CS7641 Machine Learning:
Assignment 3 Unsupervised Learning and Dimensionality Reduction

Judy Jungeun Cha
jcha64@gatech.edu

1 INTRODUCTION

The purpose of this report is to explore unsupervised learning algorithms. Clustering and dimensionality reduction algorithms are implemented to the same datasets from the assignment #1. Clustering algorithms are k-means clustering and Expectation Maximization (EM). Dimensionality reduction algorithms includes PCA, ICA, Randomized Projections (RP) and Recursive Feature Elimination (RFE). Next, we reproduce our clustering experiments by applying clustering algorithm to the new spaces generated by PCA, ICA, RP and RFE. Finally, we run neural network algorithms and compare with our earlier work from supervised learner.

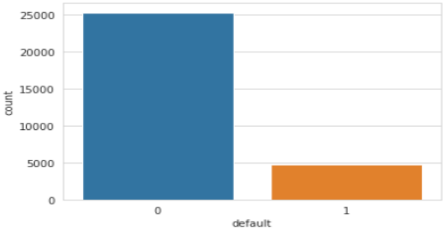
The Datasets

In the earlier assignment (#1), two datasets were Wine Quality (Wine) and Default of credit card clients (Credit).

For the Credit dataset, the goal is to identify whether default payment happens next month. With 11 different features and 30,000 of samples, there are possibilities that clustering/dimensional reduction can be effective to reduce overfitting with fewer features or/and reduce the training time compared to supervised learning. Noted 9 variables out of 11 features are categorical values (education level, marriage status, sex, history of past payment (1-6 month due). Only credit limit balance and age are numerical.

The goal of wine dataset is to identify which wine is white or red with 3 different features (alcohol, sugar level and pH) based on 6,497 samples.

Both datasets employed a binary variable, default payment next month (Yes = 1, No = 0) and wine type (white wine = 1, red wine = 0). Data cleaning and preprocessing were performed for both datasets, especially since credit data shows imbalance like below.



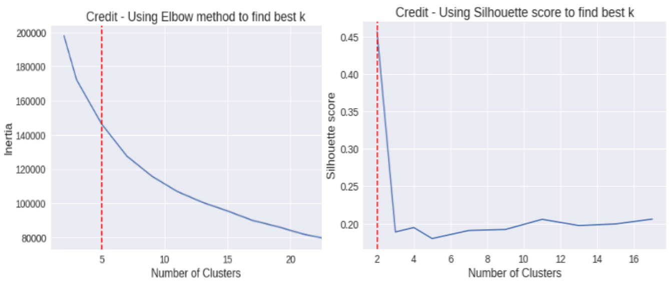
2 CLUSTERING

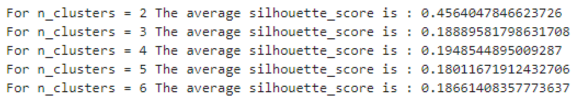
2.1 K-means clustering

K-means starts with random initial center and reinitialize with average of distances from data points to optimize the positions of centroid. K-means is one of the simplest and popular unsupervised algorithms, but we need to know the fixed number (K) of clusters beforehand which domain knowledge becomes important. To determine K, we look at inertia and silhouette score using by Elbow method.

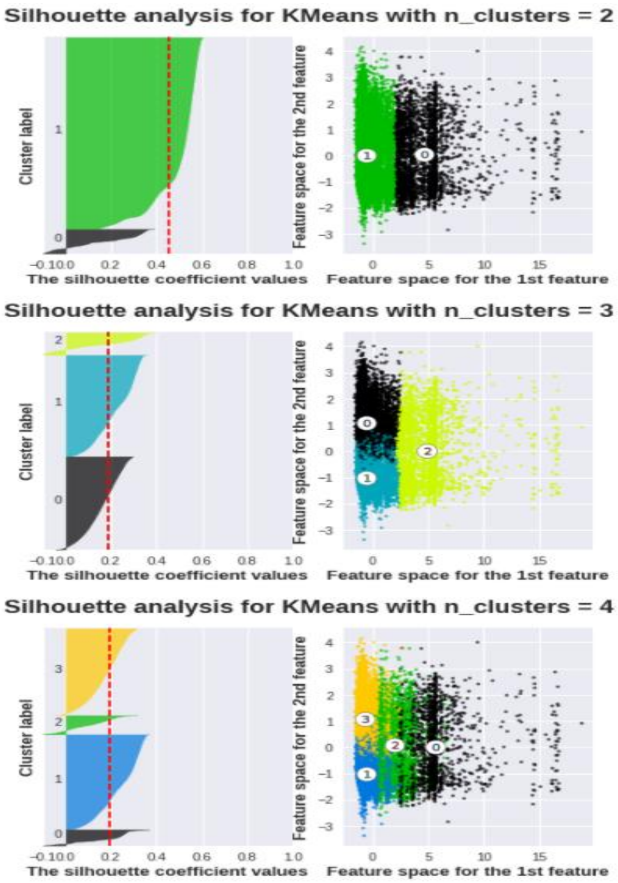
Inertia is also called mean squared distance (sum square error) between each instance and its closest centroid. Lower inertia (lower SSE) could be close to optimal number of clusters. Silhouette score is coefficient to measure consistency within clusters of datasets with ranges from -1 to 1. Therefore, higher silhouette score would be desired.

[Credit]



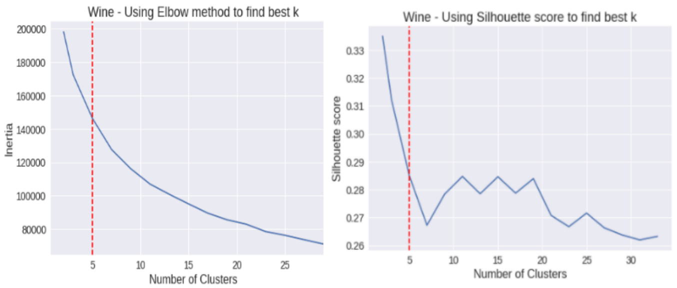


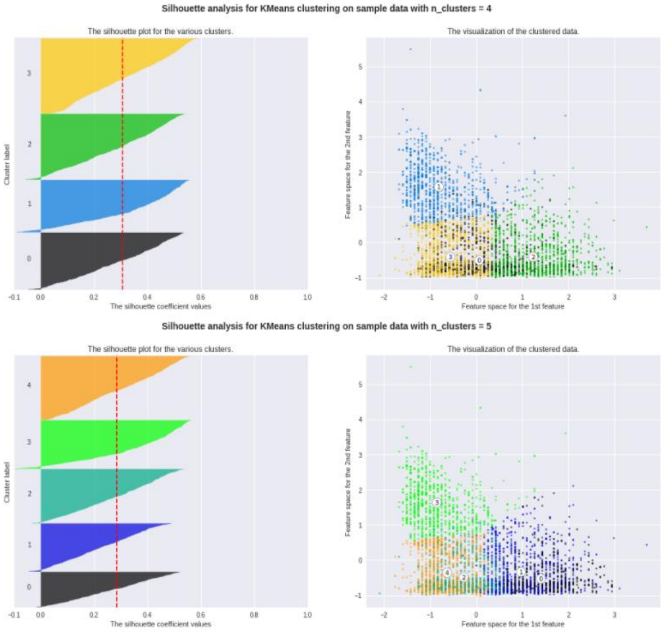
```
For n_clusters = 2 The average silhouette_score is : 0.4564047846623726
For n_clusters = 3 The average silhouette_score is : 0.18889581798631708
For n_clusters = 4 The average silhouette_score is : 0.1948544895009287
For n_clusters = 5 The average silhouette_score is : 0.18011671912432706
For n_clusters = 6 The average silhouette_score is : 0.18661408357773637
```



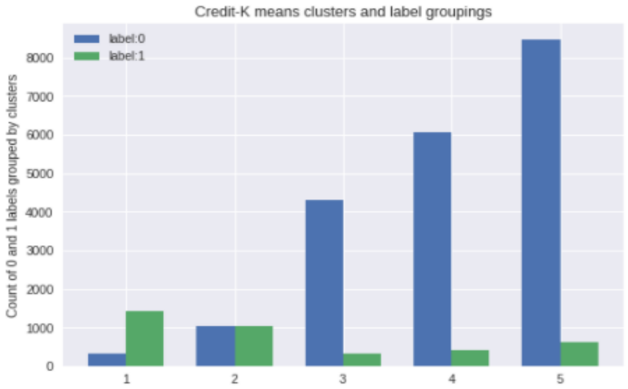
For credit, between clusters of 4 would be best based on elbow method. The average silhouette score is the highest at clusters of 2, but inertia is the highest as well at clusters of 2. Given lower inertia and higher silhouette is desired, clusters of 4 looks optimal for credit data.

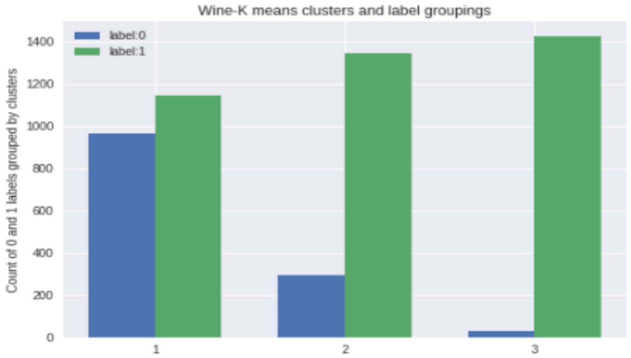
[Wine]





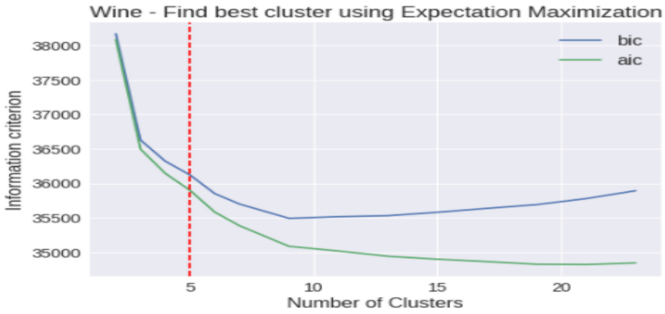
For the wine data, clusters of 4 and 5 shows more or less of the similar thickness and sizes and all the points of both clusters are bigger than average silhouette score. To balance out with inertia score, cluster of 5 would be best for wine data.

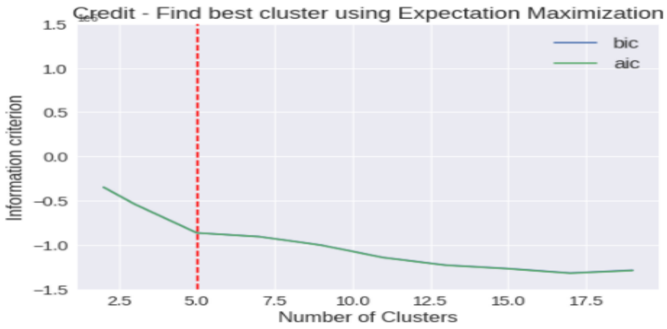




When I look at k-Means clusters and label groupings for each dataset, it seems the clusters line up with the labels. Especially wine data shows when the cluster has label 1 (white wine) the most then it has least of label 0 (red wine). It is because 3 clusters match with the number of features (alcohol, sugar level, pH). It seems when the number of clusters is closer the number of features, it has more chance clusters line up with the labels.

2.2 Expectation Maximization (EM)





The second cluster algorithm is Expectation Maximization (EM). EM is iterative method for finding maximum likelihood. Two popular criteria of EM are AIC and BIC, part of Information Criterion (IC) as below.

Akaike IC (AIC) = 2k -2l: Lower AIC via higher log likelihood or less parameters

Bayesian IC (BIC) = ln(n)k − 2l: Lower BIC via higher log likelihood or less parameters or less samples used in fitting

[log likelihood (l), # of parameters (k), # of samples used for fitting (n)]

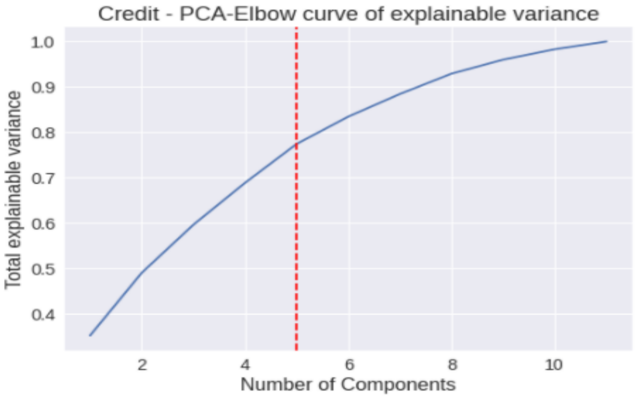
AIC and BIC indicate the amount of information lost and it should be minimizing. When k increases, # of parameters increase which means our model gets complicated(penalty). Log likelihood increases when our model explains the data better (reward). It helps us to choose balanced models between complexity and goodness of fit.

For credit dataset, cluster of 5 would be best as AIC/BIC decreasing rate slows down from 5. Noted that AIC/BIC are so close, graph shows only one line. For wine, cluster of 5 is shown as best considering silhouette score as well. The optimal number of clusters of EM seems in line with the results of K-means. It makes sense to have similar results from two different cluster algorithms.

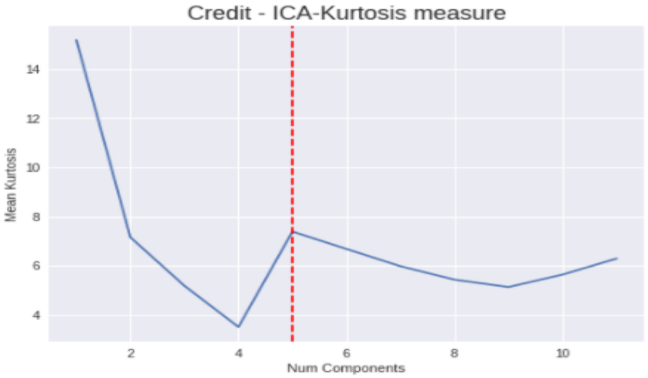
3 DIMENSIONALITY REDUCTION ALGORITHMS

When features increase linearly, amount of data we need to handle grows exponentially. This curse of dimensionality can be solved by dimensionality reduction algorithms. In this report, Principal Components Analysis (PCA), Independent Components Analysis (ICA), Randomized Components Analysis (RP) and Recursive Feature Elimination (RFE) will be discussed as a part of dimensionality reduction algorithms.

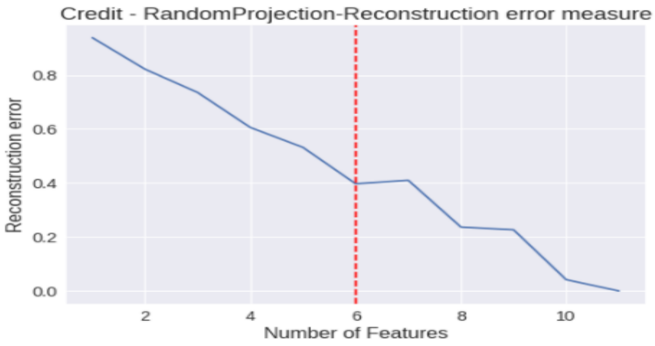
[Credit]



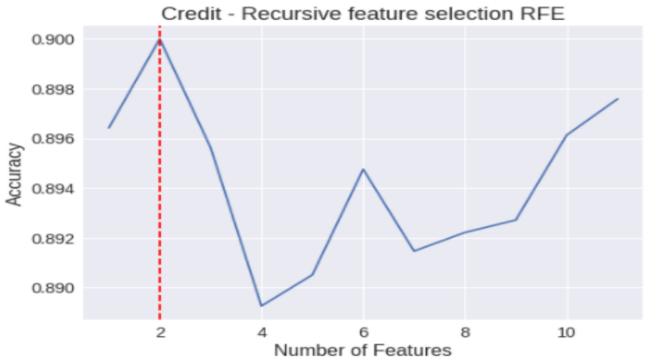
PCA reduces the dimensionality of large data with a little accuracy sacrifice. The smaller data set is easier to process and visualize and PCA is very fast compared to other algorithms. For PCA, total explainable variance is used to select the best number of features. Total explainable variance is cumulative sum of variances of all individual principal components (eigenvalues). PCA looks for maximizing variance and higher explainable variance is desired. A threshold of the total variance was set between 70 – 80% which would be components of 5 in credit dataset.



For ICA, kurtosis is used to determine the best number of features. ICA gives very good results based on assumptions that source is statistically independent, and its value has non-Gaussian distribution, which means higher kurtosis. The components of 5 have higher kurtosis. PCA and ICA shows the similar results.

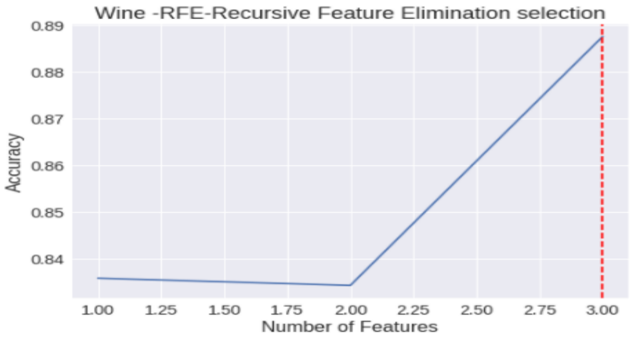


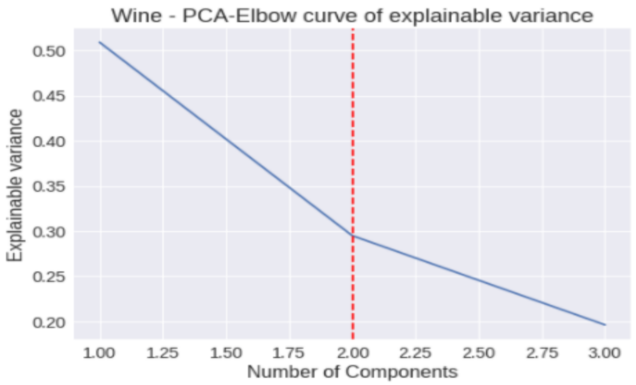
For RP, reconstruction error is calculated by gaussian random projection. Essentially RP is very similar to PCA. But PCA calculates projection matrix with eigenvector and RP uses randomly chosen vector. For credit data, 6 features would be best with low reconstruction error and overfitting consideration. Noted the graph is changed every run due to the random projection.

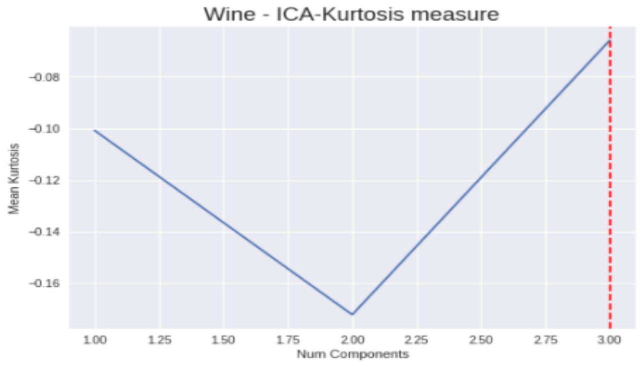


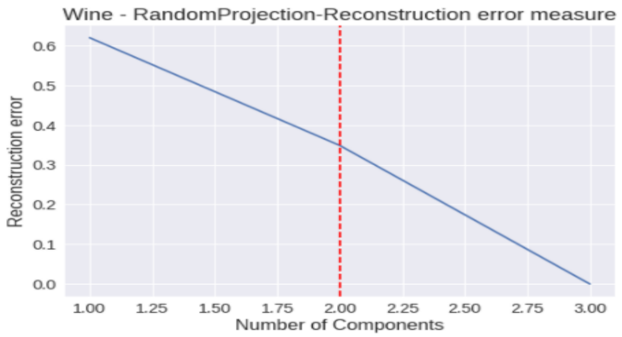
RFE selects features by recursively pruning the least important features until it reaches the desired number of features. For credit dataset, it already reached at the highest accuracy (90%) with 2 features.

[Wine]









For PCA 2 features would be the best and for ICA 3 features are the best with the highest kurtosis. Since wine data has only 3 features (alcohol, pH, sugar level), selection of the number of components for dimensionality reduction algorithms are limited compared to credit dataset. But PCA shows the reduction to 2 features with higher explainable variance from 3 variables. ICA, RFE do not give a reduction of data.

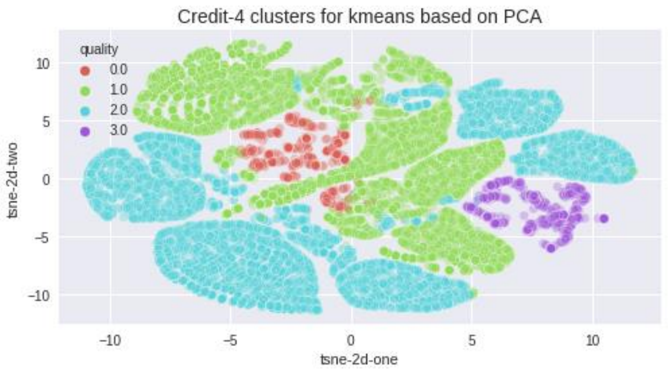
4 CLUSTERING AFTER DIMENSIONALITY REDUCTION (DR)

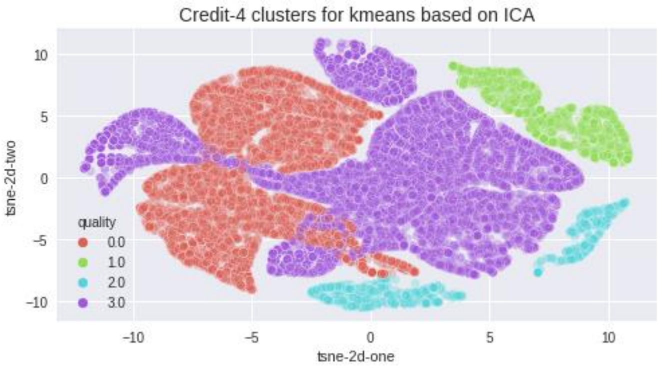
First dimensionality reduction applies to the dataset and performs clustering algorithms using k-means and EM. The best number of clusters of K-means and EM before DR were selected for comparison purpose.

[Credit]

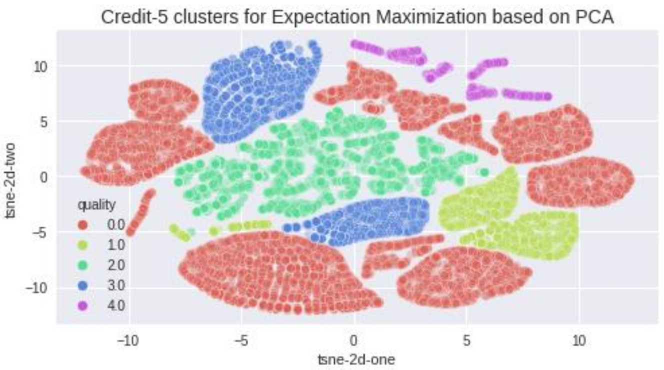
Credit Benchmark	k=4	k=5	
	K-Means	EM	
	silhouette	AIC	BIC
Original	0.1949	-864,999	-864,077
PCA-based	0.2018	282,273	283,033
ICA-based	0.2958	-92,967	-92,744
RP-based	0.1969	414,701	415,381
RFE-based	0.1948	-837,807	-836,966

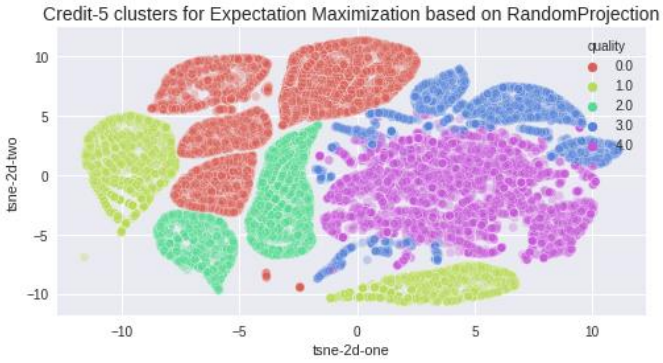
For the k-means method, when using reduced data after DR for clustering algorithms, ICA performs the best at k=4. After ICA DR, data's silhouette score gets bigger. Silhouette score is coefficient to measure consistency within clusters of datasets. The below graph show ICA has tighter form of clusters compared to others. It seems all DR algorithms has similar silhouette score with original (before DR) except for ICA which performs better. It seems we can reach the same conclusion with smaller data after DR.





Using EM, any original and RFE performs better than other algorithms at k=5 based on AIC and BIC score. RP produced the highest AIC/BIC score which performs worst at k=5.

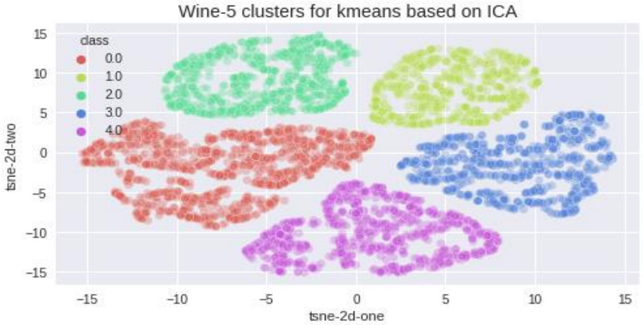


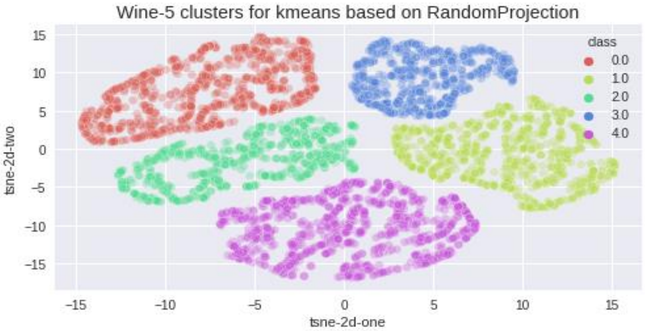


[Wine]

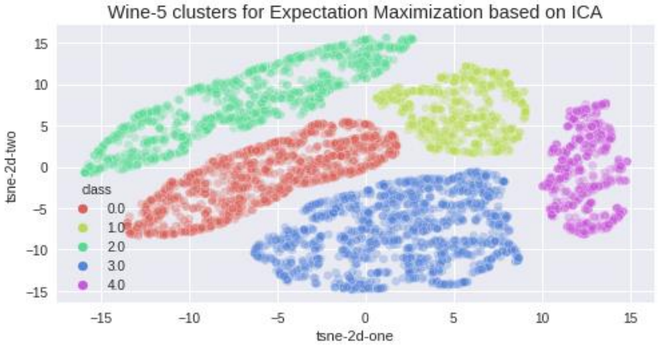
Wine Benchmark	k=5	k=5	
	K-Means	EM	
	silhouette	AIC	BIC
Original	0.2851	35,898	36,121
PCA-based	0.2851	35,620	35,941
ICA-based	0.2758	-92,967	-92,744
RP-based	0.1378	25,277	25,467
RFE-based	0.2888	35,898	36,121

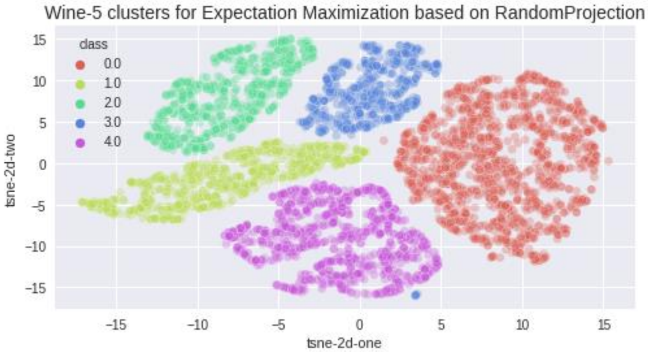
For wine data, k-Means silhouette score is all in line with original except for RP when k=5.



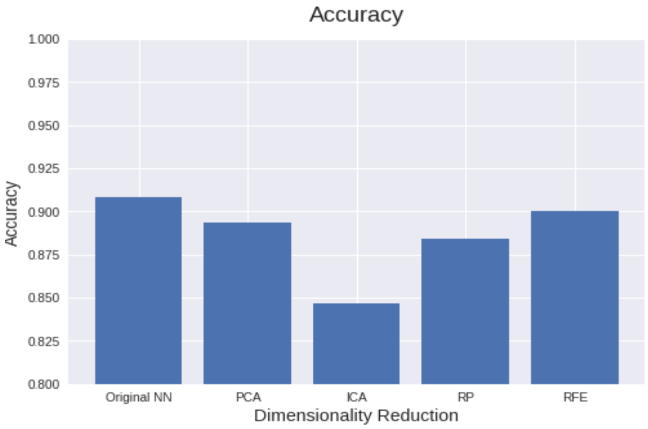


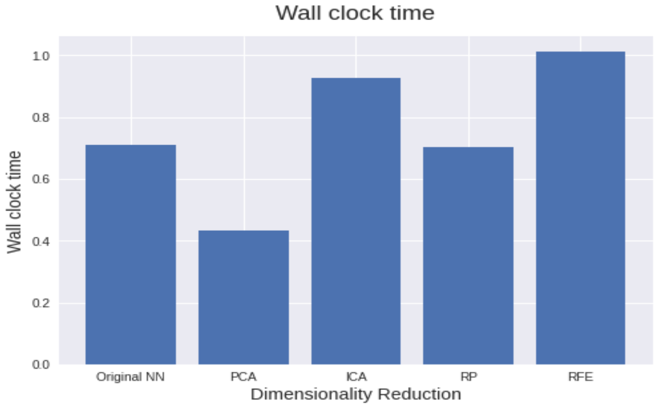
It is clear ICA performs the best for wine data when k=5. It has similar silhouette score with others and EM algorithms score is way lower than any other algorithms. Clusters plot is not clear as much as credit dataset because wine data has 1/5 number of samples compared to credit.





5 NEURAL NETWORKS ON CREDIT DATA AFTER DIMENSIONALITY REDUCTION



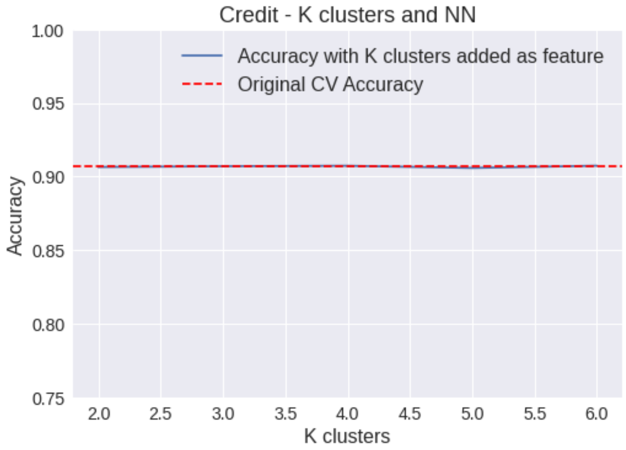


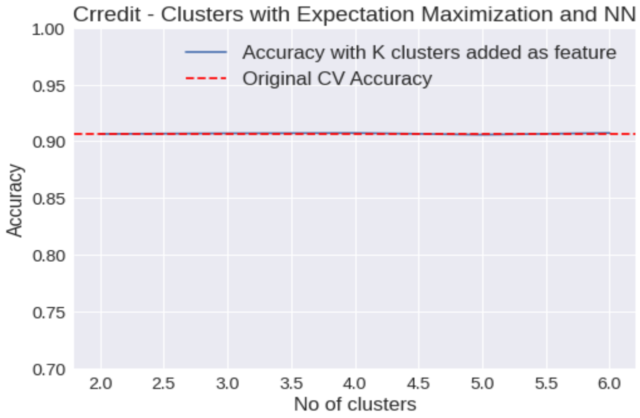
Neural network algorithms were performed on the new spaces of credit dataset created by PCA, ICA, RP and RFE. In terms of accuracy with cross validation, Original NN has the highest accuracy of 90%. This is expected because dimensionality reduction algorithms lose some data from original. But PCA, RP and RFE still show the high accuracy of 88-89% even though information reduction. This means data would have enough meaningful features even after DR.

ICA has the lowest accuracy of 85%. It might be credit dataset is not mutually independent among features (e.g., credit limit can have high correlation with marriage status, history of past payment etc.).

For wall clock time, PCA is the fastest and original and RP are next. ICA and RFE might take longer to converge because of some lost features during dimensionality reduction.

6 NEURAL NETWORKS ON CREDIT DATA AFTER CLUSTERING





Two clustering algorithms are applied to credit data set just like dimensionality reduction algorithms and the clusters were treated like new features and we rerun neural learner. For neural network leaner, a grid search with five-fold cross-validation was performed and params were 12 max iteration and hidden layer of 8. Both k-means and

EM shows high accuracy of 90% like original. For training time, original takes 0.5 seconds that is faster than time with algorithms (k-means:8.37 sec and EM:15.85 sec). Here is again due to information loss it takes longer to converge compared to original without any algorithms.