Unsupervised Learning and Dimensionality Reduction

Author: Qi Li

I. Dataset Description

Classification problem 1: Marketing

The goal of the analysis is to determine how different factors affect whether the marketing campaign of a Portuguese banking institution is successful or not. If successful, the bank's term deposit will be subscribed and it will have more stable deposits for further investments and thus increases its revenue. Similar analysis could be applied to different marketing campaigns to reach their target customers more efficiently.

Classification problem 2: Wine

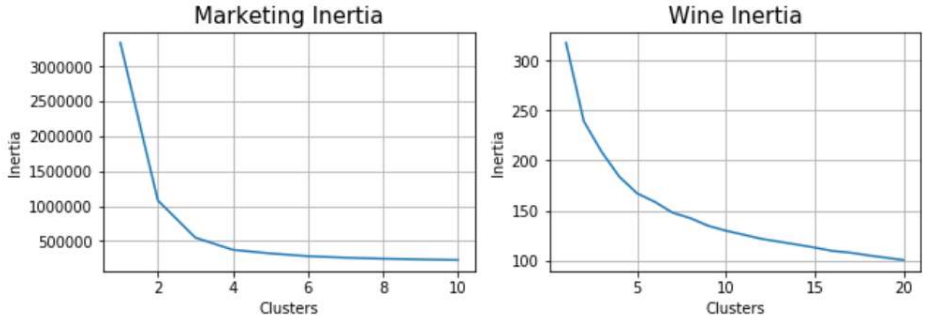
The goal is to analyze how physicochemical properties determine whether the wine is high-quality. It's interesting because for people who know little about wine, they can determine whether the wine is high-quality by checking the physicochemicals and decide whether to purchase the wine.

II. Clustering

1. K-means clustering

K should be selected depending on inertia based on different problem. Inertia measures how well a dataset was clustered by k-means and is calculated as sum of squared distances between each data point and the corresponding centroid. A good model should have low inertia and low number of clusters but usually there is tradeoff because inertia decreases when K increases.

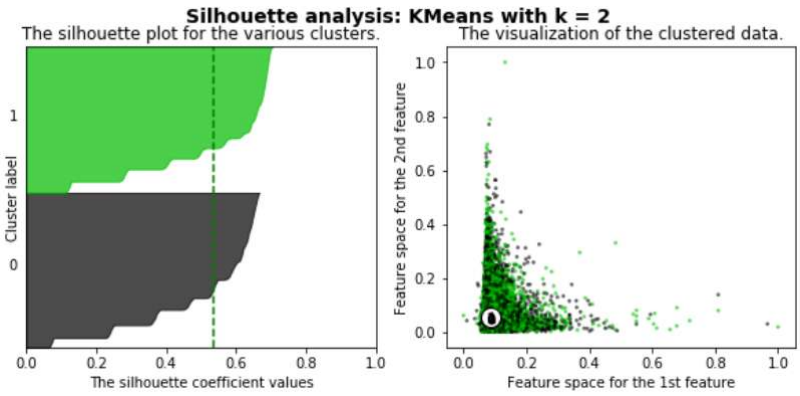
There are 2 methods to find the optimal K, the first one is Elbow method, which finds the point where the decrease in inertia begins to slow. Checking below inertia graph, best is k = 3 for marketing problem and k = 5 for wine problem. Since the data is labeled, we already know there should be 2 clusters for each dataset, and the Elbow method is not exactly matching, so I try another method so that it relies less on human judgement on selecting the best k.

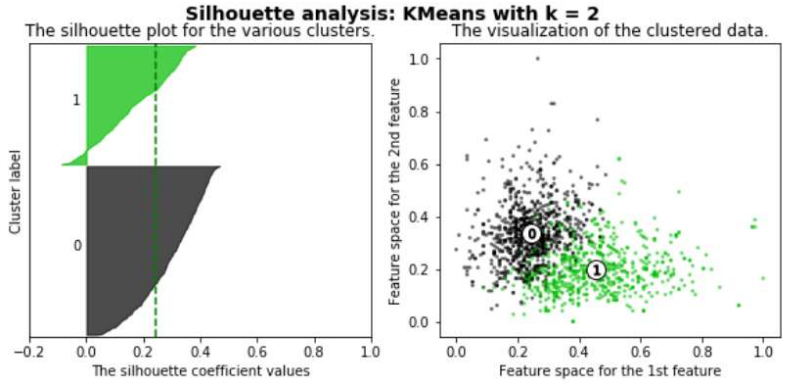


The second method is Silhouette Analysis, and it's used to study the separation distance between the resulted clusters. It displays how close each point in one cluster is to points in neighboring cluster.

(1) For the marketing data clusters, although k = 2 and k = 3 are both evenly split, and have a large portion of points inside each cluster with low silhouette values (<=0.5), we can validate internally and externally. Internally, k =2 returns a higher silhouette score, 0.54, indicating the sample has a larger mean distance between clusters. Externally, k = 2 returns a higher v-measure score of 0.0014, which is the harmonic mean of completeness and homogeneity. Therefore using Silhouette Analysis, k = 2 is a better choice.

(2) For the wine clusters, k = 2 has the highest silhouette score but as k increases, v-measure score increases majorly because homogeneity score increases a lot. This is because with more clusters, the clusters are more likely to only include points which are members of a single class.





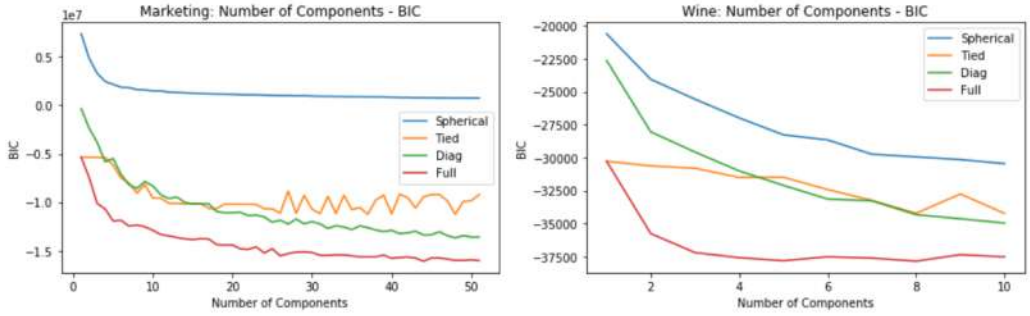
From the Silhouette Analysis, we can see that for marketing datasets, the number of data points in each cluster is similar but the centroids (for the 2 non-categorical attributes) are almost the same for k = 2 or 3 clusters. I have checked for all other attributes and have the same results because the dataset has 51 attributes and clustering depend more than 2 attributes. Also, the v-measure score, homogeneity score and completeness score are all very low. This is because out of 51 attributes, 48 attributes are categorical, and it doesn't make a lot of sense to calculate distance between binary variables and therefore could not produce results similar to the true labels. Kmeans is not a classification algorithm and does not work well for high-dimensional data, and therefore doesn't work well in this case compared to true labels.

For the wine dataset, whatever the k is, there is one cluster that has 1/3 more points than the others, that's because in the original data, low-quality wine has more instances than high-quality wine. Some of the points have negative silhouette coefficient, which means they are closer to points in another cluster than to points in its own cluster, and may indicate incorrect clustering. But the scatter plots on the 2 features make sense when k = 2, where lower right and upper left points are clustered differently. This dataset has higher homogeneity score and completeness score, because all of the variables it have are continuous and it makes sense to calculate the distance. But the scores are still low, indicating kmeans not working well for high-dimensional data (this dataset has 11 attributes), and also kmeans is sensitive to noise and wine dataset has some outliers where attributes like sulphates could be extremely high or low. Although the attributes are normalized, they still affect clustering.

2. Expectation maximization

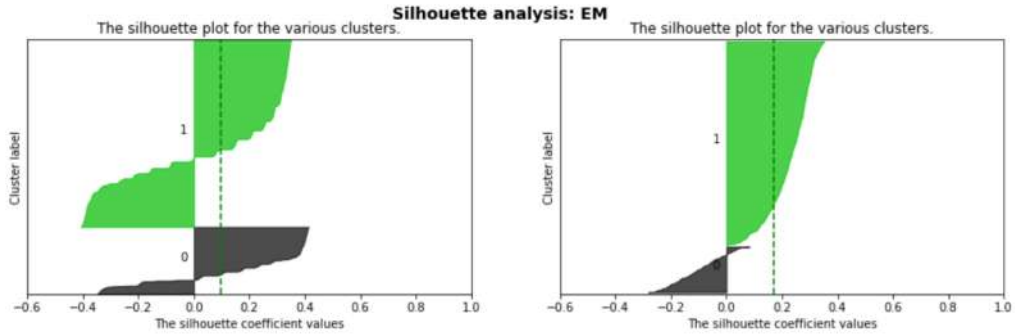
While K-means is hard assignment, expectation maximization is soft assignments because we are calculating the probability of a certain Gaussian creating a certain point. Here I use the Gaussian mixture model.

To determine the number of clusters, I use Bayesian information criterion(BIC) and also Silhouette analysis. For BIC, the lower the better, but although it penalizes for model complexity, usually the more clusters the lower the BIC so we can look at the gradient of the BIC. For both datasets, full variance has the lowest BIC. Cluster is 2 or 3 in marketing and cluster is 2 in wine.



For Silhouette analysis in marketing problem, I checked cluster 2 and 3, they all have similar v-measure scores around 0.03, but when cluster = 2 it gives the highest silhouette score, 0.10, and therefore cluster = 2. When there are 2 clusters, one of the clusters have 3/4 of the data points, and also 1/3 of all the points have negative Silhouette coefficients and therefore are closer to the other cluster than to the data points in their own cluster.

For wine problem, when cluster=2 Silhouette score is the highest. 3/4 of the data is in one cluster, and most of points in cluster 0 are having negative coefficients.

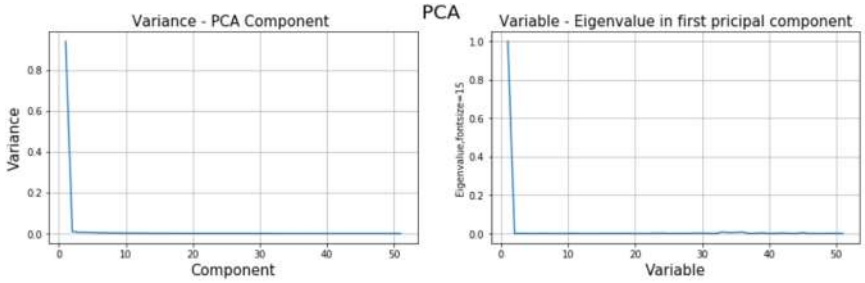


WHY???

III. Dimensionality Reduction

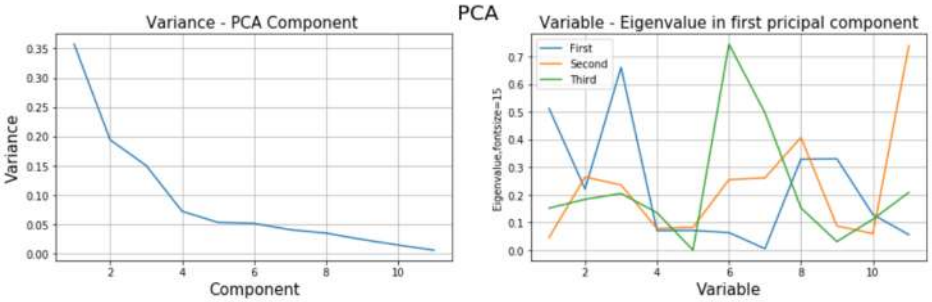
1. PCA

In marketing problem, first principal component explains 94% of the variance and in this component, variable 1 has dominant eigenvalue of 0.9998 and thus variable 1 is an important feature.



In wine problem, first, second and third component explain 70% of variance in total. Each component has

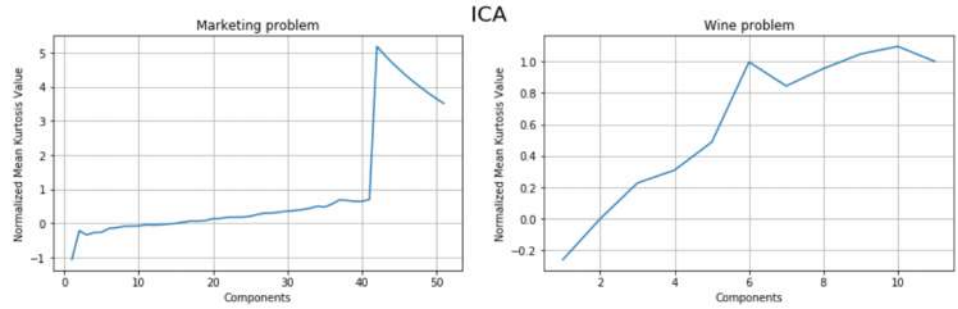
different important features: variable 1,3 are important for first component; variable 8, 11 are important for second component, variable 6, 7 are important for third component.



2. ICA

In marketing problem, to maximize kurtosis, 42 independent components are needed. It's reduced from 51 attributes.

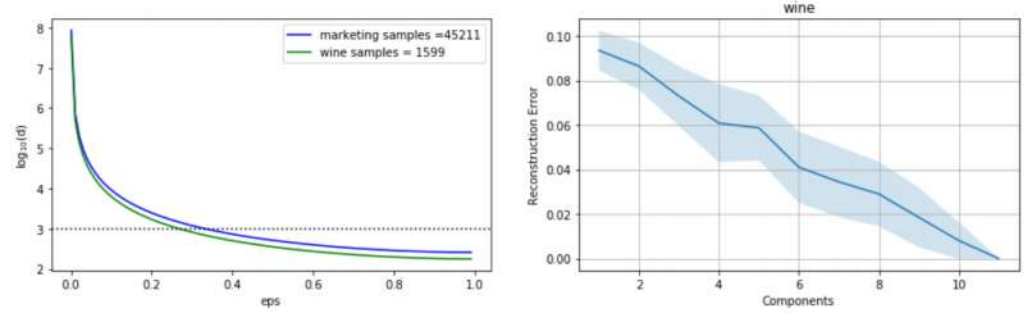
In wine problem, to maximize kurtosis, 10 components are needed. It's reduced from 11 attributes.



3. Randomized projection

The Johnson-Lindenstrauss lemma states that if the data points lie in a very high-dimensional space, then projecting such points on simple random directions preserves their pairwise distances. Also it determines the minimum dimension for a "safe" projection of higher-dimensional space to lower-dimensional space. This only relies on the number of samples and not relevant to number of features of original dataset. As shown below, as eps increases, dimension decreases, but even with eps = 1, dimension is higher than 100 for both datasets, much higher than their number of features. Eps is the parameter of Johnson-Lindenstrauss lemma that controls the number of dimensions so that the distortion in projected data is kept within a certain bound.

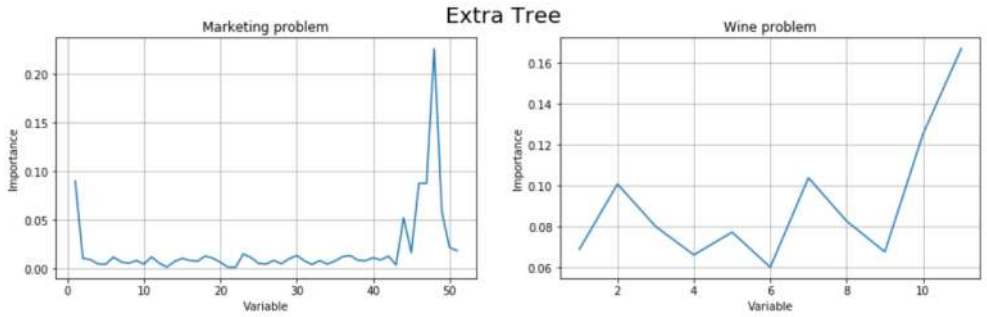
Reconstruction error and reconstruction variance reduce with number of components because more information is retained with increasing number of components.



4. Extra tree classifier

It consists of bagged decision trees and can be used to evaluate the importance of features. As below,

feature 1, 46, 47, 48 are important in marketing problem, feature 2, 7 ,10, 11 are important in wine problem.



IV. Clustering with dimension reduction

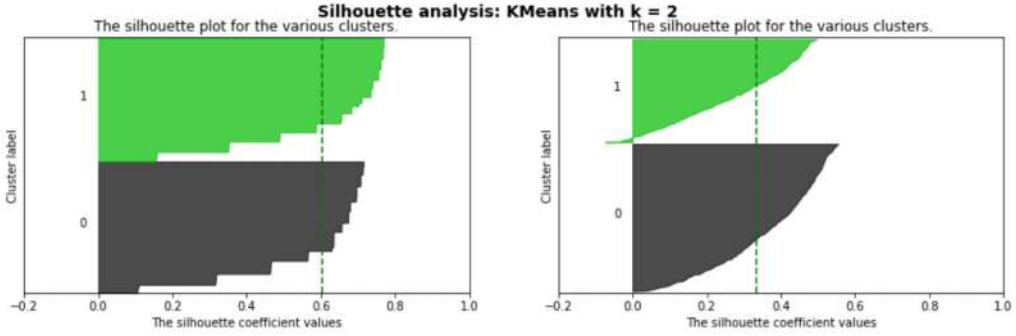
All graphs on the left are for marketing problem and ones on the right are for wine problem.

1. K-means

(1) PCA

Due to previous analysis, only first component of marketing problem is retained. Silhouette score increased to 0.6, and data is evenly distributed among clusters, but there is no improvement in v-measure. There are 3 components for wine problem, silhouette score improved from 0.24 to 0.33, data becomes more evenly distributed among clusters, and there are fewer data points having negative silhouette coefficient, but there is no improvement in v-measure.

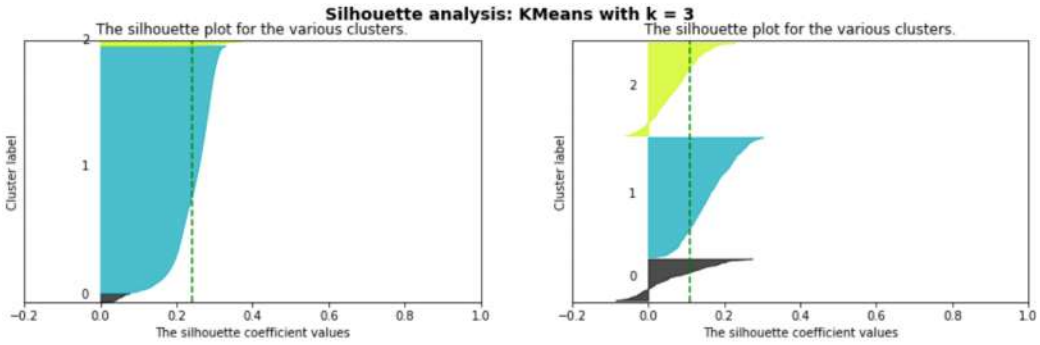
V-measure doesn't improve mainly because kmeans PCA does not provide information gain so v-measure is the same. Silhouette score improves because it has reduced dimensions.



(2) ICA

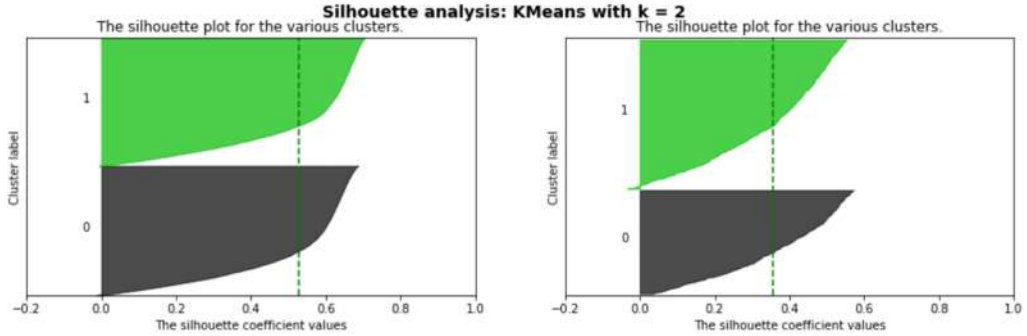
For marketing problem, there are 42 independent components and inertia keeps decreasing with clusters in a constant rate. But silhouette score is highest at 0.24 when cluster =3, much lower than without component analysis. Also one of the clusters have extremely large portion. Major reason is that most of attributes are categorical and cannot be whitened. Also data is not linear so doesn't make sense to use ICA which separates linearly mixed sources.

For wine problem, there are 10 independent components, and elbow point for inertia is 7. But silhouette score is highest at 0.11 when cluster = 3, lower than without component analysis. One of the cluster is a bit larger than the others. ICA makes sense in here since all features are linear.



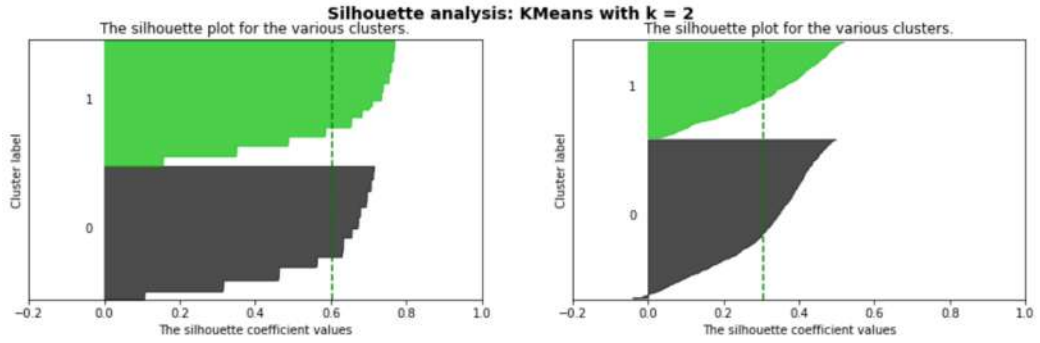
(3) RP

Although “safe” minimum dimension from Johnson-Lindenstrauss lemma are huge for both datasets, I try to use 3 components to analyze for 10 times and take their average to have a stable output. Silhouette scores improved for both datasets, cluster = 2 yield the highest silhouette score and they are evenly distributed.



(4) Extra Tree Classification

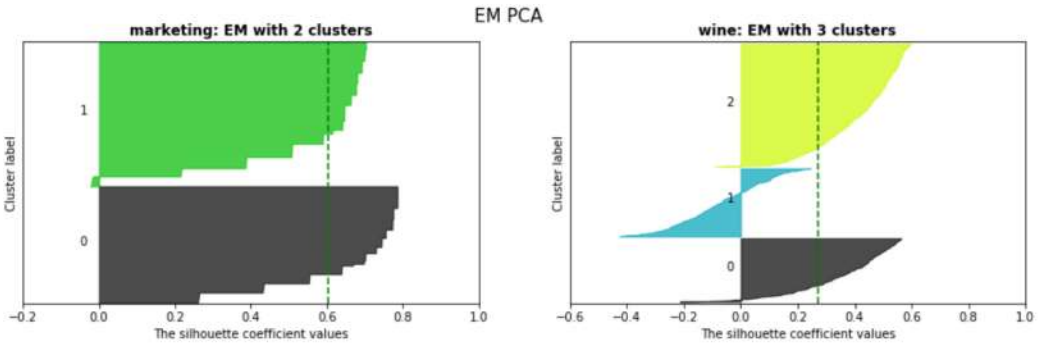
Using the most important features reduces run time for inertia and silhouette analysis. Silhouette score is improved, cluster = 2 has the highest silhouette scores and clusters are evenly distributed.



2. Expectation maximization

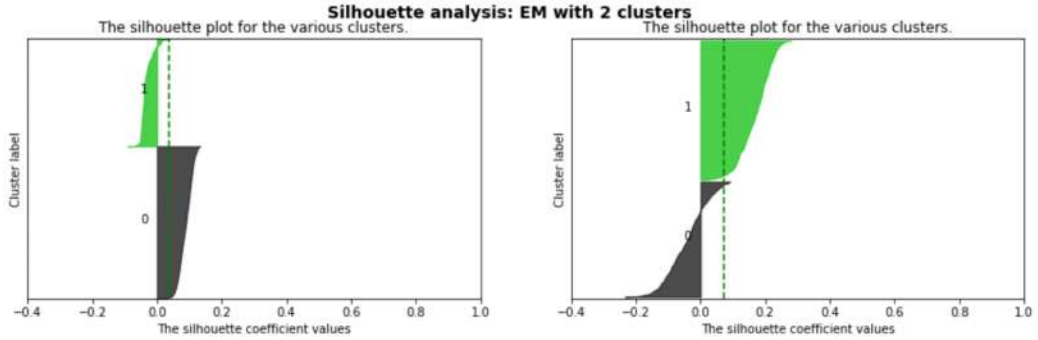
(1) PCA

For both datasets, the improvements are obvious: higher silhouette score, more evenly distributed in clusters, fewer data with negative silhouette coefficients, and data with negative silhouette coefficients has a lower scale. After using PCA in wine, the algorithm is managed to break down into 3 clusters so there is no dominating cluster. This is because it focuses on features that can yield the most information gain.



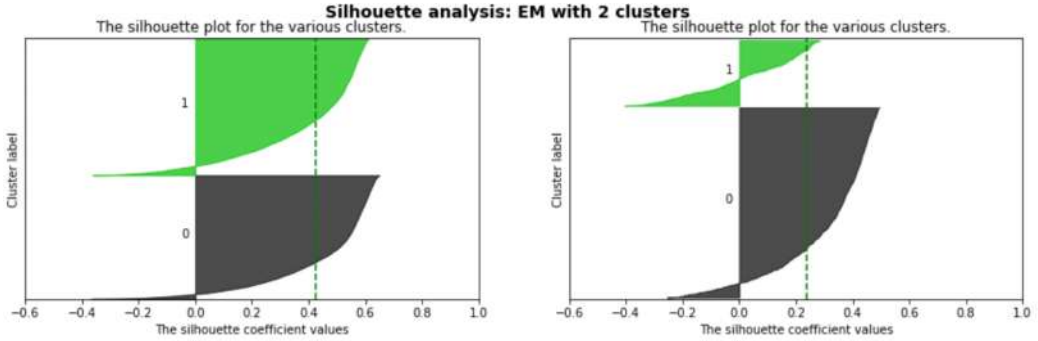
(2) ICA

For both datasets, lower silhouette score, unevenly distributed across clusters, some clusters have large portion of negative silhouette coefficients and they all have a low average silhouette coefficient than other algorithms. This doesn't yield meaningful results and doesn't seem to capture the fundamental piece of the datasets.



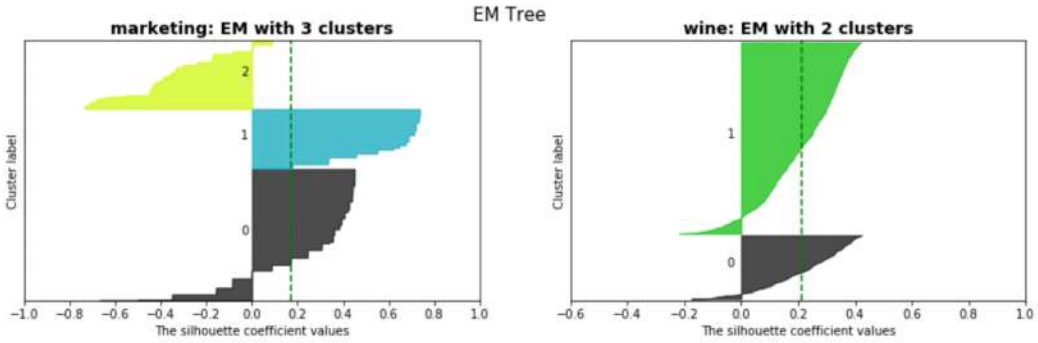
(3) RP

After running 10 times and taking the average, both datasets have higher silhouette score. Marketing problem has even distribution while wine data less even. They have only a small portion of data having negative silhouette coefficients.



(4) Extra tree classification

Although Silhouette scores are lower, both datasets have Higher v-measure scores. This is because the noise generated from other unimportant attributes have been removed and clustering will be more close to the true classifications.



3. Analysis

For marketing problem using kmeans clustering, the best dimensionality reduction algorithm is extra tree classification because PCA is not very useful in binary or categorical data, ICA does not yield meaningful results, and RP's results are not stable. For wine problem, PCA and extra tree classification are useful, because data is linear and they are able to reduce a lot of dimensionality. RP is not stable and ICA is not able to reduce dimension and categorize data correctly.

With expectation maximization, there are usually more data points that are closer to other centroids because the nature of this method is not based on distance but based on probability of distributions.

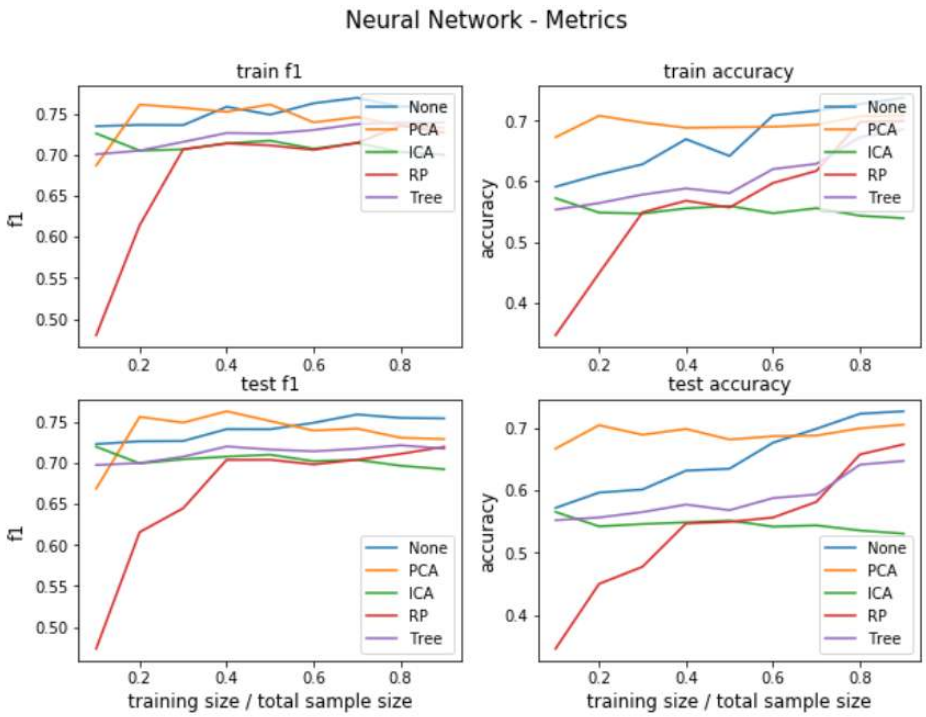
We can conclude from the results that different problems are suitable for different dimensionality algorithms:

1. PCA: When data is not categorical and has high dimensions.
2. ICA: When data is a linear mixture of independent variables.
3. RP: When the data set has huge amount of features so results show reduction in dimensionality and also stabilization.
4. Extra tree Classification: When there are a lot of features and most of the features create a lot of noise and do not contribute to the clustering results. Also it is useful when there are many complex and non-linear data types in the data set.

IV. Neural Network for Wine Problem

1. Dimensionality Reduction

After applying all algorithms, PCA has the best performance when training percentage is small because it's able to capture the most of information gain and can get rid of the noise from other not important variables. But when training percentage increases, it's not as good as original data because of information loss.



2. Clustering

I added the clustering feature to the original datasets. When training data percentage is small, expectation maximization is performing the best in terms of f1 and accuracy because this clustering feature is able to find the optimal parameters of the hidden distribution function from the given data and therefor provides more information. But when training percentage increases, this creates more noise and may be due to local maximization, it's not as good as not adding the feature.