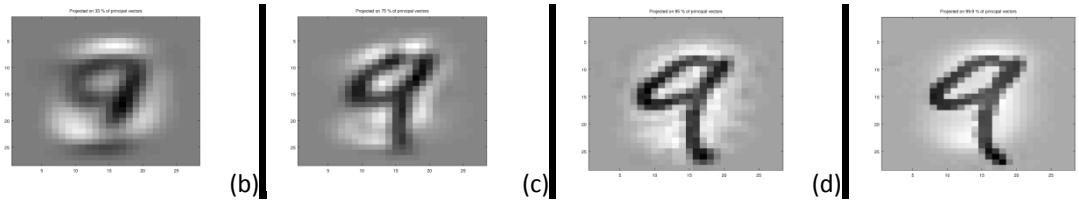


Figure 7: The images for the number nine through PCA dimensional reduction with enough principle components to capture (a) 33, (b) 75, (c) 95, and (d) 99.9 percent of the total variance

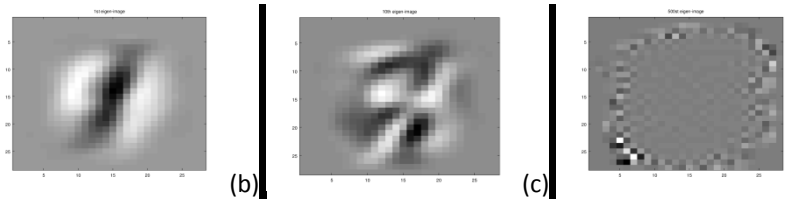


Figure 8: Images of the (a) 1st, (b) 10th, and (c) 500th principle components

Figure 9 shows the number nine after being dimensionally reduced through ICA. The subspaces used to produce the images shown in the figure have the same dimensions as the number of principle components used for the images above for PCA. These appear to be nearly if not the same as the results produced by PCA, as can be seen in Figure 7. Consequently, it cannot be said there is a notable difference between the two. However, it must be remembered that ICA does not assign a degree of importance to component signals whereas PCA does. Consequently, all component signals produced by ICA are treated as equally important. Samples of the component signals for the 4, 37, 154, and 472 dimensional subspaces can be seen in Figure 10. From this figure, it appears that with lower dimensional subspaces, component signals associate with a larger area of the images; whereas, as the dimensions of the subspace increase, a smaller portion of the image is associated with each component signal. This intuitively makes sense since, with fewer dimensions, components would need to associate themselves with larger amounts of data and the reverse is opposite for more dimensions. This idea is further illustrated using the kurtosis histograms seen in Figure 11, showing steep increases in kurtosis with increased dimensions. Because the resulting images when passed through ICA dimensional reduction cannot be said to be noticeably different than those produced through PCA, one cannot judge one to be superior to the other on the ground of accuracy. However, there is a notable difference in computation time: PCA = 9.6 seconds and ICA = 1031 seconds. Also, PCA has what may be called a significant advantage in that it does not need the user to specify the dimensions of any subspace but, instead, determines on its own the number of principle components to achieve the desired percentage of total variance.

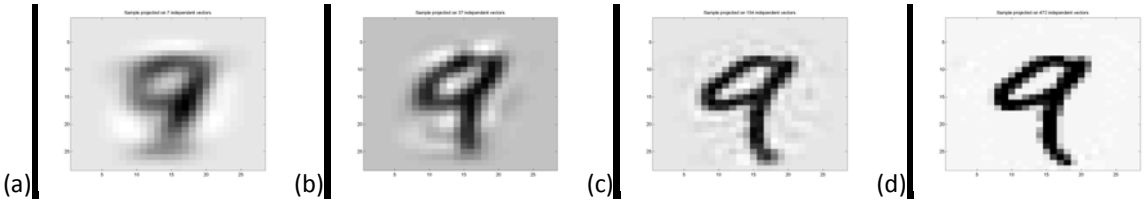


Figure 9: The images for the number nine through ICA dimensional reduction using a (a) 7, (b) 37, (c) 154, and (d) 472 dimensional subspaces

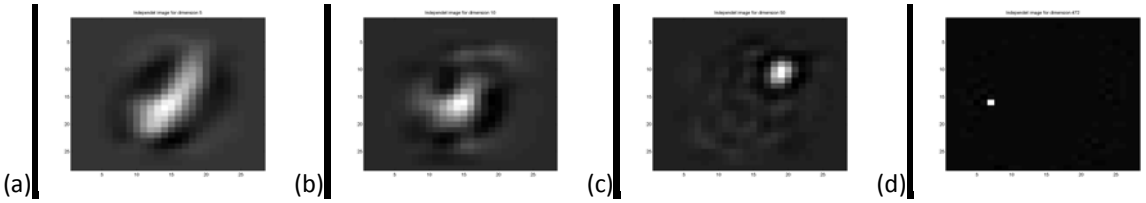


Figure 10: Sample component signals from ICA for (a) 5, (b) 10, (c) 50, and (d) 472 dimensional subspaces

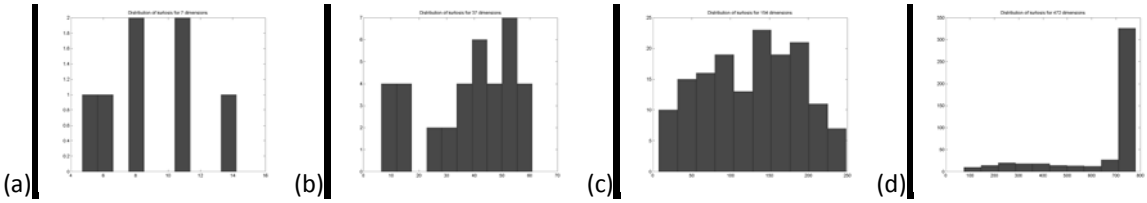


Figure 11: Histograms of kurtosis for component signals from ICA for (a) 7, (b) 37, (c) 154, and (d) 472 dimensional subspaces

As discussed earlier, randomized projection, though computationally lean, is projected onto a random matrix, thus each random projection may be quite different and produce very different clustering results. To attempt to find

the best projection, the algorithm I used runs randomized projection a predetermined number of times, calculates among the results from each run which produced the best result based on reconstruction error of data, then outputs the best result it finds out of its runs. I chose to use 1000 runs and yet they did not produce results near the functionality seen in PCA and ICA. This is obvious when judging the reproduced number in Figure 12 after being dimensionally reduced in this manner. The computational time to run the algorithm is also extremely costly, taking hours to complete. It is clear, however, that increasing the dimensionality of the subspace does improve the reconstruction, but this is not unexpected since a subspace with the same dimensions of the data should output a reconstruction exactly like the original for any of these dimensional reduction algorithms.

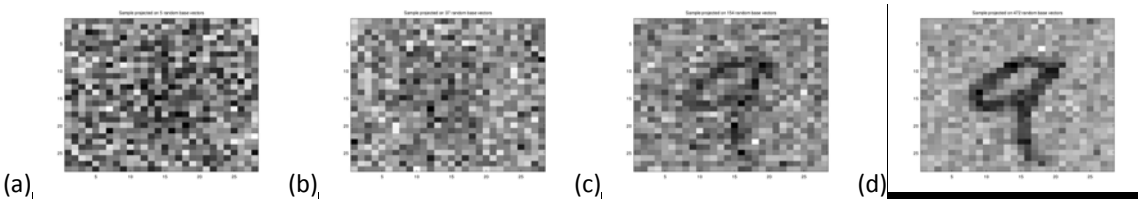


Figure 12: Images for the number nine through randomized projection dimensional reduction using (a) 5, (b) 37, (c) 154, and (d) 472 dimensional subspaces

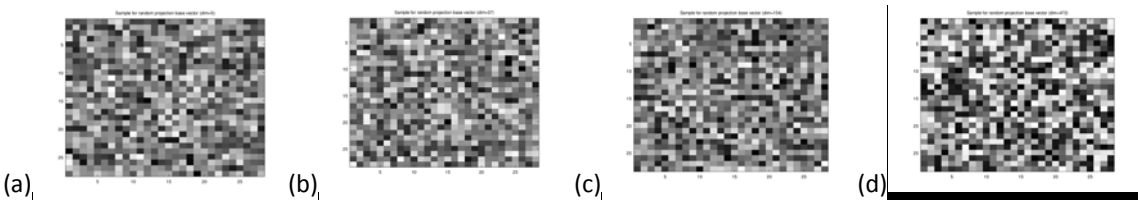


Figure 13: Samples of randomized base vectors for (a) 5, (b) 37, (c) 154, and (d) 472 dimensional subspaces

3.3 Dimensional Reduction – Cancer Diagnosis

Running PCA on the cancer diagnosis problem set produced the eigenvector distribution seen in Figure 14. According to this figure, to achieve the desired (100, 99.9, 99, 97.5, 95, 90, 75, 33) percent total variance, (30, 4, 3, 2, 2, 2, 2, 2) principle components are required. There is a large gap between 30 and 4 principle components which seems to mean that only a few of the 30 components contain pertinent information for the purpose of identification. It is seen that two components define the data very well. This supports the evidence seen previously in the clustering section that the two attributes, radius and symmetry of the cancer cells, have a strong influence on differentiating between benign and malignant cells.

Due to the results of PCA, a 2 dimensional subspace was used for ICA for this data. The results showed that the two signals produced had a high kurtosis, see Figure 15. This shows that each signal is highly independent. A comprehensive comparison between the behavior of these two signals and the two clusters formed previously would not make sense since those produced by ICA are not graded such that one signal is not chosen to carry more variance than the other, unlike components found through PCA. The results from randomized projections of a 2 dimensional subspace once again performs poorly. A greater dimensional subspace is needed; but as a consequence, it can already be said that randomized projection performs poorly in comparison to PCA and ICA.

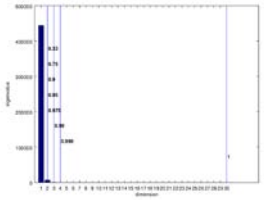


Figure 14: Eigenvector distribution of the covariance matrix of cancer diagnosis data along with the number of principle components (determined through PCA) for the specified percent total variances. Note there are lines representing

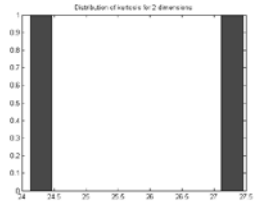


Figure 15: Histogram of kurtosis for component signals from ICA for a 2 dimensional subspace

3.4 Downsample and Attribute Reduction Dimensional Reduction

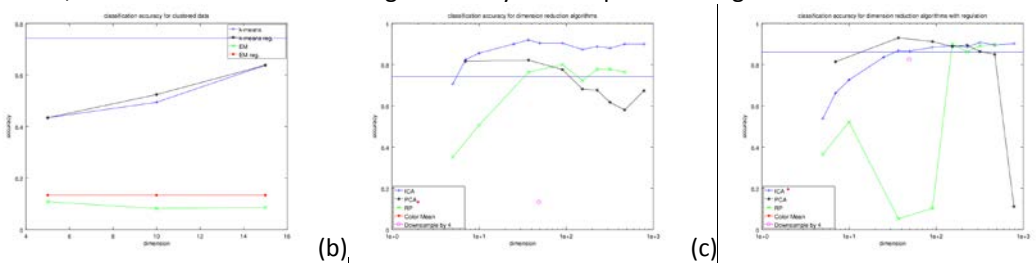
I chose to implement two naïve dimensional reduction methods. The first, aimed directly for use on the handwritten number problem, reduces the data solely to the mean and variance of the pixel intensity in the data images (there are 256 intensity levels available to each pixel). This is crude, but would provide a large payoff since this would allow for identification using only two attributes (variables). The second naïve algorithm consists of a downsample by a factor of 4. This also was thought up of more specifically for the handwritten number problem since this sort of downsample may keep the overall shape of the numbers even with the loss of detail since it would likely be the overall shape of the number that is used for identification. As has been seen, the cancer diagnosis data set can be reduced to two attributes, radius and symmetry, and produce functional results.

3.5 Neural Network – Handwritten Number Identification

In the prior assignment, it was shown that a neural network of 2 layers with 50 nodes each worked best for classification in the handwritten number identification problem. The neural network was run with and without a regularization term that is meant to deter overweighting of nodes. Because of the randomly generated parameters as discussed in the clustering section, the clustering algorithms were run 3 times each and the best result was used. In handling the training and test sets, the dimension of the subplots for each of the data reduction algorithms were kept the same. The classification accuracy for the clustering and dimensional reduction algorithms can be found graphically in Figure 16. Their computational times can be found graphically in Figure 17. For comparison, the accuracy and computational time to run the neural work as before, without clustering or dimensional reduction, as a horizontal blue line across each chart.

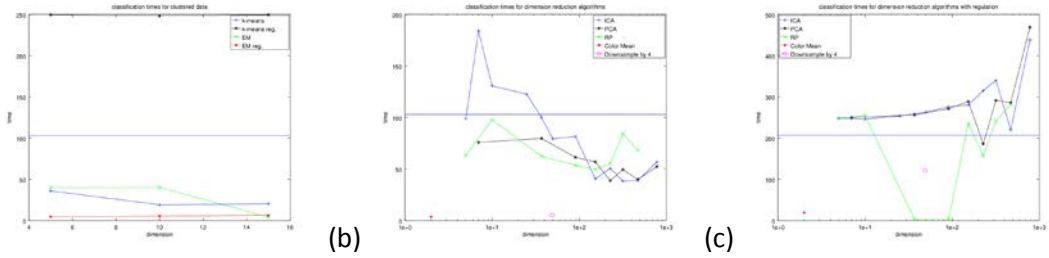
As far as accuracy goes, it can be seen that only the dimensional reduction algorithms surpass the accuracy achieved without the implementation of the present algorithms. ICA produces the most accurate results for the NN without regulation; but with regulation, it is second best to PCA up until the higher dimensions where PCA is seen to drop in accuracy. For both with and without regulation, the accuracy of PCA begins by climbing with increased dimensions up to a high dimensional point upon which time it begins to drop. This drop in accuracy at higher dimensions may be due to added noise from the higher order principal components which may amplify outliers in test data. The accuracy of ICA also improves with added dimensions but then levels out. This is because the optimum number is being reached. The components in ICA are independent, non-linear of one another and so more dimensions leads to empty components that do not harm or benefit the accuracy. The randomized project accuracy is erratic. It provides the lowest accuracy of the dimensional reduction algorithms except at higher dimensions where it surpasses PCA. As expected, it generally increases in accuracy with increased dimensions in its subspace. For the neural network without regulation, the computational times for the dimensional reduction algorithms were all less than the computational time for the neural network without the present algorithms except for PCA at lower dimensions; this was not the case with regulation added. The two poorest performers for accuracy are the cluster algorithms, with the EM clustering algorithm showing the worst accuracy (the accuracy of EM also doesn't improve with added clusters). This shows that, in this case, the soft clusters created by EM did not give it an advantage over k-means with its hard clusters. K-means, however, does show improvement with increased clusters. The low accuracy of the cluster algorithms may be attributed to the non-linear nature of the data. EM is shown to also be the worst performer per computational time between the cluster algorithms. The cluster algorithms are, however, both faster than most points of all the algorithms, with the exception of a few points from randomized projection, and faster than the classification without the use of these additional

algorithms. It may sound strange that randomized projection showed some short computation times since it was mentioned earlier that it was computationally demanding. This is because, in this case, the algorithm was only run once. It can also be noted that the times where it was the fastest, it produced its lowest accuracy. The dimensional reduction based on the mean and variance of image pixel intensities performs poorly in comparison to the classification with the other dimensional reduction algorithms and the classification without the present algorithms. On the other hand, it does still surpass the accuracy seen for the EM clustering algorithm. The downsample algorithm performed similarly to the intensity algorithm in the case without regulation; but with regulation, it performs almost as well as accurate as the neural network without present algorithms. I am not sure why this is, but I am interested in looking into it beyond the present assignment.



(a) (b) (c)

Figure 16: Accuracy as a function of number of clusters for (a) clustering and as a function of number of dimensions for (b) dimensional reduction and (c) dimensional reduction with regulation



(a) (b) (c)

Figure 17: computation time as a function of number of clusters for (a) clustering and as a function of number of dimensions for (b) dimensional reduction and (c) dimensional reduction with regulation

4 Conclusions

PCA and ICA are shown to have the best performance of the clustering and dimensional reduction algorithms in their ability to reconstruct data. They were shown to produce similar if not identical visual results for the handwritten number problem and both showed that diagnosis for the cancer cells in the cancer diagnosis problem could be performed functionally with 2 attributes. PCA and ICA had the highest accuracies of all the algorithms in the neural network problem. In fact, their accuracies far surpass those attained through the cluster algorithms. PCA and ICA do have the longest computational times of the algorithms, much longer than the cluster algorithms, but the high improvement in accuracy over clustering and less tendency for erratic behavior (which was prevalent in random projection) warrant their use. PCA may have an improvement over ICA in that it computes its number of principal components rather than depending on user defined dimension of a subspace, as is the case with ICA and random projection). Instead, the desired variance to be captured is specified. The clustering algorithms did not perform well in the neural network because of the non-linear nature of the data set.