CS7641: Analysis and Report for Assignment 3- Unsupervised Learning and Dimensionality Reduction

Mingming Zhang mzhang607@gatech.edu

Description of classification datasets

In this assignment, I used the two datasets from Assignment 1. The MINIST dataset, which includes the handwriting digits (*http://yann.lecun.com/exdb/mnist/*). The image data includes 10,000 different examples in total, each image represents one of the *10* handwriting digits from *0* to *9*, the data also contains the correct human label to indicate which number each handwriting digit is. The second dataset that I select is the wine quality dataset (*https://archive.ics.uci.edu/ml/datasets/Wine+Quality*). Here, I combined the red wine and white wine data sets, so the total number of samples are 6497 now. Because the raw wine dataset is extremely imbalanced, with each sample with discrete labels from wine score 3 to 9, it makes a very difficult task to predict the wine score directly. Thus, I reformed the data into 'Bad' and 'Good' wine, with wine score that is higher than 6, including 6, defined to be the good wine. This approach makes the prediction on a more balanced dataset compared to the original dataset.

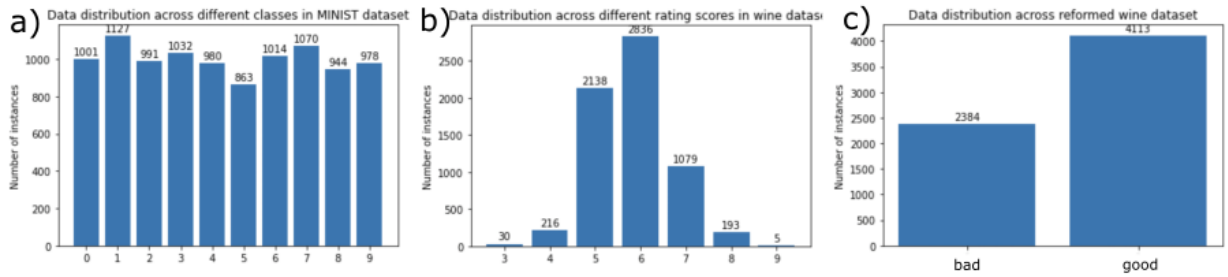

Figure 1: distribution of number of instances across different classes. a) the MINIST data set are with evenly distributed instances across 10 classes. b) wine data set is with severe imbalance issue across different wine scores from 3 to 9. c) reformed wine dataset with two classes, good and bad wine, the instances are better much less imbalanced issue compared to raw wine dataset shown in b).

Clustering on wine quality and MINIST image datasets

First, we apply k-means and Expectation Maximization (EM) cluster on the two datasets mentioned above, but without using the ground truth label in the clustering task.

When using K-means clustering, I ran a series of experiments by changing the number of clusters the algorithm is trying to classify. At each given cluster, I compute the Silhouette average score and distortion along the number of clusters. The silhouette coefficient for a single sample the dataset is defined as $s = (b - a)/max(a, b)$, where a is defined as the mean distance between a sample and all other points in the same class; b is defined as the mean distance between a sample and all other points in the next nearest cluster. Finally, I calculate the mean silhouette coefficient for all samples. Usually, the higher Silhouette coefficient indicates the model with

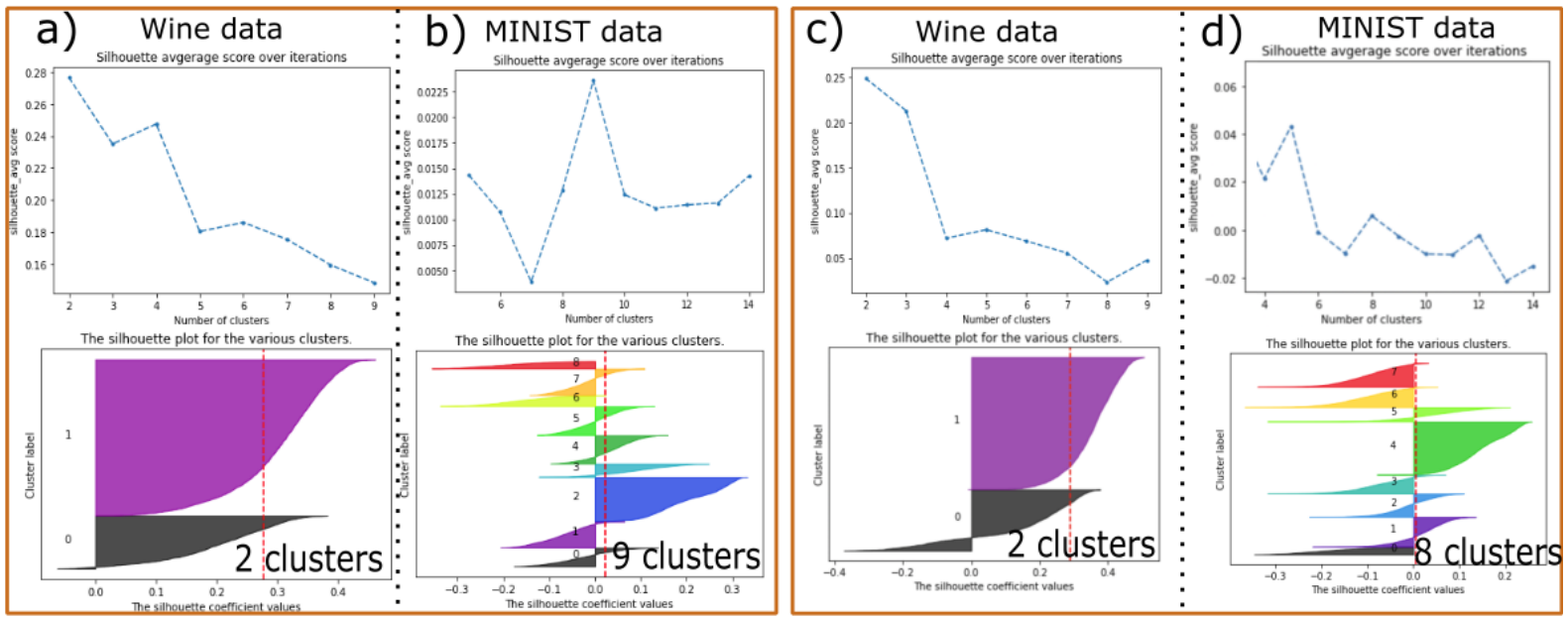

Figure 2: **K-means and EM clustering on original wine and MINIST data**sets. In each group (column), **top figure** shows the average Silhouette as the number of clustering changes; **bottom figure** shows the individual Silhouette scores for each cluster. The dashed line shows average Silhouette score across all the samples.

better defined clusters; values near 0 indicates overlapping clusters; negative values indicate a sample has assigned to the wrong cluster.

Results indicate that K-means clustering algorithm thinks there are two clusters that is in the wine dataset, because Silhouette score is highest when number of clusters is 2 (Fig.2a). This matches the true labels as mentioned in the previous section that the wine data has 'good' and 'bad' wine classes. The assigned clusters for wine dataset matched about 62.3% of the ground truth labels (Fig.3). When using MINIST dataset, it shows there are 9 classes, which seems close to our real label as we have 10 classes (Fig.2b). However, when we look at the individual Silhouette coefficients, we notice there are a

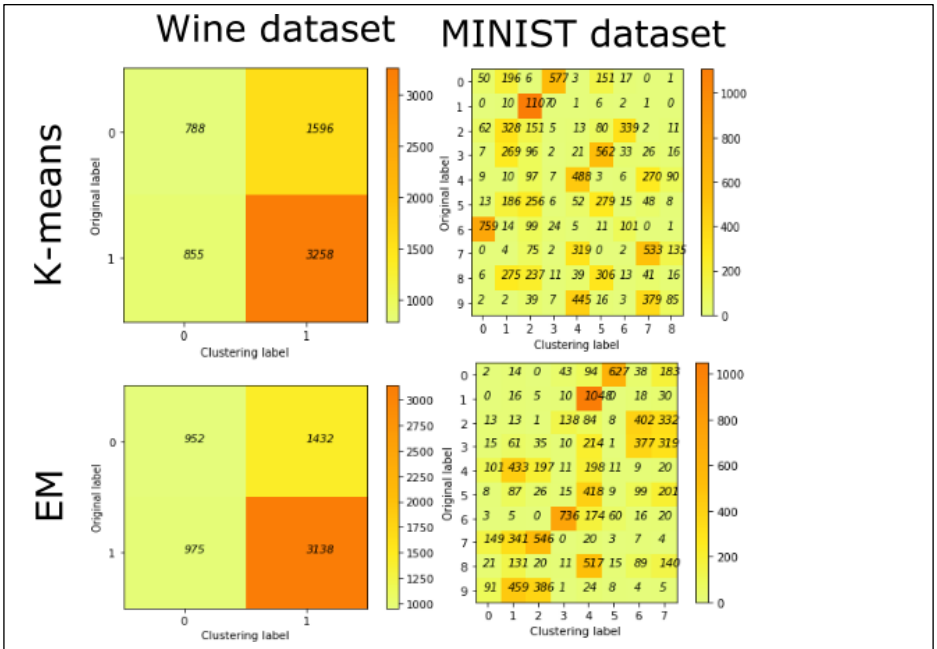

Figure 3: **K-means and EM clustering on original wine and MINIST data**sets. In K-means clustering, I selected 2 clusters and 9 clusters for wine and MINIST dataset, respectively. In EM clustering, I selected 2 clusters and 8 clusters for wine and MINIST dataset, respectively. Each color maps show the true labels V.S. clustering labels with the number of samples in each correspondence.

lot of classes that has negative coefficients (Fig.2b), which means these classes are likely to be assigned to a wrong cluster. The original and clustering label matrix shown in Fig.3 shows no consistent pattern that an original label only mainly matches a clustering label. This indicates that K-means clustering labels match the original labels poorly.

When using EM clustering algorithm on wine dataset, I observed similar results that it indicates wine dataset has two clusters and it generally matches well in the original two classes (Fig.2c and Fig3). When using MINIST data with EM clustering, I ran experiment from 4 to 14 clusters and see 5 cluster has the highest Silhouette score (Fig.2d). However, when look at individual Silhouette scores, there is one class that is with majority individual score that is positive, and the rest of other clusters with negative score. To consider an even distribution of individual Silhouette scores with more cases crossing average Silhouette score, I choose 8 clusters. Although, I must admit that the clustering results is not very good (Fig.2d). The original and clustering labels matrix indicates a more even distribution of the number of samples (Fig.3), which means the clustering labels match poorly with original labels.

Using PCA to conduct dimensionality reduction

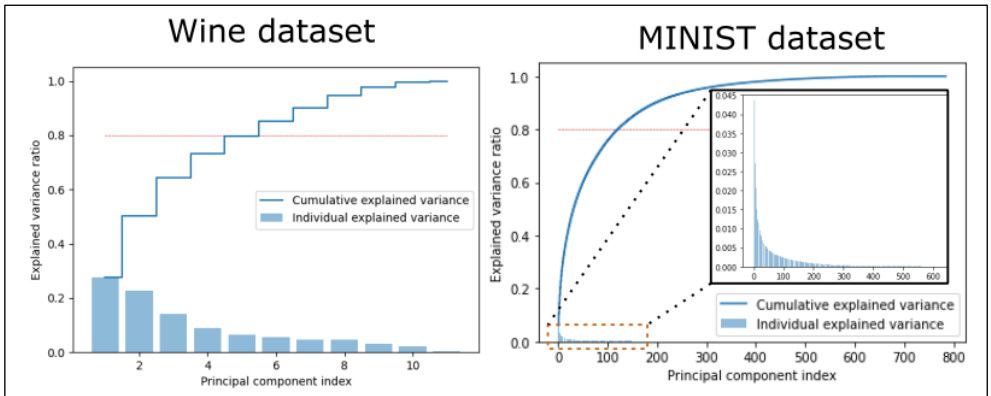

Figure 4: PCA analysis. On the wine dataset, we see that the first 5 components could have a combined explanation of 80% of the variance. On the MINIST image data set, we see that the first 121 components have a combined explanation of 80% of the variance in the dataset.

Next, I conduct PCA analysis to reduce the dimensionality of the data. Here, I set the threshold to look at when the combine first a few components could explain 80% of the combined variance in the dataset, then I will keep these first few components. From Fig.4 we can tell, the first 5 components in the wine dataset can explain about 80% of the variance. The first eigenvalue can explain about 27%, and the 5th eigenvalue can explain about 6.5% of the total variance. In the MINIST data, we will need to include the first 121 components to have the total explained variance to be about 80%. Compared to original dataset with 784 features (pixels) in the data, it has already been heavily reduced in dimensionality. Unlike observed in wine dataset, each component in MINIST dataset can only explain a little variance, with the first component explaining 6.1% and 121st component explaining 0.18% of the total variance, respectively. This is because each feature in MINIST data is a pixel in an image and thus contains only very little information if used alone.

K-means and EM clustering algorithm continues to indicate there are 2 clusters in the PCA reduced wine dataset, as in both experiments, the number of 2 cluster has the highest Silhouette score. The visualization on the clustering data from both algorithms also indicate a clearly aggregated datapoints (Fig.4a, c). When conduct experiments on MINIST dataset from 4 to 14 clusters, both K-means and EM clustering algorithm indicates that number of 4 clustering has the

highest Silhouette score (Fig.5b, d). For K-means clustering, I didn't choose 4 clusters but next high Silhouette scores at 9 clusters, because the individual Silhouette score is more evenly across the average Silhouette score (Fig.5b). For EMG clustering, I did use 4 clusters even a lot of individual samples have Silhouette score not crossing average Silhouette score. This is because the Average Silhouette score will become negative when the number of classes is higher than 6 (Fig.5d), indicating a lot of samples are likely to be assigned to a wrong class. The visualization of PCA reduced MINIST datasets are not clear when using both clustering algorithms (Fig.5b, d), matching what observed before the clustering labels are not corresponding very well with original labels.

I then compute the distance of each sample to corresponding clustering centroid. I added the distance and the clustering labels to both PCA reduced dataset as features, respectively.

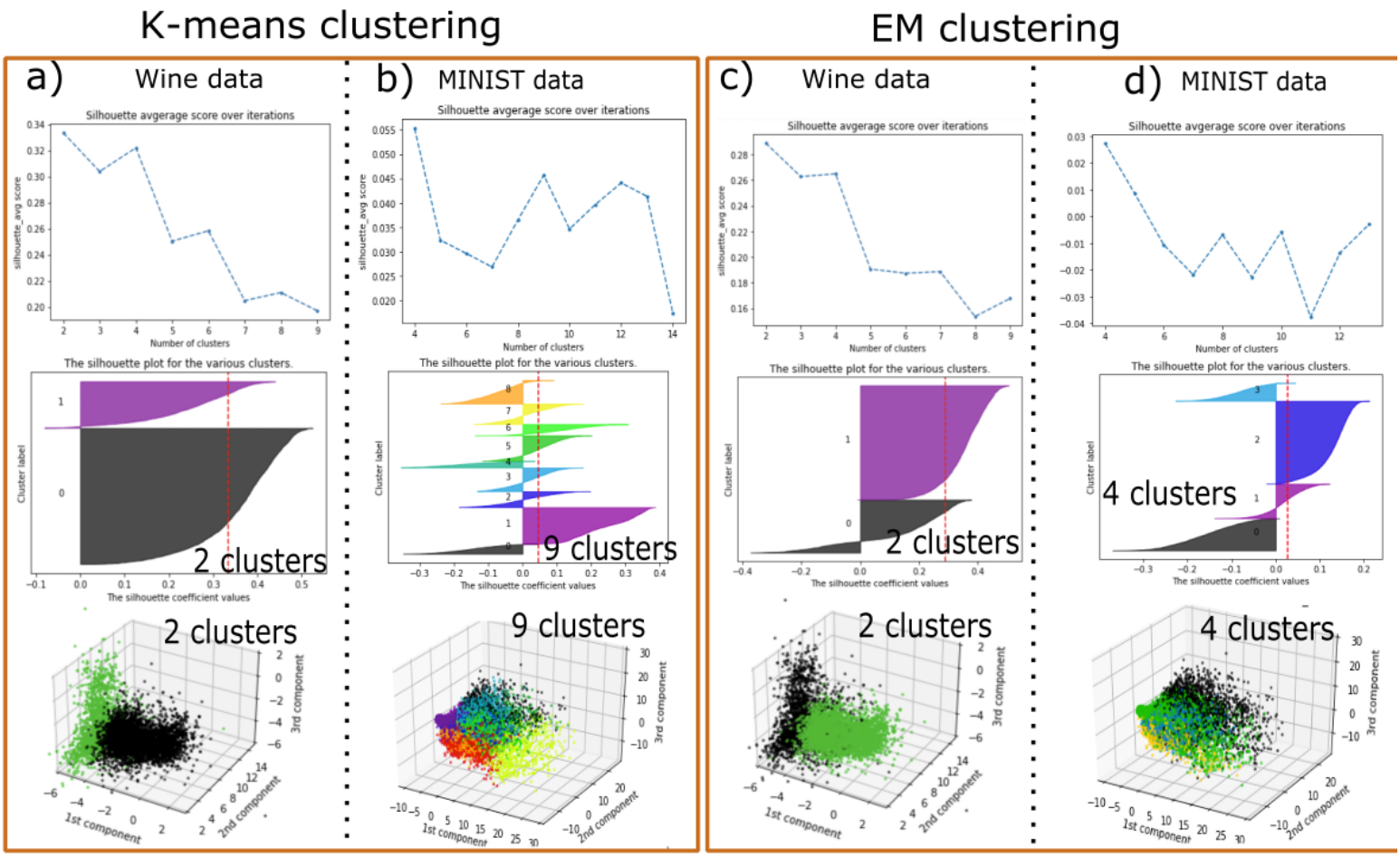

Figure 5: K-means and EM clustering on PCA reduced data. In each group (column), **top figure** shows the average Silhouette as the number of clustering changes; **middle figure** shows the individual Silhouette scores for each cluster. The dashed line shows average Silhouette score across all the samples; **bottom figure** shows the clustering visualization of the reduced data with the first 3 components.

Using ICA to conduct dimensionality reduction

For ICA experiments, I run the number of components check the average kurtosis as the number of components changes. Kurtosis measures the "peakness" of a distribution relative to a normal distribution. The average Kurtosis can be positive and negative. When Kurtosis is zero, the variable is Gaussian; and when the value is positive, the variable is said to be super gaussian; when the kurtosis is negative, the variable is said to be sub gaussian. Here we choose the number of components with the higher Kurtosis, the non-gaussianity, which is the key in ICA analysis.

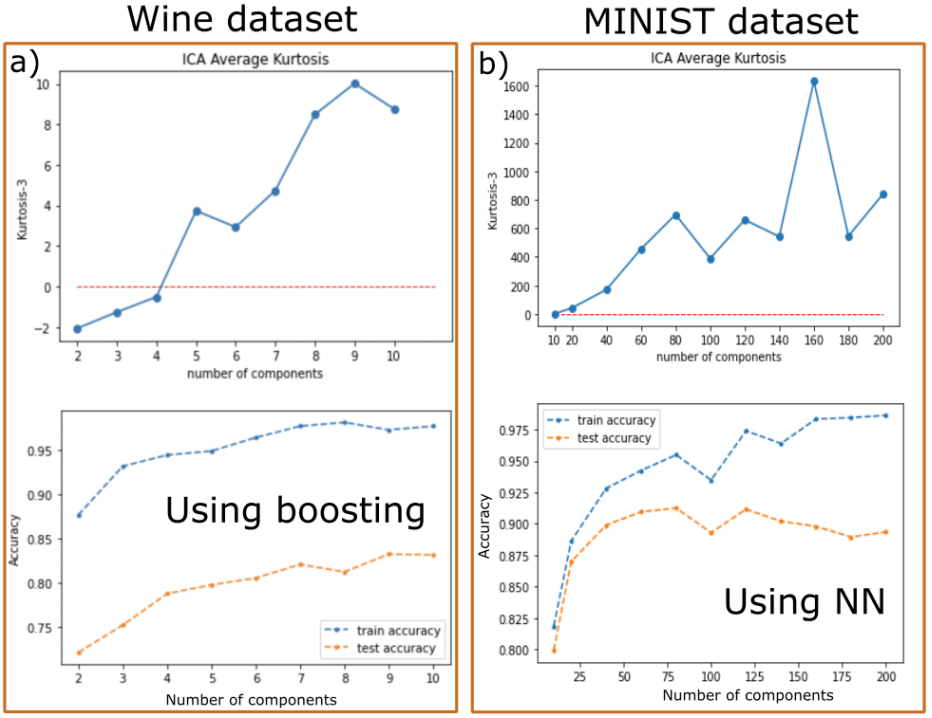

Figure 6: ICA analysis. Left column for wine dataset: top panel shows the average Kurtosis changes as the number of components; bottom panel shows using boosting algorithm to train and test the data when using different number of components. **Right column for MINIST dataset**: top panel show the average Kurtosis changes as the number of components; bottom panel shows using Neural Networks algorithm to train and test the data when using different number of components.

When using wine dataset, because wine dataset has 11 attributes, I used number of components up to 10 as the purpose of using ICA to conduct dimensionality reduction. When number of components is 9, the Kurtosis curve has the highest score. In parallel, I split the ICA reduced dataset into training (80%) and testing (20%) data and conduct experiment for each number of components using boosting algorithm (as boosting is the one performing the besting in prediction wine dataset in Assignment 1). We can see that when number of components is 9, it has the best testing accuracy at about 83% (Fig.6a). Considering the two experiments I did above, I decided to use 9 components in ICA experiment for wine dataset. When using MINIST dataset, I selectively investigate number of components from 10 to 200 and get the highest average kurtosis when the number of components is 160. I also test using neural networks to conduct training and testing on the ICA reduced dataset (NN is one of the algorithms performs well on MINIST dataset in Assignment1). We can see the performance will decrease a little when the number of components pass 120, but the degree of decrease is very little (Fig.6b). Thus, I decide use 160 components in ICA analysis for MINIST dataset.

When it comes to clustering on the ICA reduced datasets, both K-means and EM clustering indicates there are two clusters in the wine dataset (Fig.7a, c). When using MINIST dataset, K-means indicates 3 cluster has the highest Silhouette score, while with only one clusters that is mainly crossing average Silhouette score, and the other two classes are with scores close to zero

(Fig.7b). EM clustering indicates there are two clusters in the dataset and with only one cluster mainly cross the average Silhouette score (Fig.7d). This observation is consistent with before that the clustering matches well on wine dataset, while the clustering labels in the MINIST data sets matches poorly on the original labels of the image data.

I then compute the distance of each sample to corresponding clustering centroid. I added the distance and the clustering labels to both ICA reduced dataset as features, respectively.

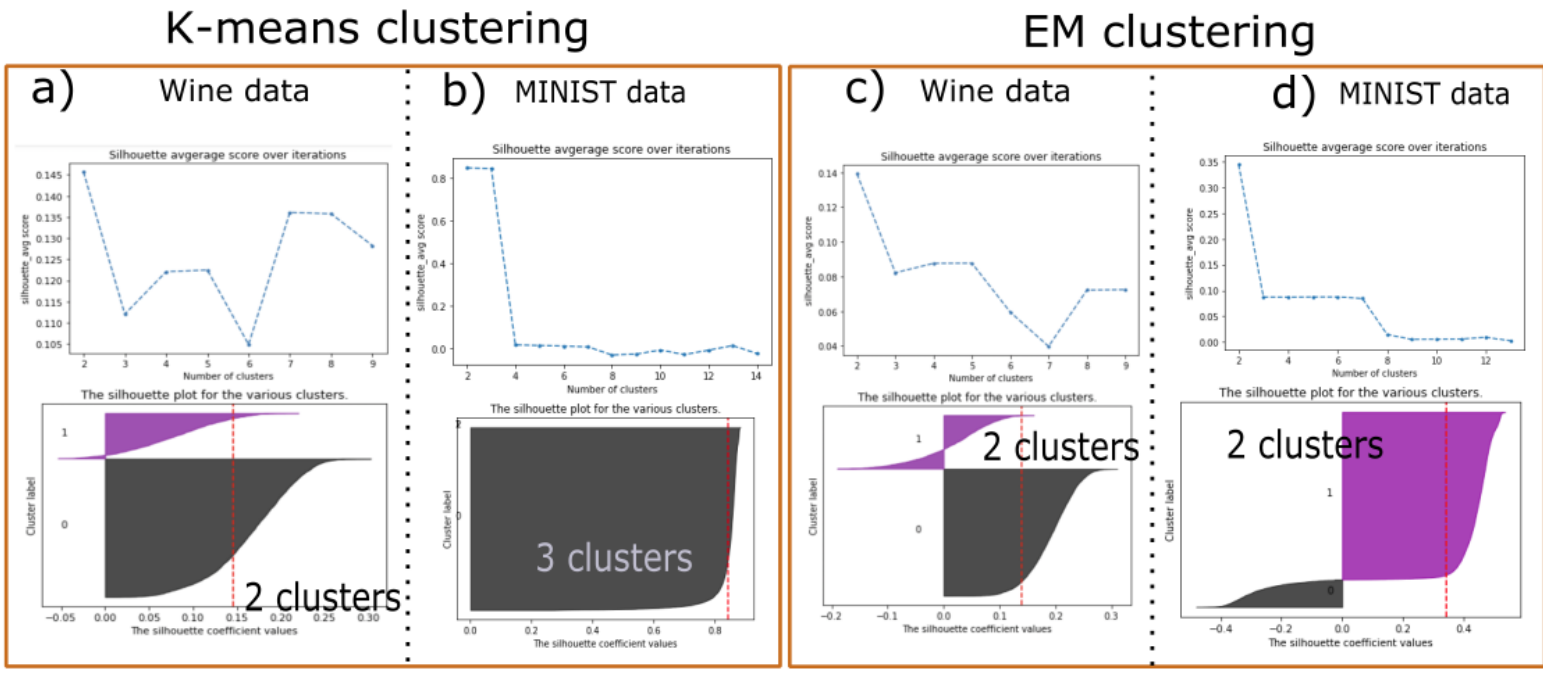

Figure7: K-means and EM clustering on ICA reduced data. In each group (column), **top figure** shows the average Silhouette as the number of clustering changes; **bottom figure** shows the individual Silhouette scores for each cluster. The dashed line shows average Silhouette score across all the samples.

Using Random Projection to conduct dimensionality reduction

The determined minimum number of dimensions of transformed data is determined by the sample size of the data. Ideally, random projection (RP) requires the dataset to have thousands of features. For wine and MINIST dataset, based on the sample size of each dataset, it will need ideally 707041 and 741772 features, respectively. However, there are only 11 and 784 features in wine and MINIST dataset, respectively. Thus, I conduct different experiment using boosting algorithm and NN algorithm to see what number of components to be used in RP could result the best performance. These two algorithms perform well in classifying each dataset when doing Assignment1. For each given number of components, I conduct RP first and split the RP reduced data into training (80%) and testing (20%) data. Then, I use the selected algorithm to conduct experiment. The number of components change from 2 to 11 in wine dataset, and the number of components change from 5 to 200 in MINIST dataset. Figure 8 shows when the number of components is 7, boosting algorithm has the best performance on wine dataset. When the

number of components reach at 160, NN algorithm maximize its performance on MINIST dataset, and the performance will not improve if increasing components (Fig.8). Thus, I decide to use 7 and 160 components for wine and MINIST datasets, respectively, in RP dimensionality reduction.

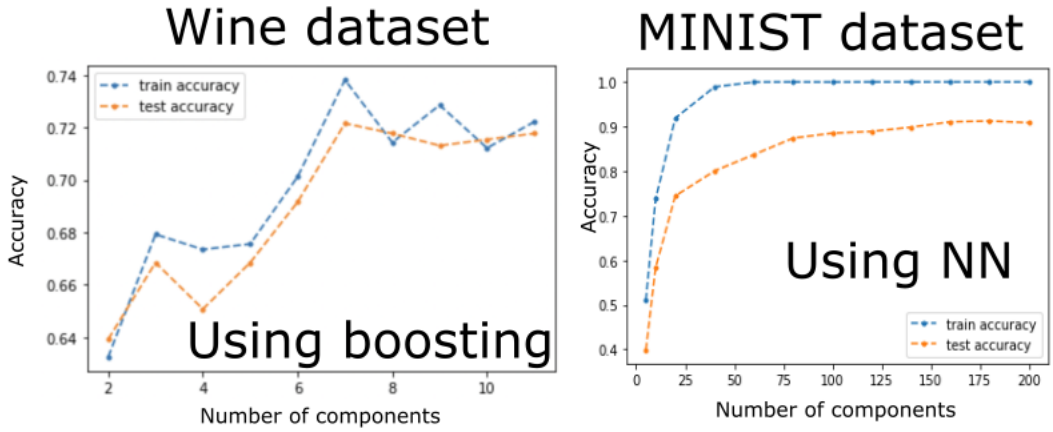

Figure 8: Random Projection (RP) analysis. **Left**: using boosting algorithm to conduct training and testing experiment on the RP reduced data as the number of components used in RP changes. **Right**: using Neural Network algorithm to conduct training and testing experiment on the RP reduced data as the number of components changes.

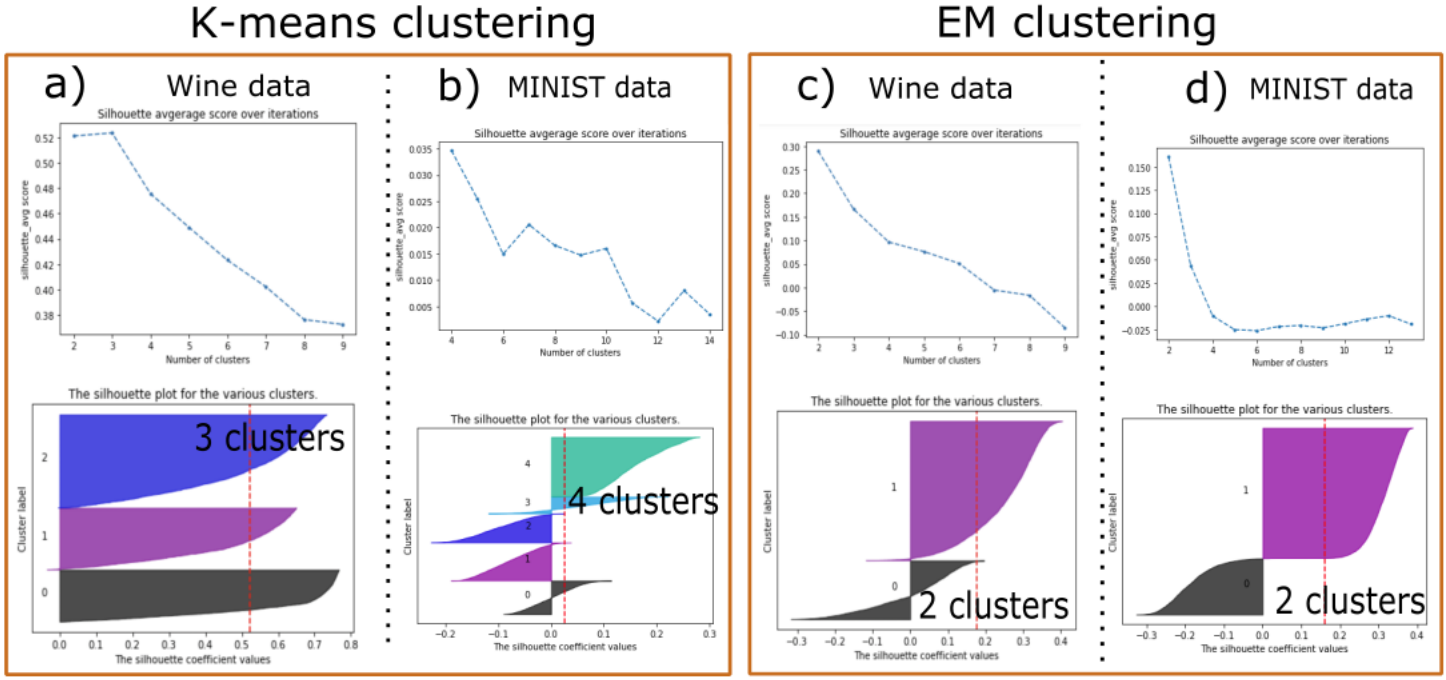

Figure 9: K-means and EM clustering on Random Projection reduced data. In each group (column), **top figure** shows the average Silhouette as the number of clustering changes; **bottom figure** shows the individual Silhouette scores for each cluster. The dashed line shows average Silhouette score across all the samples.

The results show K-means clustering indicates there are 3 clusters in wine dataset (Fig.9a) and 4 clusters in MINIST dataset (Fig. 9b). EM clustering indicates there are 2 clusters in wine dataset

(Fig.9c) and 2 clustering in MINIST dataset (Fig.9d). The clustering on wine dataset with majority of the individual samples crossing average Silhouette score means the clustering is confident is clearly distinguished; while clustering on MINIST dataset is generally with a lot of samples that are not crossing average Silhouette score or with negative score means there are a lot of samples that is likely assigned to the wrong class.

I then compute the distance of each sample to corresponding clustering centroid. I added the distance and the clustering labels to both RP reduced dataset as features, respectively.

Using Random Forest Algorithm to conduct dimensionality reduction

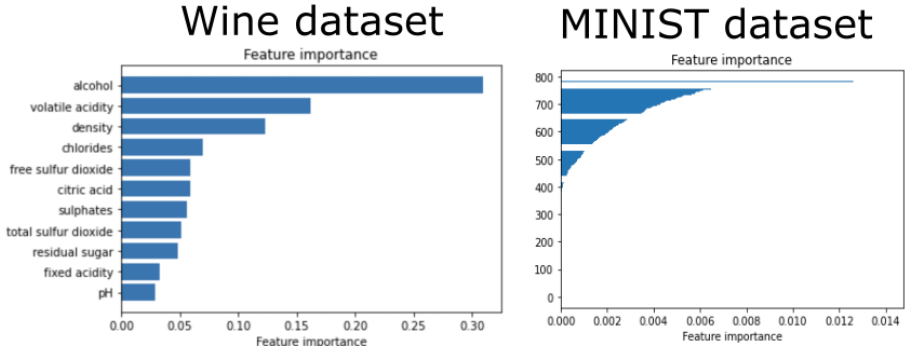

Figure 10: Random Forest (RF) analysis. **Left**: using random forest to classify the wine data and display the importance of features. **Right**: using random forest to classify the MINIST data and display the importance of features.

Last, I used random forest (RF) algorithm to conduct the experiment with true labels in both datasets. After model has been training, I pulled out the importance of the features. The results indicates that there are certain features that is dominantly important in wine dataset, such as alcohol level. In MINIST dataset, the

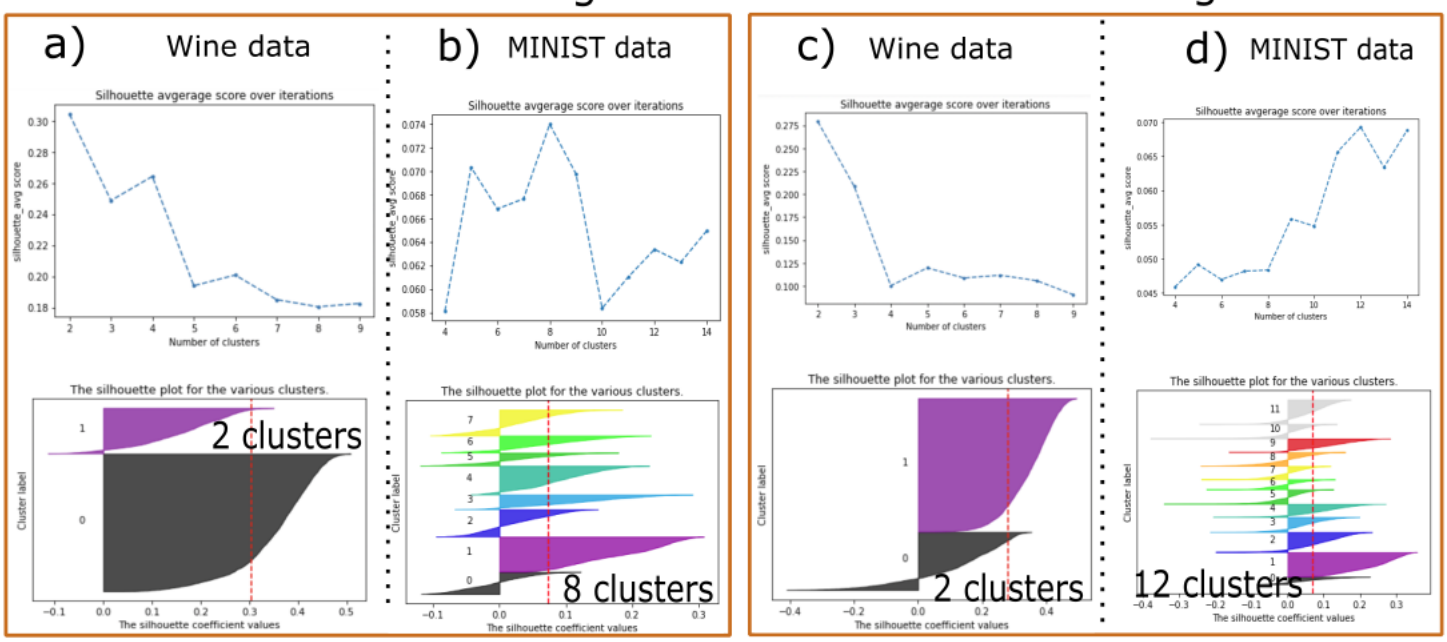

Figure 11: K-means and EM clustering on Random Forest reduced data. In each group (column), **top figure** shows the average Silhouette as the number of clustering changes; **bottom figure** shows the individual Silhouette scores for each cluster. The dashed line shows average Silhouette score across all the samples.

result show that there are about 450 pixels out of 784 pixels that contribute very little in the training of the model (Fig.10). As a results, I decide to remove the last 30% of the features in the wine dataset and remove the last 60% of the features that are least important in MINIST dataset to generate RF reduced datasets for both.

On the RF reduced wine datasets, K-means and EM clustering both indicates there are 2 clusters (Fig.11a, c). When the clustering is performed on RF reduced MINIST dataset, K-means indicates there are 8 clusters (Fig.11b) and EM clustering indicates there are 12 clusters (Fig.11d). Similarly compared to previous observations, clustering on wine dataset is more confident on the clustering results and it matches the original labels with 2 classes. While clustering on MINIST data has more individual samples that are with negative Silhouette score means the results is less confident it is correct. However, I see the classes are more evenly distributed in cross average Silhouette score on MINIST dataset compared to when using ICA and RP algorithms.

I then compute the distance of each sample to corresponding clustering centroid. I added the distance and the clustering labels to both RF reduced dataset as features, respectively.

The performance on Neural Network when conducting experiments with these datasets

Table 1: Neural Network model performance on different datasets

		MINIST data		Wine data	
		training	test	training	testing
	Original data	**1.0**	**0.924**	**0.7648**	**0.7592**
PCA	Reduced data	1.0	0.9385	0.7828	0.7546
	Reduced data + K-means features	1.0	0.93	0.7897	0.7577
	Reduced data + EM features	1.0	0.931	0.8008	0.7569
ICA	Reduced data	0.9798	0.915	0.7362	0.7392
	Reduced data + K-means features	0.9803	0.907	0.7427	0.7438
	Reduced data + EM features	0.9725	0.9125	0.7394	0.7438
Random projection	Reduced data	1.0	0.913	0.7575	0.7453
	Reduced data + K-means features	1.0	0.9055	0.7337	0.7377
	Reduced data + EM features	1.0	0.9065	0.7388	0.7353
Random forest	Reduced data	1.0	0.947	0.8274	0.7692
	Reduced data + K-means features	1.0	0.94	0.7876	0.7754
	Reduced data + EM features	1.0	0.947	0.8232	0.78

Finally, I conduct experiment on all the datasets using the neural networks. The model used here is with the same architecture, with 3 layers (30, 50, 30 units in each layer, respectively). This time, the experiment I conduct is with some modification based on Assignment 1. First, I reformed the wine dataset into ‘good’ and ‘bad’ wine classes, instead of looking at its original wine scores. I standardized both reformed wine dataset and MINIST dataset before I conduct the experiments.

Table 1 shows the performance of the models with the same neural network architecture across different datasets.

When using original wine dataset, the testing accuracy is 75.92%. We can see that, when using PCA, ICA and RP methods to conduct dimensionality reduction, the performance of the model is not better compared to using raw data. But the testing accuracy is relatively close to that when using raw data. This indicates these dimensionality reduction methods were able to capture important general information of the wine dataset, while keep the dimensionality lower compared to original dataset. When I used the random forest methods to remove the non-important features, the performance of the neural network model is with better testing accuracy compared to using original data, this indicates that RF reduced dataset was not only keep the important information, but able to remove some noise in the original data that might be affecting the model's performance.

When using original MINIST dataset, the testing accuracy is at 92.4%. When ICA and RP methods to conduct dimensionality reduction, the testing accuracy is lower than when using original data. When using PCA and Random Forest to reduce the dimensionality of the data, the final testing accuracy is over 93% and 94%, respectively, for different dataset with additional K-means and EM clustering features. From the MINIST image dataset, we know that an image has a lot background pixels at the edge of each image, which contains little information. Both PCA and Random Forest based methods were able to remove a lot of these background pixels, and thus makes the training data not just small but with more relevant information. I think it is why we are seeing improvement when using PCA and RF based dimensionality reduction. In conclusion, PCA, ICA, RP and random forest-based dimensionality reduction reconstruct both wine and MINIST data well with major important information preserved, as the performance of neural network model performs similarly when using all these different dimensions reduced datasets. The clustering that I observed when using PCA, ICA and RP are different for both wine dataset and MINIST dataset. The clustering might be similar in wine dataset, but very different when using MINIST dataset. I think the reason is that these methods use very different ways in projection their components. One interesting thing that I notice, when using RP method to conduct dimensionality reduction, the performance of the neural network model could have difference up to 2% in both of my dataset. I think the reason is that random projection uses a random matrix drawn from Gaussian distribution to project data from high dimension to lower dimension. As the random drawing from Gaussian distribution will be different each time running the experiment, the experimental results will then be different.

In addition, we can see when add additional k-means and EM clustering features (calculated distance and clustering labels for each sample), it might help the model's performance when using wine dataset and the help is very little. While using MINIST dataset, the added clustering features could worsen the model's performance. One of the reasons could be the MINIST data, even with dimensionality reduction, the dimension is still very high, like 80 to 100. The unsupervised learning method like k-means will naturally have a hard time with such high dimensionality data. Also, each feature in the MINIST data is pixels, but not different types of features like that in wine dataset. Thus, this could further prevent simple clustering algorithm to perform well directly on MINIST raw dataset or reduced dimensionality dataset.