Unsupervised Learning and Dimensionality Reduction

Dataset selection and preprocessing. Both datasets were explored in Assignment 1.

The first dataset is from a lending club. This is a binary classification problem to predict whether a loan will be paid off in full or the loan will be charged off, using the numeric features (such as loan amount and annual income) and category features (such as credit grade, term of loan, current home ownership and years of employment). The numeric attributes are scaled to range [0, 1] and the categorical attributes are transformed to binary attributes (one hot encoding) by the function get_dummies in Pandas. Since there are less risky loans than safe loans, I calculated the ratio of the sizes of risky loans and safe loans and use that percentage to under-sample the safe loans to get a balanced dataset. The preprocessed dataset contains 3896 samples, 26 attributes (two numeric attributes and 24 one-hot-encoded attributes) and two classes. This dataset is interesting because it is a very practical problem, but difficult to find a classifier to get a high accuracy prediction.

The second dataset is Pageblocks classification. This dataset contains blocks of the page layout of a document that has been detected by a segmentation process. The task is to determine the type of page block: Text (1), Horizontal line (2), Graphic (3), Vertical line (4) and Picture (5). The original dataset has 5472 samples in five classes: 4913 samples in Class 1, 329 in Class 2, 28 in Class 3, 87 in Class 4 and 115 in Class 5. To prevent the Class 1 dominating the classification score calculation, Class 1 was under-sampled to 10% of the original size (491 samples). The balanced dataset has 1050 samples, 10 numeric features and 5 classes. The values of ten numeric attributes of the training set were scaled to zero mean and unit variance by the StandardScaler (in the sklearn.preprocessing). This is an interesting dataset, because even after under-sampling it still has five classes with quite different number of samples, the classification accuracy of which may be dominated by classes with large number of samples. I am curious to see whether these five classes can be classified correctly.

Part 1. Clustering on the data sets.

Algorithm implementation and distance measure.

For both datasets, the k-means clustering is performed using the sklearn.cluster.KMeans module. On more detail, the k-means problem is solved by Lloyd's algorithm, using 'k-means++' to initialize the centroids, and 10 times of the k-means algorithm (n_init = 10) will be run with different centroid seeds. The Euclidean distance is used as the similarity measure, because it is common, intuitive and sklearn only support this for k-means. The expectation maximization (EM) clustering is performed using the GaussianMixture from the sklearn.mixture module, with covariance_type = 'full' (each component has its own general covariance matrix) and using k-means method to initialize weights (init_params='kmeans').

How to choose k?

For both datasets, the Elbow method and the average silhouette method are used to choose k, and the adjusted mutual information (AMI) score between real label and the predicted label by clustering was used to validate the selection. For the Elbow method, the sum of squared error (SSE) on each k is computed for K-means clustering, but the log probability (LL) on each k is computed for the EM clustering. To better compare the trend of the results of k-means and EM algorithms, a negative sign is applied on the SSE of k-means to display the graph. The silhouette score is calculated by the sklearn.metrics silhouette_score function. A small silhouette score indicates the sample is on or very close to the decision boundary between two neighboring clusters.

For the safe/risky loan classification problem, no k value indicates a "knee" position in the Elbow plots of both k-means and EM, as shown in Figure 1-1A and -1B. With the average silhouette method, both k-means and EM have largest average silhouette value when k = 15, shown in Figure 1-1C. However, K-means clusters have highest AMI score when k =2 and the score decreases monotonically, as shown in Figure 1-1D. We can also see the EM clusters have highest AMI score when k =3, and AMI score decreases as the k increases when k > 3. The high AMI scores when k = 2 or 3 make sense, because this dataset has two classes. However, since when performing clustering we should not provide class labels, we cannot choose k value using AMI score between true labels and clustering labels. Therefore, k = 15 is used for later analysis.

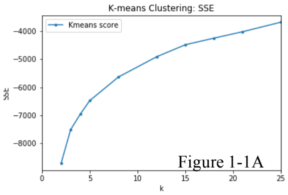


Figure 1-1A

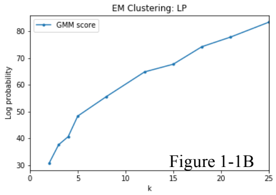


Figure 1-1B

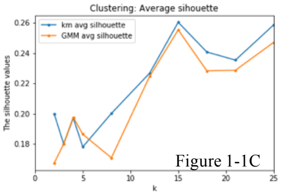


Figure 1-1C

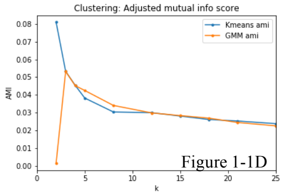


Figure 1-1D

For the Pageblocks classification, similar as the safe/risky loans, the Elbow curve does not have an obvious "knee" position (see Figure 1-2A and -2B). The EM clustering show the highest average silhouette value when k = 6 and k-means show highest silhouette when k =12, as shown in Figure 1-2C. Since this dataset has 5 classes, the result of average silhouette analysis of EM clustering lines up better with the real number of classes. The AMI score when k = 6 is also highest for EM clustering and K-means clustering (see Figure 1-2D).

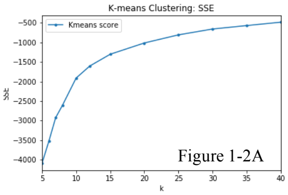


Figure 1-2A

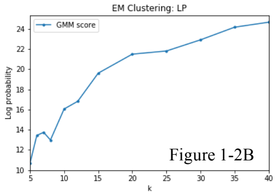


Figure 1-2B

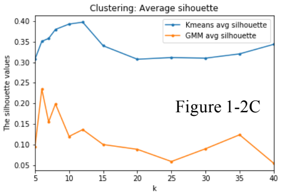


Figure 1-2C

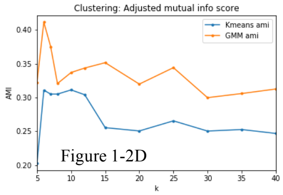


Figure 1-2D

The kind of clusters I got.

Since both my datasets have high dimensions, displaying the clustering results in every pair of feature dimensions is impossible. Instead, I display the clustering result as scatter points in two dimensions by using the t-distributed Stochastic Neighbor Embedding (t-SNE) technique. The sklearn.manifold.TSNE is used to plot.

For the Loans dataset, as discussed above, k = 15 is applied for the k-means clustering and the EM clustering. As shown in Figure 1-3A and 3B, even though the same cluster number is used for k-means and EM clustering, the clusters generated by these two algorithms are not exactly same. This is because the EM is a soft clustering based on probability distribution over clusters which guarantees non-decreasing likelihood, but k-means is only based on Euclidian distances (in my experiments) which guarantees error decreases. In both cases, on the two t-SNE dimensions, the majority of the scatter points have similar pairwise distances, which might be the reason why both clustering algorithms cannot find a good way to cluster the samples into two clusters for this binary classification problem.

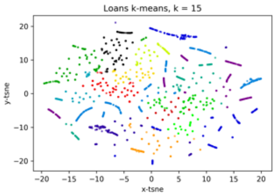


Figure 1-3A

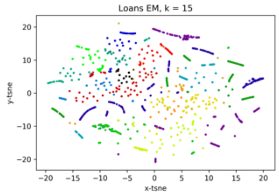


Figure 1-3B

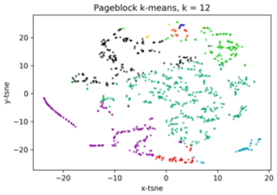


Figure 1-3C

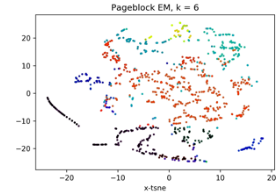


Figure 1-3D

For the Pageblocks dataset, as discussed above, k=12 and k = 6 is applied for k-means and EM clustering, respectively. On the two t-SNE dimensions, the sample points are naturally better clustered than the data of Loans problem. Therefore, it is easier to cluster this dataset with higher accuracy and less cluster numbers. Noticeably, both clustering results have some very small cluster containing very few sample, but also have very big cluster takes up more than 30% of the total samples. This is reasonable, because the five classes in the Pageblocks dataset have quite different size: 491 samples in Class 1, 329 samples in Class 2, 28 samples in Class 3, 87 samples in Class 4 and 115 samples in Class 5.

Discussion.

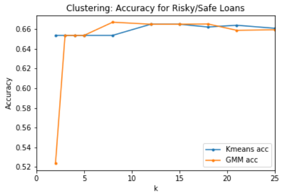


Figure 1-4A

The clustering results are converted to classification labels and then compared with the true labels of each dataset. For the Loans dataset, the k-means clustering has classification accuracy between 0.65 to 0.66 with the cluster number k ranges from 2 to 25 (Figure 1-4A). However, when k =2 the EM clustering results in a very low accuracy at 0.52, barely beats random guess, but when k ≥ 3 the accuracy varies between 0.65 and 0.67. The reason might be: The EM clustering is a soft clustering method which computes possibility of each sample be assigned to every cluster. When only two clusters exist, the huge amount overlapping samples will have similar possibility being assigned to each class; therefore, the EM clustering result is just like a random guess which has classification accuracy 0.5. With the optimal k value, the classification accuracy is not too much lower than those obtained by the supervised learning algorithms (0.66 to 0.70) in Assignment 1. The low accuracy is partially due to the dataset itself. As discussed in Assignment 1, this dataset is noisy, so no matter what algorithms are used they are learning on a lot of noise.

For the Pageblock dataset, the EM clustering has significantly higher classification accuracy than the k-means clustering for different choice of k, only except for k =8 (Figure 1-4B). Even though the average silhouette method and the AMI score indicate that k=12 and k =6 is the best choice for k-means and EM, respectively, the classification accuracy is very low when k ≤ 12. Even with k > 15, the best accuracy the EM clustering can reach is about 0.86 much worse than the classification accuracy achieved by the supervised learning (0.93 to 0.96) in Assignment 1. There is some space to improve the clustering accuracy.

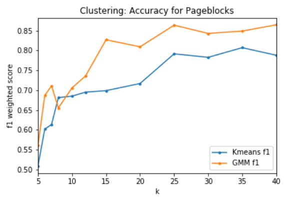


Figure 1-4B

The relatively low classification accuracy of both datasets compared to supervised learning classification may be a result of the curse of dimensionality. To improve the performance, instead of using Euclidean distance as the similarity measure, alternative metrics could be used to evaluate the similarity between samples, such as the Manhattan distance and cosine similarity. However, the k-means clustering in sklearn only implemented with Euclidean distance as the similarity metrics. The GaussianMixture of sklearn also does not have alternatives to avoid the curse of dimensionality. Even though the distance metric cannot be changed within sklearn package, dimensionality reduction techniques can be applied to transform the original data and reduce the bad impact from curse of dimensionality.

Part 2. Dimensionality reduction on the two datasets.

Algorithms and Experiments

For the principal component analysis (PCA), the sklearn.decomposition.PCA module was applied to the two data sets. The "n_components" parameter was set to equal to the number of features of the data.

The sklearn.decomposition.FastICA module is applied to the two datasets for the independent component analysis (ICA). The kurtosis is computed by the pandas.DataFrame.kurt function, which returns unbiased kurtosis over requested axis using Fisher's definition of kurtosis (kurtosis of Gaussian == 0.0).

The randomized projection (RP) analysis is done by using the SparseRandomProjection module of the sklearn.random_projection. The main theoretical support behind the efficiency of random projection is the Johnson-Lindenstrauss lemma, which states that "a small set of points in a high-dimension space can be embedded into a space of much lower dimension in such a way that distances between the points are nearly preserved". To evaluate how well the dimensions and distribution of random projections matrices are controlled to preserve the pairwise distances between samples of the dataset, the pairwise distance correctness rate and the reconstruction error rate were examined for different number of projected dimensions. The reconstruction error is the mean squared error between the original data and the reconstructed data (projecting the projected data back to the original dimensions). Since this is a randomized projection method, the mean and standard deviation of the pairwise distance correctness rate and the reconstruction error rate are obtained by rerunning the RP analysis for 10 times.

The Random Forests method is selected as the feature section algorithm. The RandomForestClassifier in sklearn.ensemble is used to analyze the importance of each feature.

To validate whether the dimension numbers determined by Elbow method are optimal, I run neural network learners on the datasets transformed by each dimension reduction method. To obtain reliable accuracy, k-fold cross-validation is performed for neural network classification on the transformed data. The sklearn.neural_network.MLPClassifier with hyperparameters solver='lbfgs', activation='identity', alpha=0.1, hidden_layer_sizes=(50,) was used for the validation of every dimension reduction approach.

Safe/Risky Loans.

Principal component analysis (PCA).

PCA transforms the data to linear planes that maximizes variance and thus generates a matrix in which the eigenvalues (covariance) are maximized. The distribution of the eigenvalues for Loans dataset is plotted in Figure 2-1A. The eigenvalue decrease quickly from dim=1 to 7, after 7 the eigenvalue slowly decrease. When dim=7, the cumulative eigenvalue from 1 to 7 is only 68.8% out of the total cumulative eigenvalues. At dim=15, the cumulative eigenvalue from 1 to 15 is 92.5% out of the total cumulative eigenvalues; and at dim = 22, the cumulative eigenvalue is 99.4%. The optimal choice of the projection dimension for the Loans dataset should be somewhere between 15 and 22.

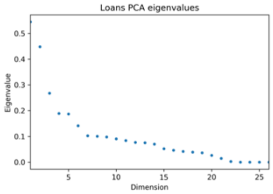


Figure 2-1A

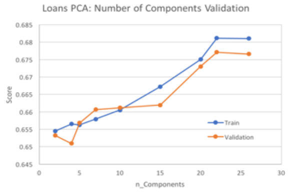


Figure 2-1B

The train and cross-validation accuracy obtained by different principal component input to neural network learner are plotted in Figure 2-1B. With dimension number 20, 22 and 25, the accuracy of training and cross-validation are similar, much higher than the accuracies using dimensions less than 20.

Independent component analysis (ICA).

ICA was originally developed to deal with the cocktail-party problem, which is to solve the unobserved sources of the observed variables. The assumptions to solve this type of problem are: unobserved sources are statistically independent and the observed variables are linear combinations of those independent sources. The key to a successful ICA estimation is that the independent components must be nongaussian. In another word, the nongaussianity determines the quality of ICA estimation. For this feature extraction problem, the nongaussianity of ICA is measured by kurtosis. For the Loans dataset, the maximum kurtosis is 13.77 when 22 decomposed dimensions (n_components = 22), as shown in Figure 2-2A. With dimension number 22, 23 and 25, the accuracy of training and cross-validation are similar, much higher than the accuracies using components less than 20 (Figure 2-2B). Therefore, 22 components will be used for later analysis in Part 3.

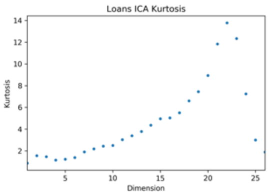


Figure 2-2A

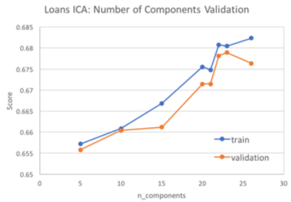


Figure 2-2B

Randomized projection (RP).

The average pairwise distance correctness rate and the average reconstruction error of 10-times experiments are displayed as in Figure 2-3A and Figure 2-3B, with the standard deviation of them as the error bars (The standard deviation = positive error bar length = the negative error bar length). The standard deviation (stddev) of pairwise distance correctness becomes small when more than 15 randomly projected dimensions are applied (Figure 2-3A). When using 2 projected dimensions, the stddev of pairwise distance correctness is about 25.8% of the average correctness; while using 26 projected dimensions, the stddev of pairwise distance correctness is about 3.3% of the average correctness. As shown in Figure 2-3B, the variations of RP analysis are not very big with different number of projected dimensions. When project to 21 dimensions, the average pairwise distance correctness rate is almost as high as projected to 26 dimensions and the average reconstruction error is only about 0.03. High classification accuracy is also achieved using the transformed data from 21 projected dimensions in neural network learner.

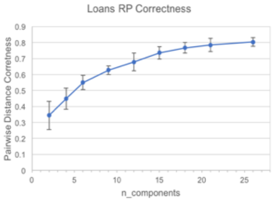


Figure 2-3A

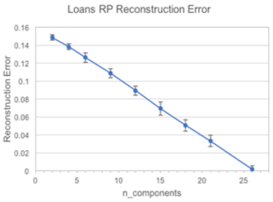


Figure 2-3B

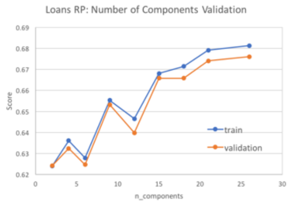


Figure 2-3C

Feature selection by random forest (RF).

The importance score of each feature of the Loans dataset is shown in Figure 2-4A. There are four features of the Loans dataset are significantly more important than others: the loan amount (feature 0), annual income of the applicant (feature 1), loan term 36 months (feature 9), and loan term 60 months (feature 10). This result is reasonable, because from common sense these four features are the crucial factors to for a bank to judge whether a loan is safe. The validation results in Figure 2-4B show that after the five most important features are used in neural network learner, the cross-validation accuracy reaches a plateau. Therefore, the 5 most important features are used for the clustering in Part III.

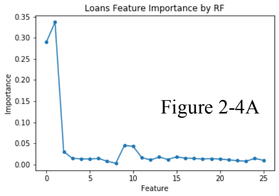


Figure 2-4A

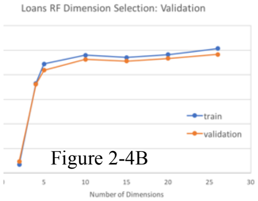


Figure 2-4B

Pageblocks

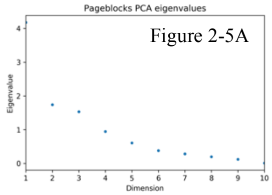


Figure 2-5A

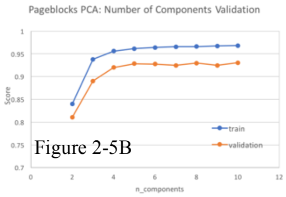


Figure 2-5B

PCA. For the Pageblocks dataset (Figure 2-5A), a "knee" position shows when dimension = 6, where the cumulative eigenvalue is 93.8% out of the total cumulative eigenvalue. The optimal projection dimension should be in the range of 6 to 10. The training and cross-validation accuracy obtained by different principal dimensions evaluated by neural network learner are plotted in Figure 2-5B. The best training and cross-validation score are achieved with 4 or more principal components. Therefore, 6 principal components are sufficient for later analysis.

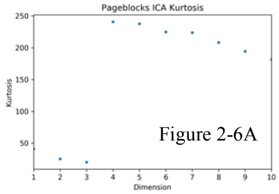


Figure 2-6A

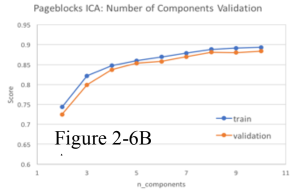


Figure 2-6B

ICA. For the Pageblocks dataset, the kurtosis was very small when n_components ≤ 3, but immediately jumps to the highest when n_components = 4 (Figure 2-6A). When n_components > 4, the kurtosis slowly decreases but still at a relative high level. The best training and cross-validation score of neural network classifier are achieved with 8, 9 and 10 components (Figure 2-6B). But the accuracy score achieved using 4 independent components, is not much worse than the scores achieved by 8 decomposed components. Therefore, 4 independent components will be used for the clustering in Part 3.

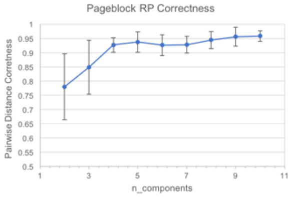


Figure 2-7A

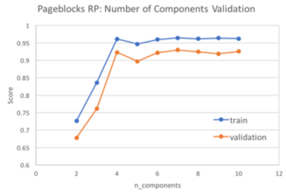


Figure 2-7B

RP. As discussed for the Loans dataset, the standard deviation of pairwise distance correctness rate from 10-times experiments are displayed as error bars on the pairwise distance correctness curve and the reconstruction error curve. The standard deviation (stddev) of pairwise distance correctness becomes small when 4 or more randomly projected dimensions are applied. When using 2 and 3 projected dimensions, the stddev of pairwise distance correctness is about 14.8% of the average correctness; when using 4 projected dimensions, the stddev of pairwise distance correctness is about 2.8% of the average correctness. Using data transformed by 4 or more projected dimensions in the neural network classifier results in high accuracy scores (> 0.90). Therefore, transformation with 4 randomized projected dimensions is a reasonable choice for clustering in Part III.

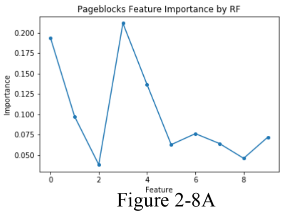


Figure 2-8A

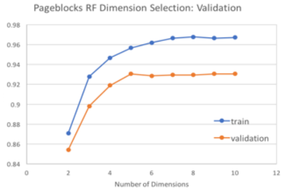


Figure 2-8B

RF. The random forest results indicate that four features are significantly more important than others, they are: the height of the block (feature 0), the length of the block (feature 1), the eccentricity (= length / height) of the block and the percentage of black pixels within the block (blackpix / area). The validation results show that after 5 most important features are used in neural network learner, the cross-validation accuracy no longer increases. Therefore, the 5 most important features are used for the clustering in Part III.

Part 3. Reproduce clustering on the two datasets after dimensionality reduction.

Safe/risky loans.

Clustering on PCA transformed data. As discussed in Part 2, the original data of Loans is transformed to 20 principal components for the clustering. With the transformed data, the k-means clustering SSE score (Figure 3-1A) and EM log probability (Figure 3-1B) still do not have obvious elbow position. However, the AMI scores

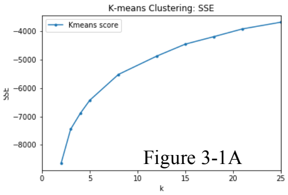


Figure 3-1A

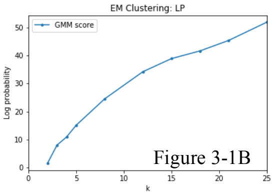


Figure 3-1B

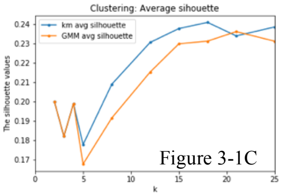


Figure 3-1C

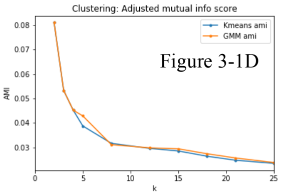


Figure 3-1D

of EM clustering becomes almost identical to that of K-means clustering (Figure 3-1D). Unlike the clustering with original data where the EM clustering has very low AMI score when k =2 (see Figure 1-1D), with PCA the AMI score when k =2 is higher than that of any other k. After PCA the silhouette score reaches maximum with k =18 for k-means and k=21 for EM (Figure 3-1C), but without PCA both k-means and EM have highest silhouette score when k=15 (Figure 1-1C). With k=18, the the scatter plot of clusters projected on the t-SNE dimensions are displayed in Figure 3-1E and 3-1F. Even after PCA, the sample points on the two t-SNE dimensions are not better clustered compared with the original data in Figure 1-3A and 1-3B. The reason might be the data itself has similar variance isotropically, so PCA cannot find a principal component dimension on which data has significantly larger variance than on other dimensions.

Figure 3-1E

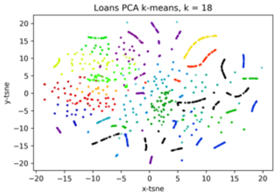
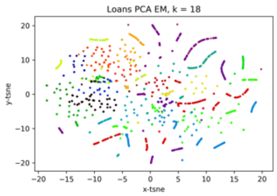


Figure 3-1F



Clustering on ICA transformed data. As discussed in Part 2, original data is transformed to 22 independent components by ICA for the clustering. With the transformed data, the k-means clustering SSE score curve shows an unobvious elbow position at k=15 to 18 (Figure 3-2A), and EM log probability (LL) curve shows an unobvious elbow position with k=18 (Figure 3-2B). As shown in Figure 3-2C, both clustering algorithms have highest average silhouette score at k =12 (k=15 without ICA, see Figure 1-1C in Part 1). However, the AMI scores of EM has highest value at k=3, but the AMI scores of K-means is close to 0 when k<10 (Figure 3-2D). Since k=18 gives the elbow position in k-means SSE plot and in EM LL plot, the scatter plot of 18 clusters projected on the first two t-SNE dimensions are displayed in Figure 3-2E and 2F. The clustering for k-means and EM after ICA are almost identical, but are quite different from those in Part 1 (Figure 1-3A, B). However, the sample points are almost uniformly distributed on the 2D t-SNE plane, indicating ICA cannot improve the clustering on this problem.

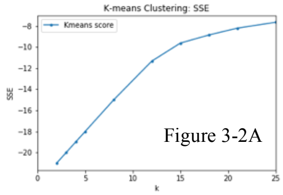


Figure 3-2A

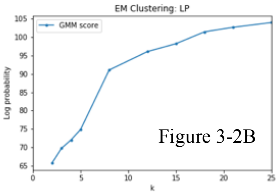


Figure 3-2B

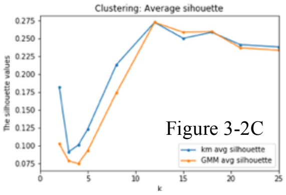


Figure 3-2C

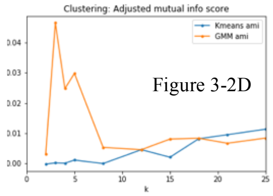


Figure 3-2D

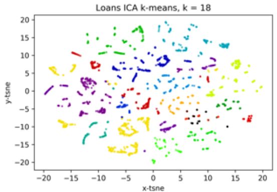


Figure 3-2E

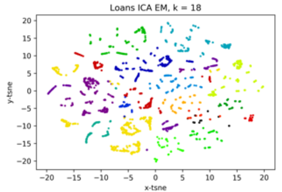


Figure 3-2F

Clustering on RP transformed data. As discussed in Part 2, original data is transformed to 21 randomized projected dimensions for the clustering. With the transformed data, the k-means clustering SSE score and EM log probability still do not have obvious elbow position (Figure 3-3A, B). The average silhouette score for k-means increase as the k value increases (Figure 3-3C); the silhouette score for EM has similar trend, but the score is highest when k=18. The AMI score curves of k-means and EM has opposite trends as those in Part 1 (Figure 1-1D). In Part 1, the k-means has best AMI score when k=2, but after RP transformation k-means has the highest AMI score when k = 3; while the EM has the highest AMI score when k =3 in part 1, but here it has the highest AMI when k =2. Since EM has highest average silhouette score when k=18, 18 clusters are generated by EM method and k-means algorithm, as shown in Figure 3-3E and 3-3F.

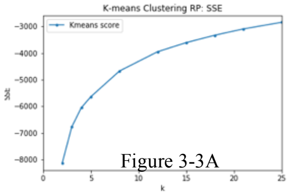


Figure 3-3A

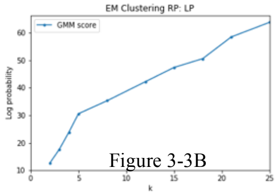


Figure 3-3B

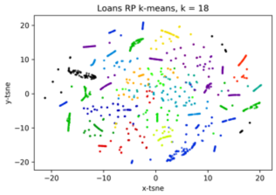


Figure 3-3E

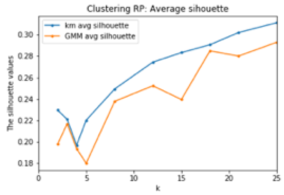


Figure 3-3C

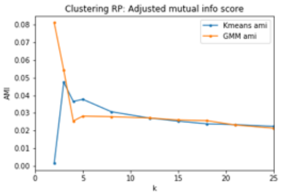


Figure 3-3D

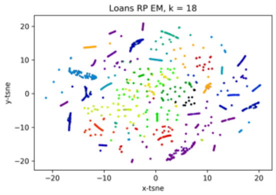


Figure 3-3F

Similar as the ICA transformed data, the sample points are almost uniformly distributed on the 2D t-SNE plane, indicating RP cannot improve the clustering on this problem.

Clustering on dimensions selected by RF. As discussed in Part 2, original data is transformed to 5 most important dimensions for the clustering. Distinct from all clustering discussed above, there is an obvious elbow position for k-means when k = 3 (Figure 3-4A) and an obvious elbow position for EM when k = 4 (Figure 3-4B). The average silhouette score is also high (0.7~0.8) when k = 3 or 4 for both methods (Figure 3-4C), and much higher than the score of original data, the PCA, ICA and RP transformed data (0.2~0.3), indicating better clustering. The AMI scores decrease monotonically as the k value increases (Figure 3-4D). Considering all these indicators, both k-means and EM should have k = 3, which is much less than using original data and PCA/ICA/RP transformed data. As shown in Figure 3-4E and 3-4F, with k=3, the cluster result of k-means and EM are identical. Interestingly, sample points are better clustered than original data and all the other transformed data. In both Figure 3-4E and F, a very obvious margin divides the bottom light green cluster and the black and blue clusters on the top. This result indicates that the 5 most important features selected by RF (loan amount, annual income of the applicant, loan term 36 months, loan term 60 months and credit grade A) are most informative and less noisy. All the other features may add noise to the clustering, which makes the clustering becomes harder and suffers from curse of dimensionality.

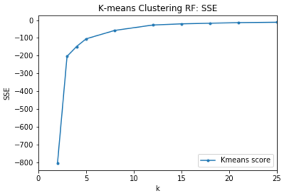


Figure 3-4A

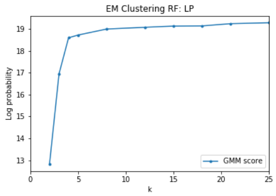


Figure 3-4B

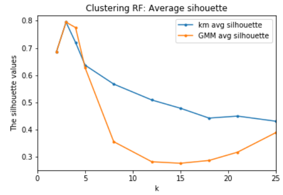


Figure 3-4C

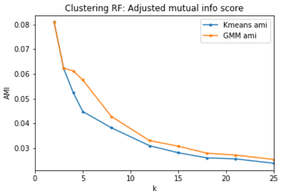


Figure 3-4D

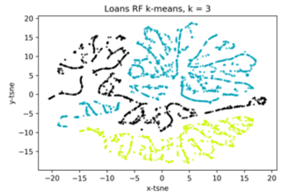


Figure 3-4E

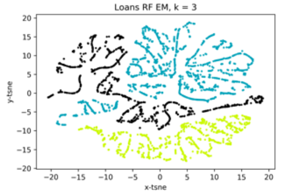
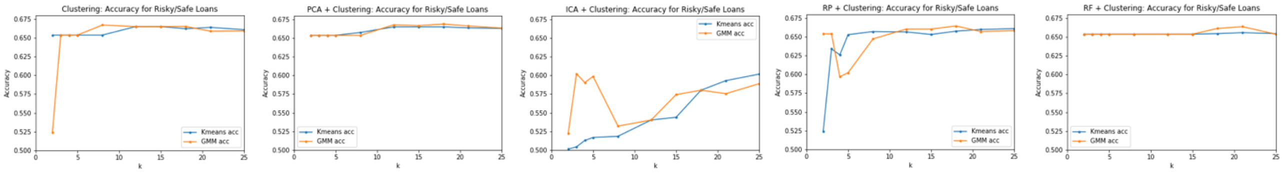


Figure 3-4F

Conclusion.

From the comparison of the clustering results of the original data (part 1) and the dimensionality reduced data, we can see that the Loans dataset is noisy (except for the the most important 5 features) and has isotropically variances in the whole space. Therefore, except for RF, dimensionality reduction (DR) techniques cannot improve the clustering too much. The cluster labels are compared with the true labels of the dataset below. The DR techniques cannot improve the clustering accuracy comparing with the original data, but PCA and RF transformed data are more robust with different number of clusters. ICA decreases the clustering accuracy in the full range of k. The two clustering algorithms k-means and EM do not have quite different clustering accuracy, except for ICA.



Pageblocks.

Clustering on PCA transformed data. As discussed in Part 2, original data is transformed to 6 principal components for the Pageblocks' clustering. With the transformed data, the k-means clustering SSE score (Figure 3-6A) and EM log probability (Figure 3-6B) still do not have obvious elbow position. Both k-means and EM have highest silhouette score when k =10 (Figure 3-6C), and the AMI scores are also high when k=10 (Figure

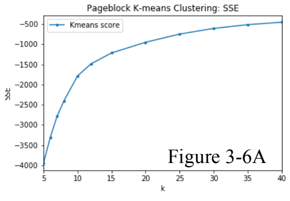


Figure 3-6A

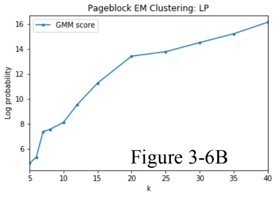


Figure 3-6B

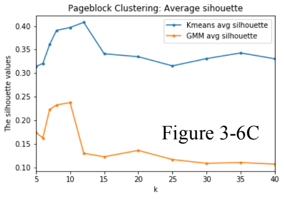


Figure 3-6C

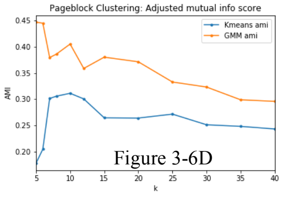


Figure 3-6D

3-6D). Unlike in Part 1, the k-means and EM clustering has high silhouette score and AMI score when k = 6. Using k=10, the scatter plot of clusters projected on the 2D t-SNE plane are displayed in Figure 3-6E and 3-6F. Though the same cluster number is used for k-means and EM clustering, the clusters generated by these two algorithms are not exactly same. The samples are better clustered after dimensionality reduction by PCA than the original data in Part 1 Figure 1-3C and D. This is reasonable, because PCA projects data to principal component that maximizes the variance of the data.

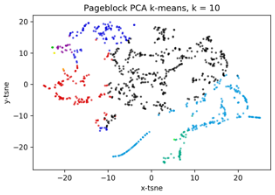


Figure 3-6E

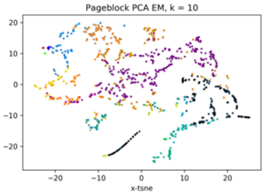


Figure 3-6F

Clustering on ICA transformed data. As discussed in Part 2, original data is transformed to 4 independent components for the clustering. On the k-means SSE curve (Figure 3-7A), there is no obvious position that SSE no longer grows. On the EM log probability curve (Figure 3-7B), when k ≥ 25, the log probability does not increase significantly. For k-means clustering, when k =8, it has the best silhouette score (Figure 3-7C) and AMI score (Figure 3-7D). For the EM clustering, when k = 25, it has the third highest AMI score but relatively low silhouette score.

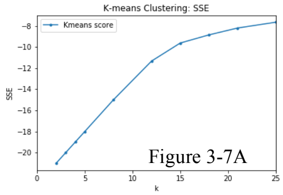


Figure 3-7A

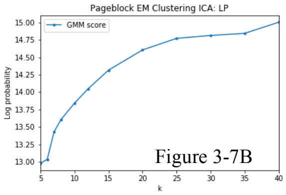


Figure 3-7B

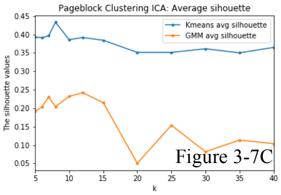


Figure 3-7C

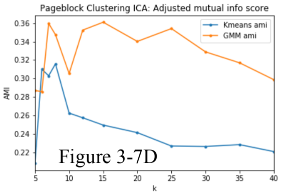


Figure 3-7D

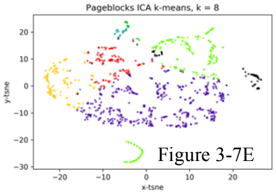


Figure 3-7E

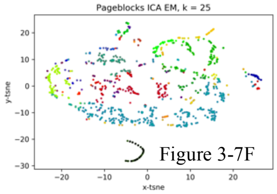


Figure 3-7F

Therefore, the scatter plot of k-means with 8 clusters and of EM with 25 clusters are shown in Figure 3-7E and 3-7F. After ICA, the samples are not more separated than the original data in Part 1 (Figure 1-3C and D), but more clear cluster boundaries are defined.

Clustering on RP transformed data. As discussed in Part 2, original data is transformed to 4 randomized projected dimensions for the clustering. Both the SSE curve of k-means (Figure 3-8A) and LP curve of EM (Figure 3-8B) reach a plateau when k =35. However, both k-means and EM have best silhouette score (Figure 3-8C) and AMI score when k = 7 (Figure 3-8D).

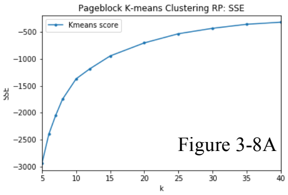


Figure 3-8A

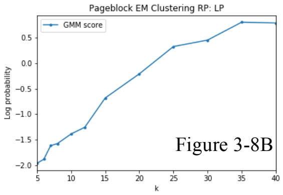


Figure 3-8B

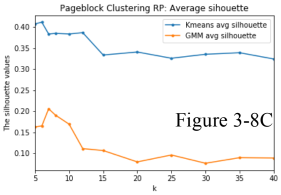


Figure 3-8C

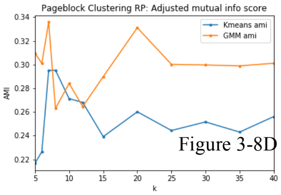


Figure 3-8D

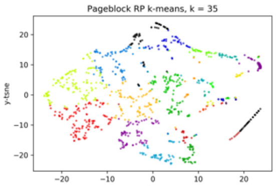


Figure 3-8E

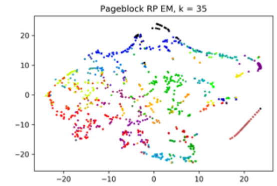


Figure 3-8F

The scatter plot of k-means and of EM with 35 clusters are shown in Figure 3-8E and 3-8F. After RP, the samples are not better clustered than the original data in Part 1 (Figure 1-3C and D), though the distribution changed.

Clustering on features selected by RF. As discussed in Part 2, original data is transformed to 5 most important dimensions for the clustering. With the transformed data, the k-means clustering SSE score and EM log probability still do not have obvious elbow position. Both K-means and EM have best silhouette score and AMI score when k = 8.

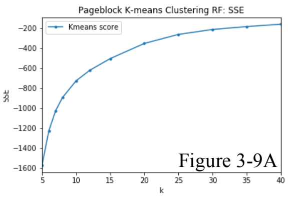


Figure 3-9A

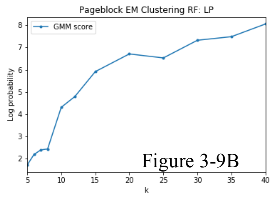


Figure 3-9B

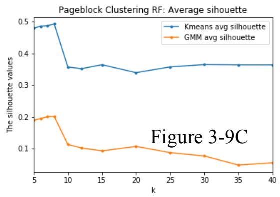


Figure 3-9C

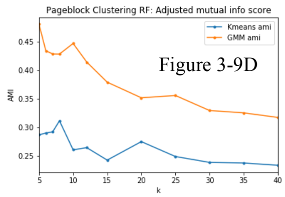


Figure 3-9D

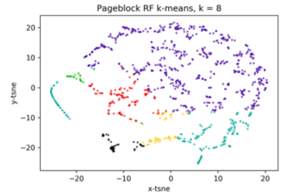


Figure 3-9E

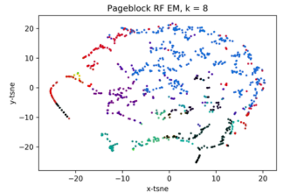
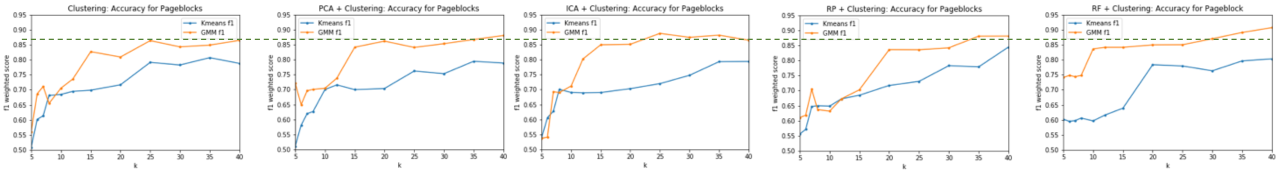


Figure 3-9F

The scatter plot of k-means and of EM with 8 clusters are shown in Figure 3-8E and 3-8F. With only the dimensions selected by RF**,** the sample points are more uniformly scattered on the t-SNE plane than the original data in Part 1 (Figure 1-3C and D), indicating the five eliminated features are actually helpful to cluster the data.

Conclusion.

From the comparison of the clustering results of the original data (part 1) and the dimensionality reduced data, we can see that the Pageblock dataset is more sensitive to the DR approaches than the Loans data. The reason might be that the Pageblocks dataset has all numeric features which are very dense, unlike the Loans data has many one hot encoded categorical features. The cluster labels are compared with the true labels of the Pageblocks dataset below. The accuracy always increases as the cluster number increases until reaches the plateau. The EM algorithm perfoms better than k-means for all the cases. As indicated by the horizontal dash line, the EM clustering on DR data can get higher accuracy with sufficiently number of clusters.



Part 4. Neural network learning after dimensionality reduction.

The Pageblock classification dataset is selected for this task and Part 5, because all the features of this dataset are numeric, for which the dimensionality reduction methods have more significant effects. For both part 4 and part 5, the sklearn.neural_network.MLPClassifier with hyperparameters solver='lbfgs', activation='logistic', alpha=0.1, hidden_layer_sizes=(50,) was used for the original data and the data after dimensionality reduction (DR). To obtain reliable accuracy, k-fold cross-validation is performed for neural network classification on the original and the transformed data. The result of the original data is the benchmark for this section.

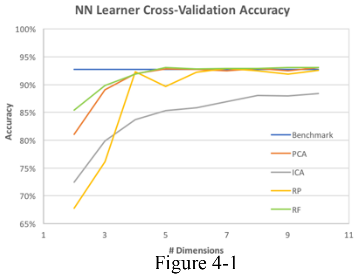


Figure 4-1

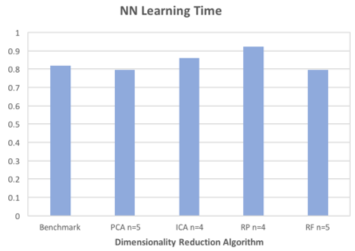


Figure 4-2

The cross-validation accuracy of neural network (NN) learner using data transformed from each of the 4 DR algorithms is compared with the benchmark in Figure 4-1. The clock time of NN learning on the dimension reduced data are shown in Figure 4-2. The classification accuracy of ICA is far below the accuracy of other algorithms with all different dimension numbers, by about 4% when dimension number is 10. The data transformed by PCA performs best, even with only 5 transformed dimensions its accuracy (92.86%) beats the benchmark accuracy (92.76%). Moreover, the running time on PCA transformed data (5 dimensions) is shorter than that of the benchmark. The data with features selected by RF performs as good as PCA transformed data.

With 5 dimensions, its accuracy outperforms over the benchmark by 0.29%, and with less running time. The performance of NN learner on the RP transformed data is also decent. With 6 or more dimensions, the accuracy of NN learner on the RP transformed data is similar as the benchmark. Even though the NN learning needs longer time for the RP transformed data on this dataset, but the data transformation time of RP is faster than PCA and ICA. This advantage will be more obvious when the original data has a great number of features and samples.

Part 5. Neural network learning on data after dimensionality reduction and clustering.

The k-means and EM clustering are performed on dimensionality reduced data, the same process as Part 3. Since the cluster numbers are larger than 2, the cluster labels are one hot encoded and then combined with the dimension reduced data or the original data as the input for the NN learning. As shown in Figure 5-1, using original data combined with clustering results (the pale blue and pale orange bars), the NN classification always has similar accuracy as the benchmark (NN learning on original data without DR and clustering in Figure 5-1). The original data with clusters generated by ICA transformed data and by RF transformed data performs a little better than the other two. However, different DR data combined with clustering results (the dark blue and dark orange bars in Figure 5-1) has significantly different accuracy. The data whose features selected by RF performs the best, no matter it combines with k-means clusters or EM clusters it outperforms the benchmark. The PCA transformed data with k-means clustering has the same accuracy as benchmark, while the same with EM clustering has worse accuracy than the benchmark by about 1.24%. RP transformed data with k-means clustering and EM clustering has worse accuracy than the benchmark by 0.76% and 1.90%, respectively. The ICA transformed data performs much worse than the benchmark, by 11.05% with k-means and by 4.76% with EM.

The running time of each combination of DR and clustering is displayed in Figure 5-2. Most combinations have similar running time as the benchmark. Only the RP transformed data always has longer running time than the benchmark. NN learning on ICA transformed data with EM clusters has very short running time, but bad accuracy. The running time of RF selected data and original data with EM cluster labels is significantly shorter than benchmark. Interestingly, they are also the combinations resulting in best classification accuracy.

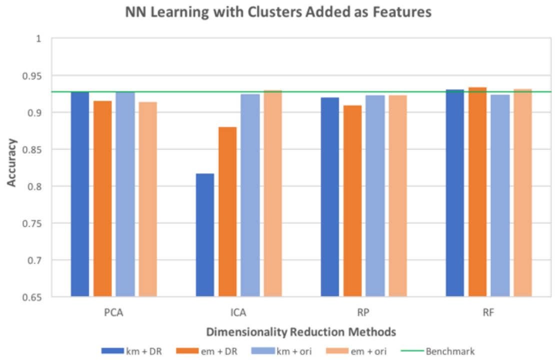


Figure 5-1

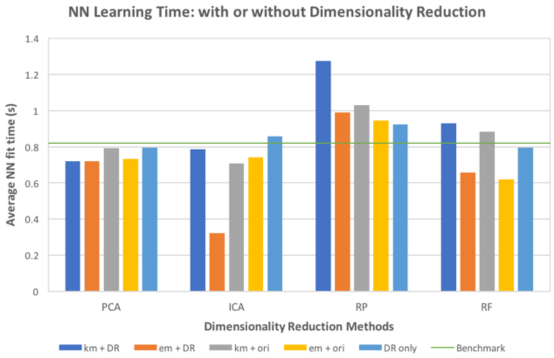


Figure 5-2

Conclusion.

Using the clusters obtained from the DR transformed data as new features is not necessarily helpful to improve the classification accuracy of supervised learning. Because the number of clusters generated from different DR transformed data can be quite different from the real number of classes in the dataset. The clusters from RF selected data improve the performance of the NN learner because 8 clusters are determined from the RF selected data for both k-means and EM. This cluster number is relatively consistent with the true number of classes which is 5 in this Pageblocks dataset. More clusters are generated by the other DR transformed data, 10 clusters for PCA, 25 clusters for ICA and 35 for RP. The biased cluster labels may mislead the supervised classification.