Dantong Zhu
CS 7641
Project 3

Unsupervised Learning and Dimensionality Reduction Write-up

The Datasets

We are going to use the forest fire dataset and (red) wine quality dataset in this project. The forest fire dataset is about forest fires in the northeast region of Portugal. There are 517 data points total, each with 12 attributes, including location, time, weather, and etc. The original label for each data point is the buring area of each fire, or non-fire in the case buring area = 0. The goal is to predict the burning area of a some fire given the 12 attributes. In our classification problem, we label a data point as 0 if the burning area is strictly smaller than 5, and as 1 otherwise. The wine quality dataset is about wine samples from the north of Portugal. There are 1599 data points with 11 attributes of chemical information for the wine, including acidity, sugar, pH, and etc. The label for each data point is the quality of the wine, an integer score between 0 and 10. The goal is to using the chemical information attributes to predict the quality score of the time. Both datasets are downloaded from the UCI library, and more details about the datasets can be found in README.txt.

Overview of the project

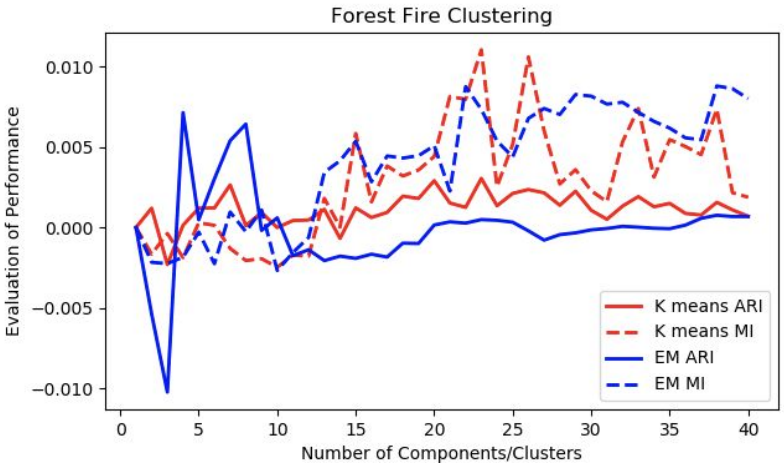
There are 5 tasks in this project. For Task 1-3, we are going to run clustering and feature transformation algorithms on both the forest fire and the wine quality datasets, and also run clustering algorithm on the data after feature transformations for both datasets. For Task 4-5, we will focus on the forest fire dataset and run neural network on some new data obtained from feature transformation and clustering results. In Assignment 1, we concluded from the neural network experiments that the neural network for the forest fire dataset works the best, considering both classification accuracy and time consumption, when there are 10 hidden layers with each layer containing 50 nodes. Thus, in Task 4 and 5, we will fix the structure of the neural network to be of 10 hidden layers each containing 50 nodes. For every experiment, we always run 5 trials for the same set of parameters.

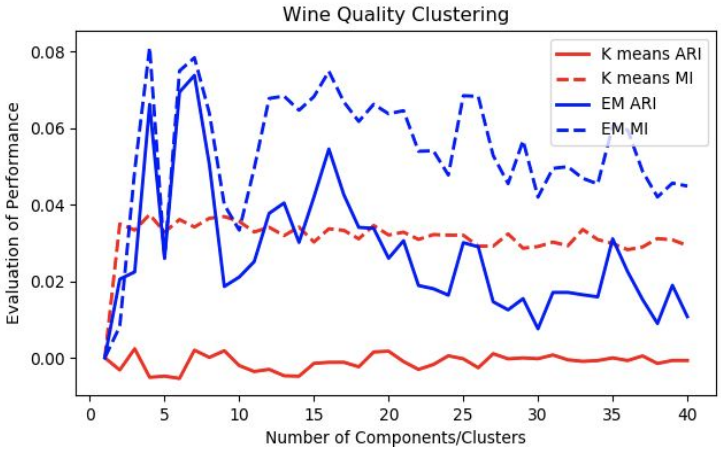
Task 1: Run the clustering algorithms on the datasets

1.1 Clustering vs. true labels

We run both K means and Expectation Maximization clustering algorithms on the datasets with varying the target number of clusters. To evaluate the clustering result, we use two different metrics: the adjusted rand index (ARI) and the adjusted mutual information (MI). The ARI compares similarity between two clusterings by computing all pairs of data points and counting pairs that are assigned in the same or different clusters. ARI is equal to 0 for random independently labeled data, and equal to 1 for identical clusters. The adjusted mutual information (short as MI in this write-up) is obtained from the mutual information between two clusterings with accounting for the fact the mutual information of two clusterings with big number

of clusters tends to be big, regardless of how much information they actually share. For each dataset and each clustering algorithm, we computed the ARI and MI scores between the clustering partition and the original label of the data. We plotted the results in the graphs below.





From the plot results, we obtain the following analysis:

(1) Expectation Maximization in general gives a clustering that is more similar to the true labels of the dataset, for both metrics, i.e. blue curves are in general above the red curves. Recall that EM is just a generalization of K means in which each data point might be considered to be in multiple clusters with certain probabilities. From the experiment result, we see that taking care of the “fractional” clusters does help improving the clustering result for both datasets.

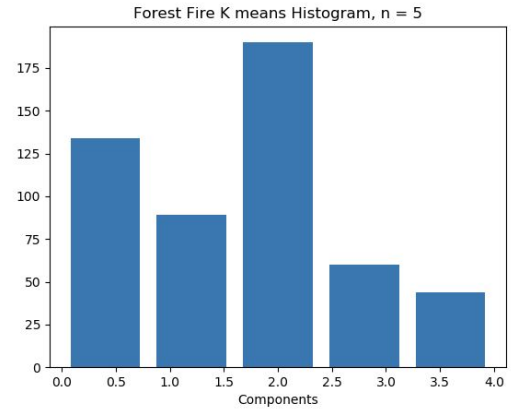
(2) In general, both evaluation results have more variance when the number of clusters are small and become more stable when the number of clusters gets better. This is because when the number of clusters is small, each cluster is more important over the whole dataset. Also there is more fluctuation in the forest fire dataset than the wine quality dataset. We believe this is because there are way more data points in the wine quality dataset than the forest fire dataset.

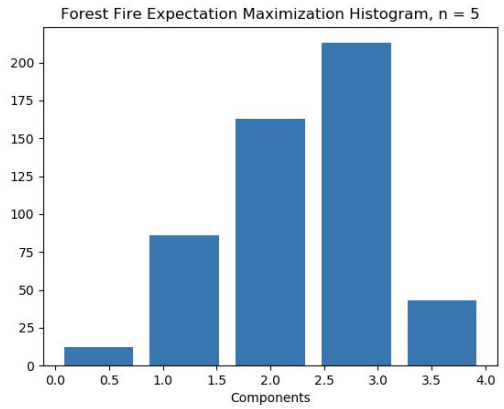
(3) When comparing the evaluation scores between the two datasets, we observe that the scores for the wine quality dataset is way higher than the forest fire dataset. We believe this is related to that the classification problem for forest fires is naturally just harder than the wine quality one, for that the two datasets have the same number of attributes but wine quality has way more data points thn forest fire.

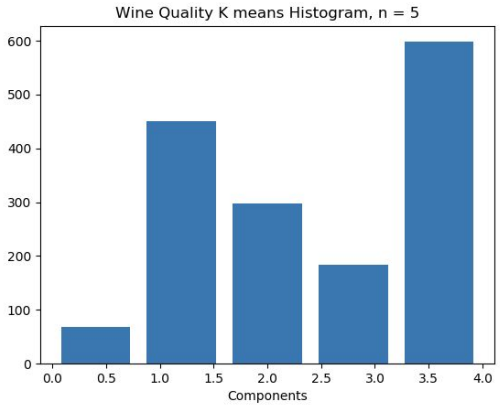
(4) Roughly speaking, the similarity metrics achieve the best when there are 5 clusters and stopped. We therefore will use 5 clusters in later experiments when the number of clusters is fixed.

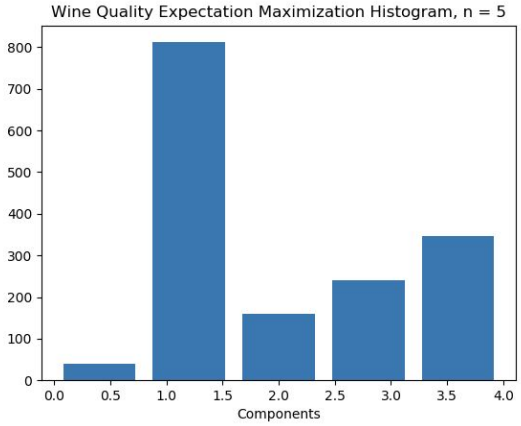
1.2 Histograms of the Clusterings

We also looked into the clusterings when there are 5 clusters, which gave the best similarity to the true labels according to the above. We then obtain the following four histograms, two for each dataset corresponding to K means and EM clusterings.







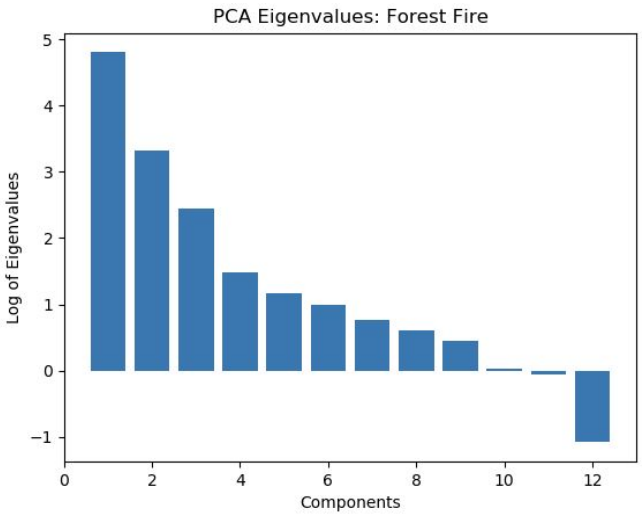


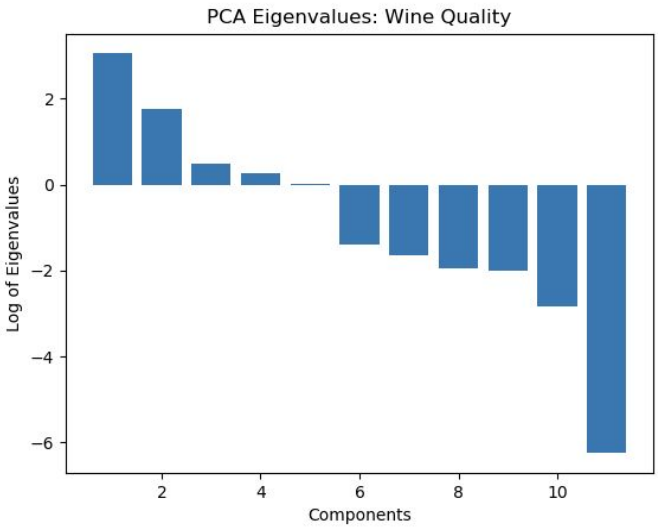
Even though it is still hard to visualize what the clusterings are really like, at least we can see some distribution of the data points over the clusterings from the histograms above. From the first part of this task, we know that EM produces clustering that is more similar to the true labels. Here we observe that for both datasets, Expectation Maximization algorithm gives a more skewed distribution. This shows that for our datasets, EM as a soft clustering algorithm is better at catching a skewed distribution of data points than K means.

Task 2: Apply the dimensionality reduction algorithms to the two datasets

2.1 PCA: plot eigenvalues

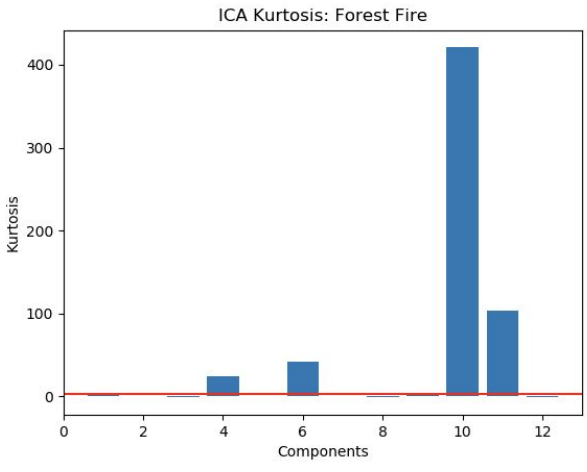
We first applied PCA on both datasets. To see how the PCA results are, we plotted the eigenvalues for the covariance matrices for both datasets.

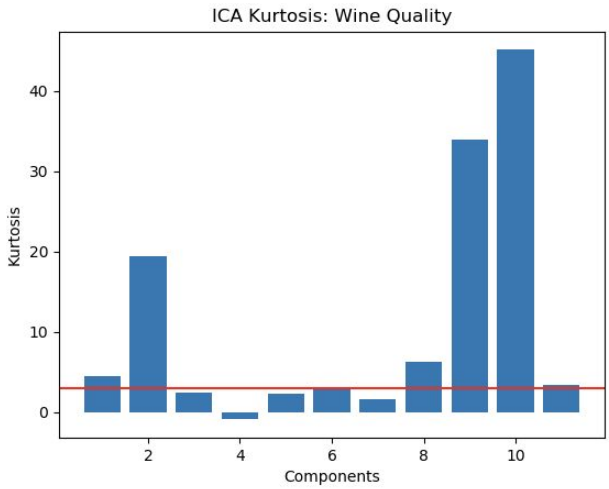




Since the eigenvalues of covariance matrices for both datasets decrease exponentially from the biggest to the lowest, we in fact plotted the natural log of the eigenvalues in the graphs. Looking at the two plots of eigenvalues, we clearly see that the forest fire dataset has way more big eigenvalues then the wine quality dataset. This implies that, in the d-dimensional space for each dataset where d = number of features, there exist a lot more independent vector directions, i.e. the eigenvectores, who have big variance for the projected data in the forest fire dataset than the wine quality dataset. Because of this, we should expect that applying PCA to the features in the forest fire dataset loses less information than applying PCA to the wine quality. In fact, it turns out PCA works really well comparing to other feature transformation algorithms we do in this project for the forest fire dataset. We are going to discuss more into this in Task 4.

2.2 ICA: plot Kurtosis

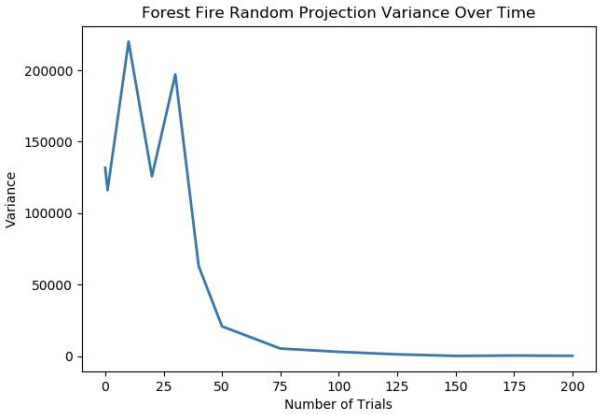


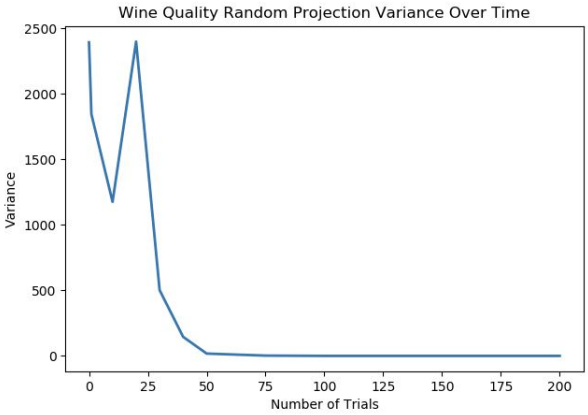


To see the ICA results, we plotted the kurtosis for the distribution for every new feature. We also drew a horizontal line at 3 which is the kurtosis for a Gaussian distribution. It turns out that in the forest fire dataset, there exists one ICA component that has kurtosis greater than 420, and

some other ones around 103, 42, and 24. This probably telling us these ICA components with high Kurtosis is probably related to some real independent features in the world, related to the forest fires. On the other hand, the highest kurtosis for an ICA component for the Wine Quality dataset is only about 42, way smaller than the precious 420. We concluded that ICA is a better feature transformation algorithm on the forest fire dataset than the wine quality dataset. We will see more details proving it in Task 3.

2.3 Random Projection (RP): plot variance over number of runs



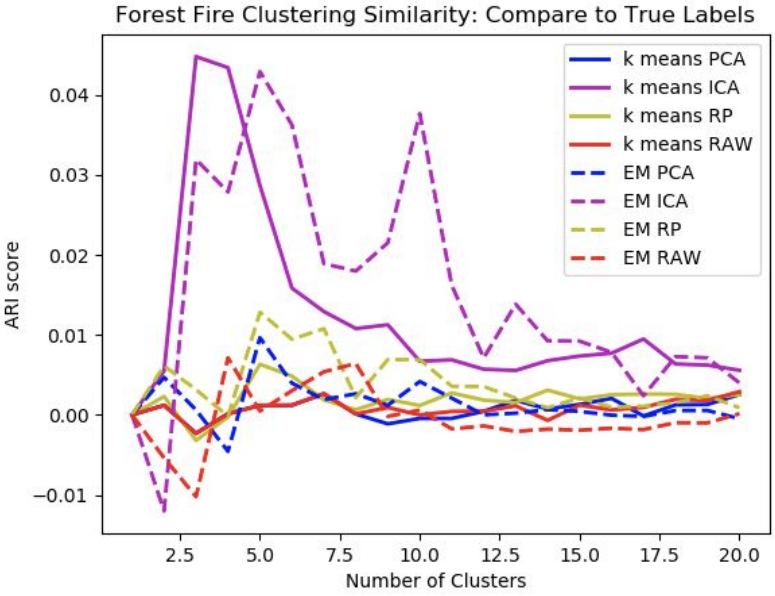


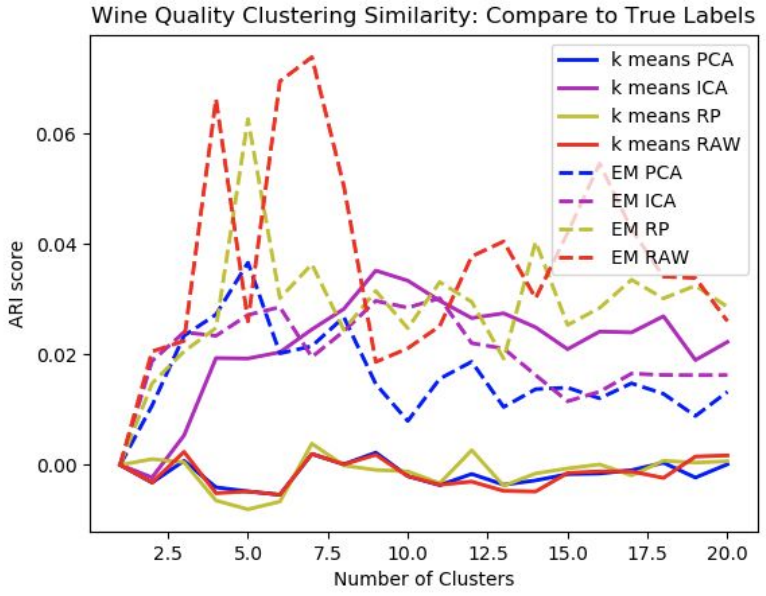
For Random Projection, we tried running it for multiple times and plotted the variance of the result features. As the graphs implies, we observe that the variances for the raw features for both datasets are very high. They both decay very quickly as the number of times of running Random Projection increases, with some fluctuation at the beginning and stay small once the number of runs exceeds 75. This shows that the data is going to become more like random noise when the number of runs is too big, which is no good for the purpose of classification or clustering. We will talk more about Random Projection to smaller dimension later.

Task 3: Reproduce clustering experiments on the data after feature transformation

To do Task 3, we chose parameters for each feature transformation that work well for the datasets, according to results from Task 2. Specifically, in PCA we picked the 6 components corresponding to the 6 highest eigenvalues of the covariance matrix for forest fire and 4 components for wine quality. For ICA, we picked the components that have kurtosis higher than 3 for each dataset. For random projection, we simply let it run for 10 times for each dataset. On these new features, we then apply both K means and Expectation Maximization. We compare the new clustering to both the true labels from the dataset and the clustering obtained from the raw features in Task 1, which are 3.1 and 3.2 below respectively. We use adjusted rand score (ARI) as the similarity metric between two clusterings (labelings) in this task.

3.1 Compare to the True Labeling





We first compare all the clusterings, from both raw features and new features, to the true labeling of the data, with the number of clusters varying between 1 and 20. We obtained the following observation and analysis:

(1) For both dataset and almost every feature transformation algorithm, EM gives a clustering more similar to the true labeling then the K means, i.e. for the same color the dotted curve is in general higher than the solid curve. This agrees with our result in Task 1, which can be explained by that soft clustering is better for both datasets.

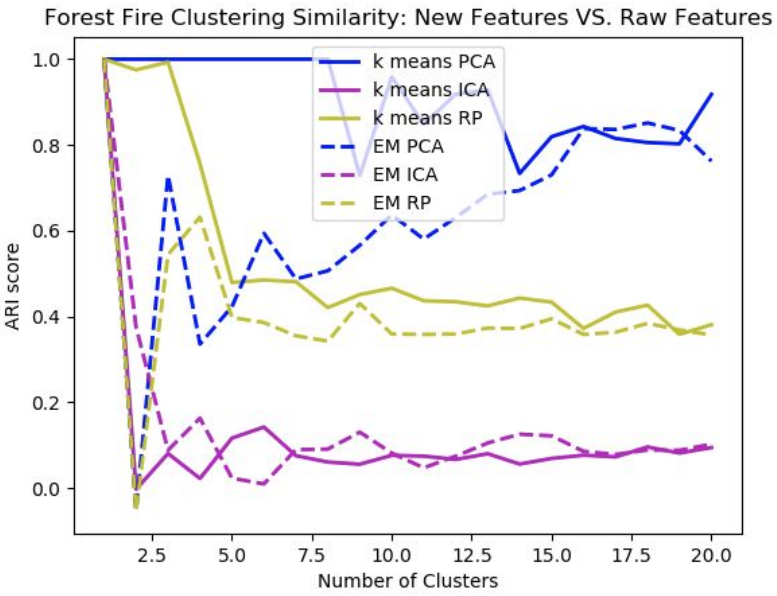
(2) In the forest fire dataset, ICA works much better than other feature transformation algorithms and the raw features. It is not true for wine quality though. We believe this is related to the fact that the forest fire dataset gets some ICA component with kurtosis as high as 420, which is about 10 times the highest kurtosis for wine quality. This also gives us more confidence to believe that the ICA components with high kurtosis are indeed capturing something meaningful.

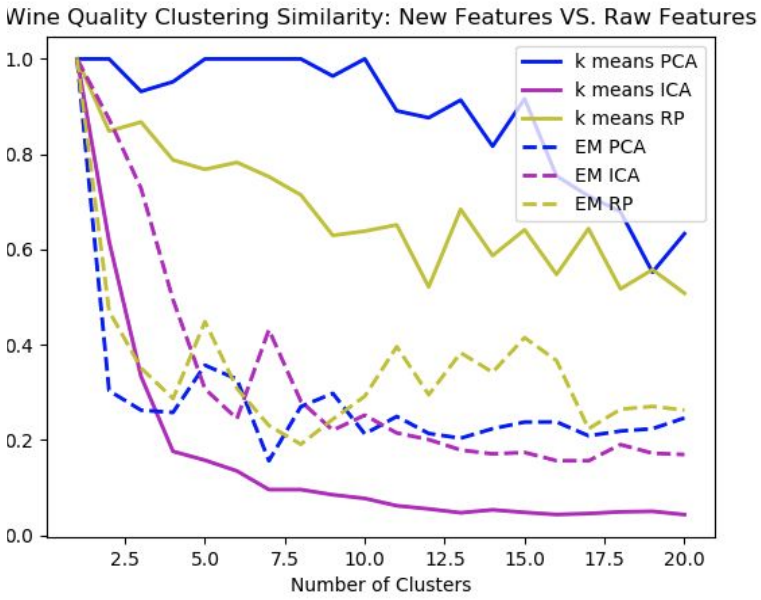
(3) In the forest fire dataset, clusterings obtained from PCA, Random Projection, and raw features are not so much different, in terms of similarity to the true labeling. Recall that the forest fire dataset has many high eigenvalues. Even though clustering from the PCA results does not seem much different here, we note that we will see PCA gives very good classification result through neural network in Task 4, way better than ICA. This shows us that having higher ARI score between the clustering and true labeling not necessarily imply better classification accuracy.

(4) In the right graph for the wine quality dataset, we see that the clustering that is the most similar to the true labeling, in terms of ARI, is obtained by EM on the raw features. This shows that for this wine quality dataset, none of PCA, ICA, or Random Projection can give a clustering that is a lot more similar to the true labeling. We believe this is also

partially because the clustering algorithms could already get good results on the raw data, as shown in Task 1.

3.2 Similarity between clusterings of raw data and clustering of new data (after feature transformation)





We also compared the new clustering results from the data after feature transformation and the clustering from the raw data. It turns out that

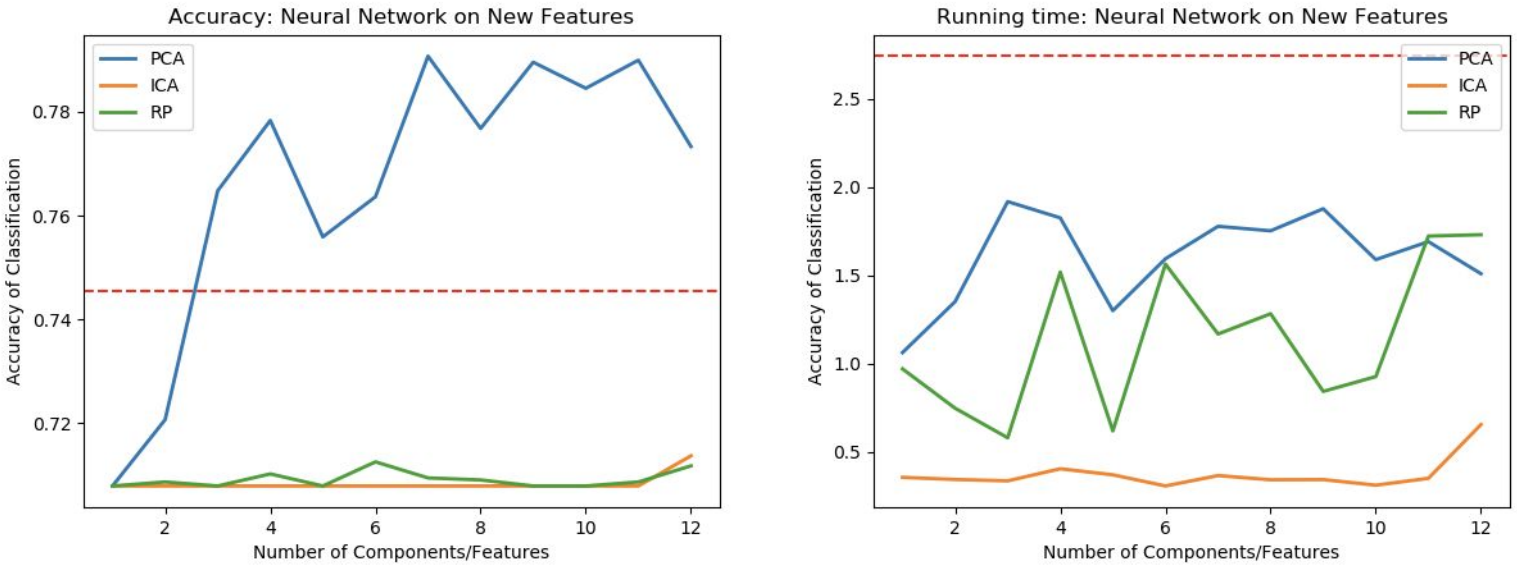
(1) For both datasets, PCA gives the most similar clustering results between new data and raw data, followed by Random Projection, and ICA gives the most different results.

(2) The big difference for ICA shows consistency with the fact that, the purple curves and the red curves are relatively further away from each other in the previous plots in 3.1. And we also talked about that in the case for forest fire, the big difference in fact leads to “something good,” meaning that the clustering from ICA results is more similar to the true labeling than the clustering from raw.

(3) Similarly, the small difference for PCA agrees with that the red curves (raw) and blue curves (PCA) in the plots in 3.1 are in general close to each other. In fact, when the number of clusters is less than 10, for both datasets, the raw K means clustering and the PCA K means clustering are identical for most cases, as one can see that blue solid lines in both graphs are as high as 1 when the number of clusters is small. This shows that PCA just does not change as much the K means clustering results.

Task 4: Run neural network on the new features

We now run neural network on the data after feature transformations for the forest fire dataset. Recall that in assignment 1, the structure of 10 hidden layers where each layer contains 50 nodes gives the best classification performance, around 0.745. When running neural network for the new data, we keep the same structure of network.



We run experiments on data obtained from different feature transformations and obtained the results for accuracy and running time above. We now present the analysis based on the graphs here:

(1) Very clearly, neural network on the PCA features gives the best classification accuracy, way better than both ICA and Random Projection, as well as the raw data. We believe this is because the covariance for the forest fire dataset has many very high eigenvalues. This not only means that we captured important information by transforming to the PCA components, but also that transforming to the PCA components does not really lose much information from the original dataset. On the other hand, ICA and Random Projection do not really improve the classification result for our forest fire dataset.

(2) In terms of running time, we see that neural network on PCA results is the slowest, followed by Random Project, and ICA is the fastest. This may make PCA not as efficient. However, note that the raw data (the rest dotted line on top) takes way more time for any of the three feature transformation methods. This shows that every feature transformation saves us about at least 30% time.

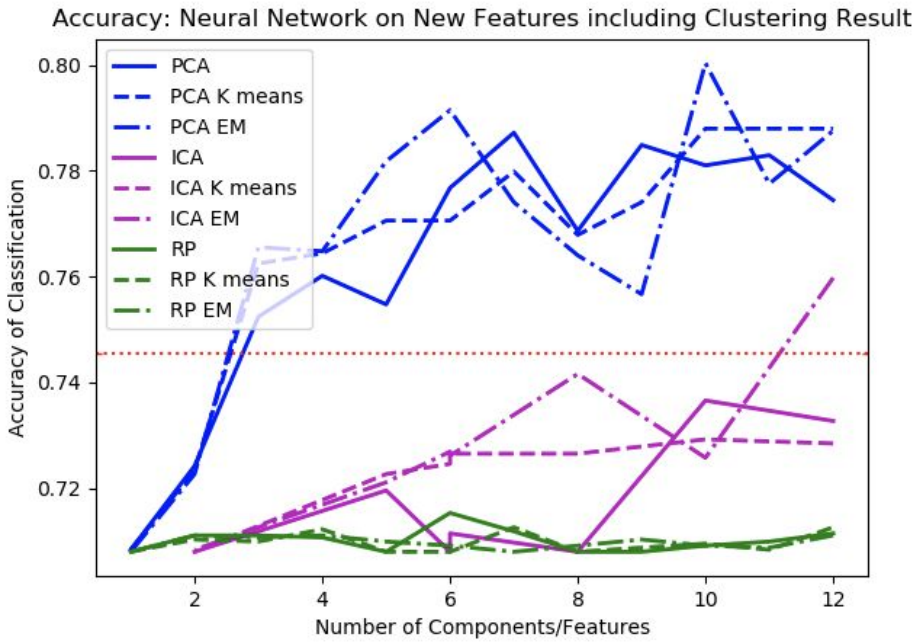
(3) In addition of (2), observe that the time curve for PCA in fact does not go up when the number of inputs/features to the neural network increases - it actually even goes down till the end. Note that with the same structure of the network, the running time is determined by both the number of inputs and how fast each algorithm converges. Thus, we believe that with more PCA components, the neural network converges faster, and it

is because the forest fire dataset's covariance matrix has so many very high eigenvalues.

(4) We conclude that, when using neural network to do classification, the PCA features give great performance in terms of both accuracy and running time.

Task 5: Treat the clustering result from Task 3 as an extra feature, rerun neural network

Now, we add clustering results on new features to the new data as an additional feature and rerun the neural network. We compare the new classification results for both K means and EM clustering with the classification result on the data obtained from the feature transformation algorithms directly.



We plot the accuracy result above, and present the following observation and analysis:

(1) Adding the clustering result to the features, for either K means or EM, does not change much of the performance, i.e. lines of the same color stay close to each other.

(2) The performance with the additional clustering feature seems to be better in some cases, especially for the ICA features. Even though the best performance on the ICA results with additional clustering feature is still far away the results from the PCA components, the clustering feature increases the accuracy the most for ICA components than PCA and Random Projection. We believe that this is related to the result in Task 3 where ICA features show more similarity to the true labeling of the dataset. Note that the additional clustering feature improves less on classification accuracy for PCA result, which is also related to that result in Task 3 that PCA does not really change the clustering of the data points.

(3) When the accuracy with additional feature is better, the EM tends to perform better than K means, which again agrees with our previous experiments and results.