CS7641 – Machine Learning

Assignment 3

Jingyao Zhu

jzhu398@gatech.edu

1 Overview

In this report, two different unsupervised clustering algorithms including K-means and expectation maximization and four dimensionality reduction algorithms such as PCA, ICA, Randomized Projections and truncated SVD have been applied to two datasets to solve a real-world problem.

I. Wine Quality Dataset

(http://archive.ics.uci.edu/ml/datasets/Wine+Quality)

The first dataset is available which is from UCI machine learning repository about white wine quality data. It has 4,898 varieties overall. There are 12 different properties of the wines, only one of which is about Quality and the rest are on chemical properties of the wines including density, acidity, alcohol content, etc. All features will be the inputs to make the cluster.

Based on many articles mentioned, the worldwide distribution of wine is 31 million tonnes. It is giant. But there is no fundamental way to tell which metrics are a kind of good quality of the wine. How can you differentiate the wine? The big question arises. Many experts think that its smell, flavor, and color are the keys. Is it true? Here the use of clustering comes. If wine features have relationships with its quality, the clustering may help us to find which is good wine because it will group different qualities based on node location.

II. Bike sharing demand

(https://www.kaggle.com/c/bike-sharing-demand/overview)

The second dataset is about bike-sharing systems. As many articles mentioned, automated renting bicycles could make human life so different in daily life throughout a city. But this system could have many processes such as obtaining membership, rental, and bike return. Rent a bike from one location and return it to another location. There are 12 features including duration of travel, departure location, arrival location, etc. The dataset could be used to forecast the frequency of rental bikes (high or low) in the Capital Bikeshare program in Washington, D.C.

This dataset has been used in HW1 and HW2. For the bike-sharing demand dataset, the original problem is to use the current dataset corresponding with weather information to predict demand expected as a regression problem. But it is hard to predict a specific number or range in real life because of the high error rate. Then according to the available cutoff online, I decide to change the scope of the problem from predicting demands to predicting high or low demands. Then the problem changes to a binary classification problem.

It can help companies who offer sharing bikes to predict the impact of marketing in different locations according to reliable predictions. Also, the ML models can help bike-sharing companies focus on different market strategies to avoid malicious competition. Moreover, companies may be able to launch differentiable services based on different demands and lower the cost to maximize the profit margin.

2 Analysis of Wine Quality Data

i. K-means clustering

1. Explanation of the method

K means is one of the clustering algorithms that aim to group different sets of objects in such a way that objects share the same characters. The clustering algorithm is an unsupervised method in which no label is available for samples. K-means splits wine into different clusters based on the features available. In this case, inertia or within-cluster sum-of-squares criterion is applied to K means algorithm. Based on the calculation, minimum inertia is the measurement of whether K means is doing well. But it is based on a good assumption which is clusters are convex (maybe not always true). In Figure 1, inertia is smaller and smaller with the increasing number of clusters. Generally speaking, the lower value is, the better algorithm performs. Zero is always optimal. Besides that, there are many ways to help measure whether clusters are good or not. In Figure 2, there are four criteria to help pick the best K. When the Silhouette Coefficient scores are high, the k means model may provide better-defined clusters. From the graph, clusters may have more and more overlapped information when the number of K grows. It suggests picking K less than 7. Homogeneity shows that each cluster has members only from one single class and completeness show that the same cluster includes all members of a given class. The larger, the better. When taking them together, the suggestion would be K less than 7. NMI tells that the mutual information among clusters does not go down with the number of K grows.

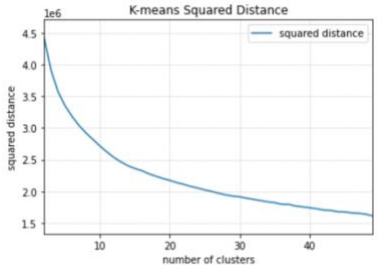


Figure 1: k means distance

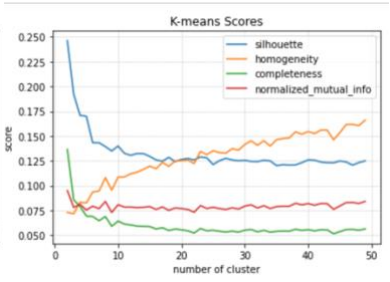


Figure 2: k means measurements

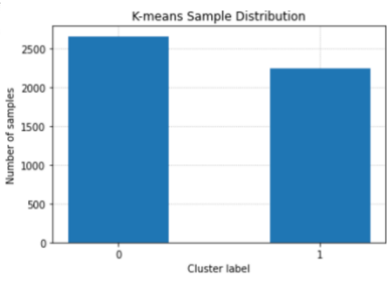


Figure 3: k = 2 cluster

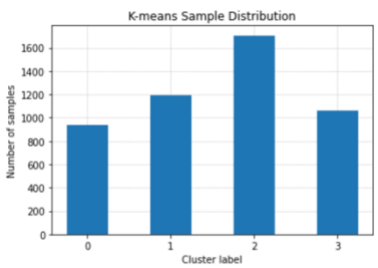


Figure 4: k = 4 cluster

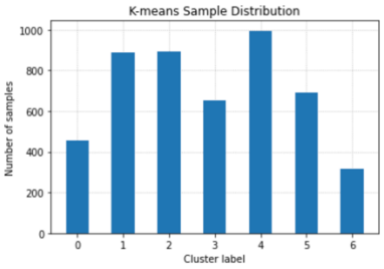


Figure 5: k = 7 cluster

2. Cluster description and results

Overall, 4 clusters will be used in K means because the K = 4 performance in four criteria is the best and inertia score is okay. But I still graphed k = 2, 4 and 7 because k = 2 means "good" and "bad" in reality and k = 4 is the best performance and k = 7 is the max k. Figure 3, 4 and 5, shows that k = 2, 4 and 7 cluster number and its distribution of samples in different clusters. I think k = 4 makes sense because 4 clusters could mean "Great", "good", "fair" and "bad" for wine and the distribution is close to Gaussian distribution.

ii. Expectation Maximization clustering

1. Explanation of the method

An expectation-maximization algorithm is a method for leveraging maximum likelihood estimation to potential features. The maximum likelihood estimation is a way to density estimate distributions and possible parameters. An expectation-maximization algorithm first estimates the values and optimizes the model. Then repeat two steps till the problem gets resolved. It is one of the unsurprised clustering algorithms and helps group into different clusters. In Figure 6, inertia is smaller and smaller with the increasing number of clusters. It suggests cluster number is less than 10. Besides that, there are many ways to help measure whether clusters are good or not. When the Silhouette Coefficient scores are high, the EM model may provide better-defined clusters. From the graph, clusters may have more and more overlapped information when the number of clusters grows. It suggests picking K less than 10. Homogeneity shows that each cluster has members only from one single class and completeness show that the same cluster includes all members of a given class. The larger, the better. When taking them together, the suggestion would be less than 10. NMI tells that the mutual information among clusters does not go down with the number of K grows. In Figure 7, AIC scores lower with more clusters. In other words, the larger number of clusters provides great model performance. But BIC shows that the EM model has the lowest score and best model performance when the number of clusters is around 10.

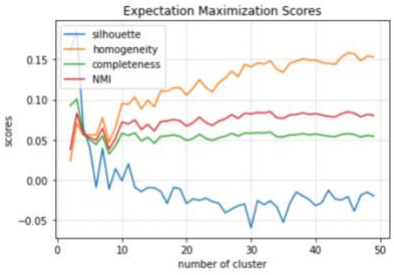


Figure 6: EM measurements

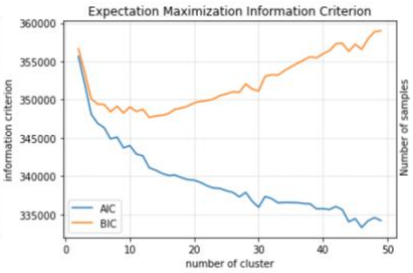


Figure 7: EM information criterion

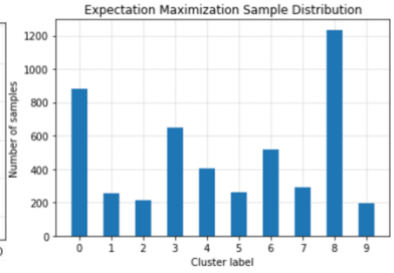


Figure 8: EM cluster distribution

2. Cluster description and results

Overall, 10 clusters will be used in EM because the AIC and BIC performance suggest 10 and 10 are also suggested in four criteria. Figure 8 shows 10 cluster numbers and their distribution of samples in different clusters. Each cluster has samples from 200 to 1200.

iii. Dimension reduction PCA

Principal component analysis (PCA) is one of the dimension-reduction methods that the main idea is to find eigenvalues to project high dimensional space to a low dimensional space while preserving as much of the data's variation as possible. In other words, the first principle component can keep characters of the non-projected data in projected data. The definition is to find the best-fitting line which is to minimize the average squared distance from the points to the line. When components keep growing, the singular value will become smaller and smaller. At the same time, more eigenvectors keep as much of the data's variation as possible. It is a tradeoff. In Figure 9, the suggestion is to use 2 as the number of eigenvectors because it can explain more than 70% variances and it does not give a pretty small value of eigenvalues.

iv. Dimension reduction ICA

Independent component analysis (ICA) is another method to reduce the dimensions of the large dataset with many variables. ICA components are not working the same as PCA's because they are to max statistically independent in each domain. Thus, it is not always necessarily uncorrelated. In Figure 10, average kurtosis goes up when the number of components goes up. But the greatest ratio is happening in 9. The meaning of the average kurtosis is to say how many meaningful results the model captured. Thus, ICA will suggest using 9 components.

v. Dimension reduction Randomized Projections

Randomized Projections (RP) is a super cost-effective way to reduce dimensions. In definition, data in high dimensions is to project onto low dimensions by applying a random unit length matrix. Based on the concept from Johnson-Lindenstrauss Lemma, this kind of projection will not introduce a significant distortion in the data and not be sensitive to impulse noise. In Figure 11, RP suggests 10 to be the best when reconstruction error provides the lowest. Running multiple times won't help lower the error compared to a single run.

vi. Dimension reduction Truncated SVD

The idea of Truncated SVD is very similar to PCA. It used eigenvalues and eigenvectors to decompose the original features but Truncated SVD is useful on large spadatasetsaset, which cannot be centered. To avoid exploding the memory usage, Truncated SVD came up. In Figure 12, a good suggestion is about 4 components because it can provide around 70% explanation and there is a good jump(change) from 3 to 4. Singular values are around 1000 (fair value).

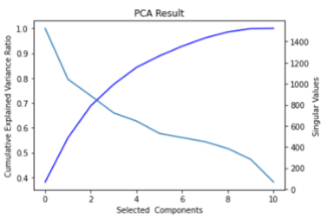


Figure 9: PCA results

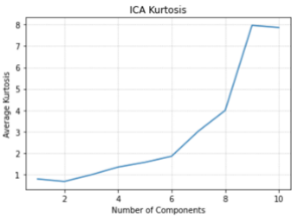


Figure 10: ICA results

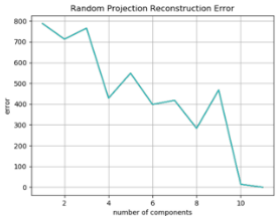


Figure 11: random projection results

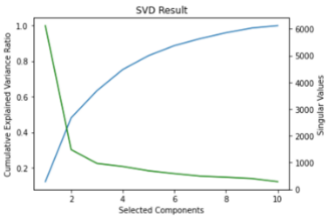


Figure 12: SVD results

vii. Clustering after Dimensionality Reduction

Overall, 2 components for PCA, 9 components for ICA, 10 components for RP, and 4 components for truncated SVD will be applied to re-run two clustering algorithms.

K means: Based on the inertia, all four dimension reduction algorithms and original data show that inertia is smaller and smaller with more number of clusters (does not change). But all four dimension reduction algorithms could provide lower inertia. In other words, all four dimension reduction algorithms perform better in clustering problems compared to the original one. PCA, RP, and SVD give a better score in silhouette compared with the original one. ICA provides optima results in inertia but the case is that ICA does not provide good results in silhouette. Take consideration of homogeneity and completeness together. Only ICA has better results compared to the original one. Even though their scores are different but their patterns are the same, Cluster = 4 is still the best number to pick.

The reason for "no change" in the number of clusters is that dimension reduction algorithms are projecting information, which means that they do not change overall input information even though the dimension has been changed. But under four dimension reduction algorithms, the clustering model has much more clear evidence to conclude.

EM: Original features provide higher AIC and BIC compared with four dimension reduction algorithms. Other measures are also supported better model performance to four dimension reduction algorithms in clustering problems. Take consideration of homogeneity and completeness together. The number of clusters 10 will be suggested. In silhouette score, the number of clusters should be less than 10. In this case, 10 will be used for four dimension reduction algorithms in the clustering problem.

I think the reason for "no change" in the number of clusters should be the same. It is all about input overall information (not about the dimension of the information)

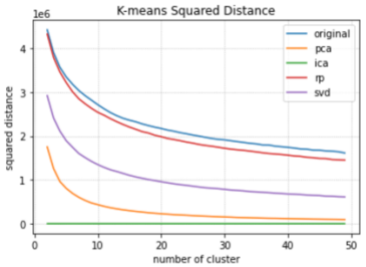


Figure 13: k means distance

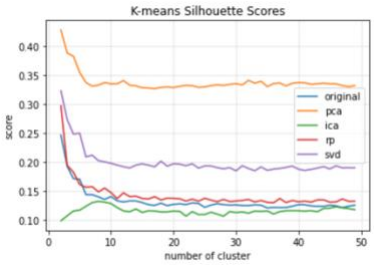


Figure 14: k means measurement(s)

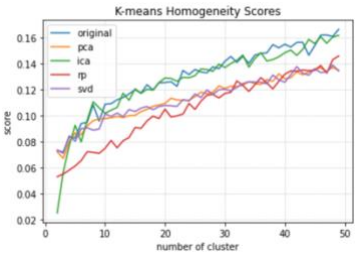


Figure 15: k means measurement(h)

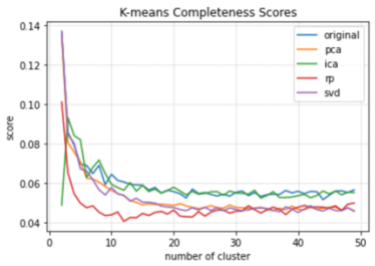


Figure 16: k means measurement(c)

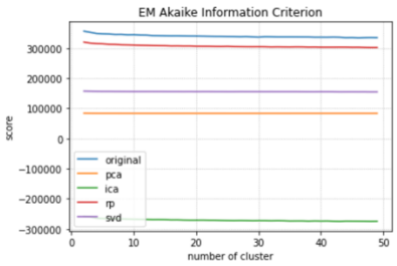


Figure 17: EM AIC

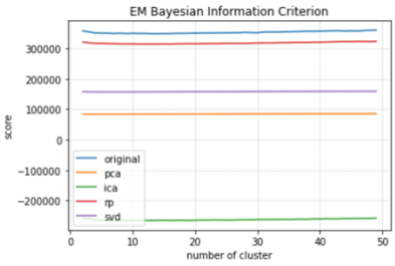


Figure 18: EM BIC

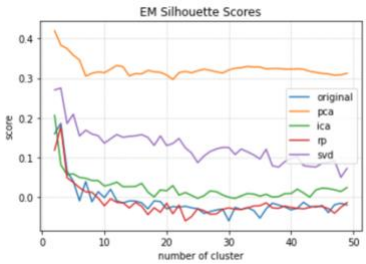


Figure 19: EM measurement(s)

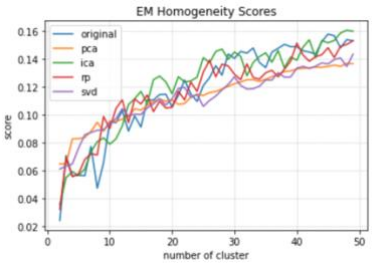


Figure 20: EM measurement(h)

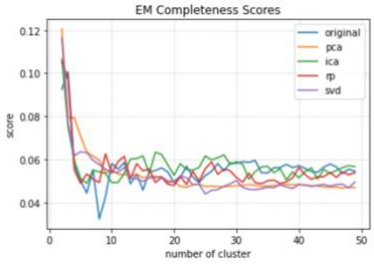


Figure 21: EM measurement(c)

3 Analysis on bike sharing demand

i. K Means

In the first problem. I have introduced K Means and how it works. In this bike demand problem, K means will group different demands of bike into different clusters based on current features. In Figure 22, inertia is smaller and smaller with the increasing number of clusters. Generally speaking, the lower value is, the better algorithm performs. In Figure 23, NMI and completeness suggest k = 3 because a higher score provides higher model performance. Even though k = 3 did not show the highest score in homogeneity, 0.889 is also a very compatible score compared with k = 8 or 12. The advantage for k = 3 is much smaller and computation cost is less than k = 8 or 12. Silhouette indicates that k = 2 or 3 will provide better results. Thus, 3 clusters will be used in K-means clustering because it has relatively high scores and less computation cost for all the measures after trade-off. In Figure 24, cluster distribution shows that each cluster is range from 100 to 700.

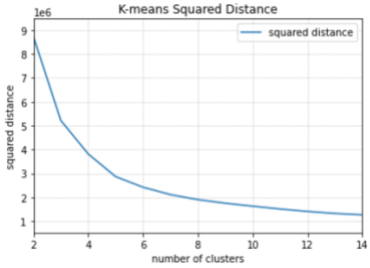


Figure 22: k means distance

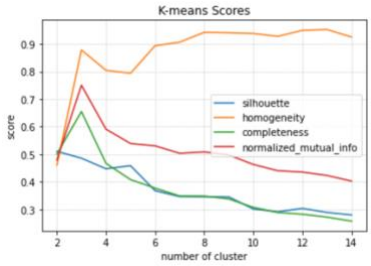


Figure 23: k means measurements

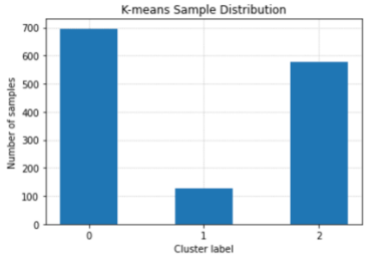


Figure 24: k = 3 cluster

ii. Expectation Maximization

I have been explained the EM algorithm in problem 1. Therefore, I will only general overview of it. Gaussian mixture applied expectation maximization algorithms to get different groups into different clusters. In Figure 25, Silhouette, homogeneity, completeness and NMI suggest the number of clusters is 2 or 6 because they provide higher scores. But in Figure 26, the lower AIC and BIC scores, the better model performs. AIC and BIC get smoothly after cluster number = 6. But AIC and BIC of cluster number = 2 results are still compatible. In Figure 27 and 28, they are shows the distribution of clusters in different cluster numbers. When cluster = 6, one of the clusters only contains 47 samples, which may be too small. Thus, 2 clusters will be used for the EM algorithm since measurement scores are high and computational cost is low. Its range is from 600 to 700.

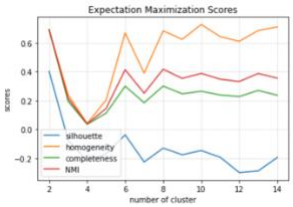


Figure 25: EM measurements

Figure 26: EM information criterion

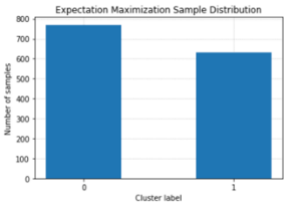


Figure 27: k = 2 cluster

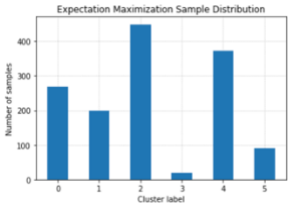


Figure 28: k = 6 cluster

iii. Principal component analysis (PCA)

It is very similar to problem 1. In Figure 29, When eigenvectors grow, eigenvalues become smaller and smaller. But the total explanation goes up. 2 components will be suggested for PCA because it will provide around a 97.5% explanation rate.

iv. Independent component analysis (ICA)

It is very similar to problem 1. In Figure 30, has two peaks, which are 2 components and greater than 10 components. Its meaning is that projection reflects meaningful results. Generally speaking, Kurtosis climbs up when components are greater than 4. Considering computation cost and kurtosis score, 2 components will be applied for ICA.

v. Randomized Projections (RP)

It is very similar to problem 1. It is still re-run 10 times. In Figure 31, 20 components show the lowest error. But considering the computation cost, 18 components will be used for RP.

vi. Truncated SVD

It is very similar to problem 1. It does not apply to centralized data. In Figure 32, 3 components will be applied because it provides a 97.5% explanation rate and the singular value is a fair value (around 300).

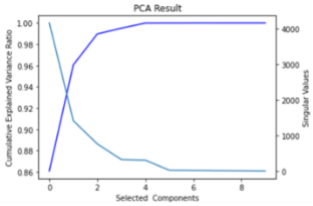


Figure 29: PCA

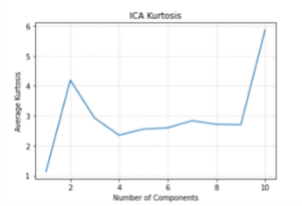


Figure 30: ICA

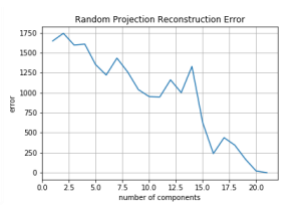


Figure 31: RP

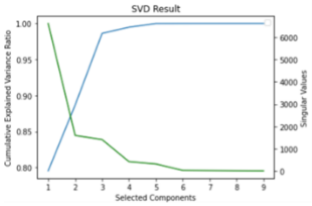


Figure 32: SVD

4 Clustering after Dimensionality Reduction

Overall, 2 components for PCA, 2 components for ICA, 18 components for RP, and 3 components for truncated SVD will be applied to re-run two clustering algorithms.

i. K Means

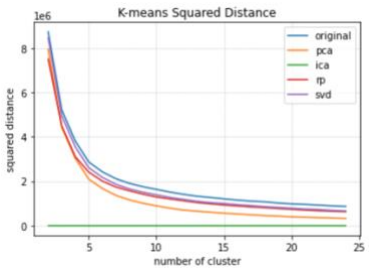
In the first problem. I have introduced all of the concepts. Here I will go recommendation directly. Based on the inertia, all four dimension reduction algorithms and original data show that inertia is smaller and smaller with more clusters. Even though original features show better results in inertia but the difference is slight. PCA, RP, ICA, and SVD give a better score in silhouette compared with the original one. Think completeness and homogeneity together. Except for ICA, the other three have better performance compared to the original one. The possible solution could be k = 3 or k = 5 since their scores are quite close. Considering the computation cost, k = 3 wins.

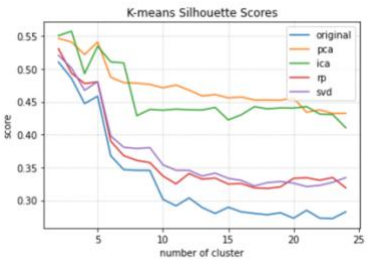
The reason for "no change" in the number of clusters is that dimension reduction algorithms are projecting information. I have mentioned this in problem 1. It means that they do not change overall input information even though the dimension has been changed. But under four dimension reduction algorithms, the clustering model has much more clear evidence to conclude.

ii. Expectation Maximization

It is not very clear whether original features provide better AIC and BIC compared with four dimension reduction algorithms since the results are a mix. Luckily, other measures are supported that better model performance to four dimension reduction algorithms in clustering problems because they all provide higher scores. The number of clusters 6 will be suggested since silhouette score, completeness score, and homogeneity score are the highest.

I think the reason for "change" in the number of clusters is that dimension reduction shows much strong evidence to go with 6 when previously the results of 2 and 6 are very close in higher dimensions. In the previous case, the computation cost is the main consideration to go with 2. But we could make the recommendation based on key measurements after applying dimension reduction.





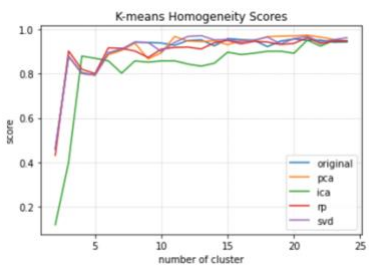
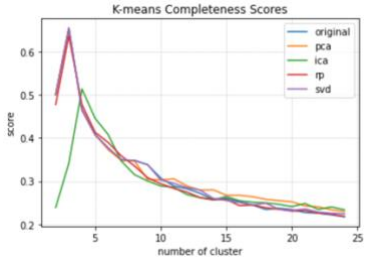
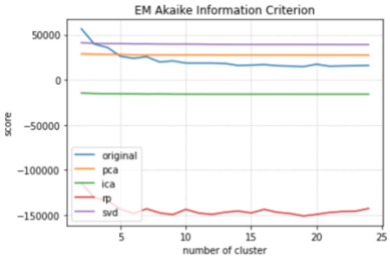


Figure 33: k means distance

Figure 34: k means measurements(s)

Figure 35: k means measurements(h)





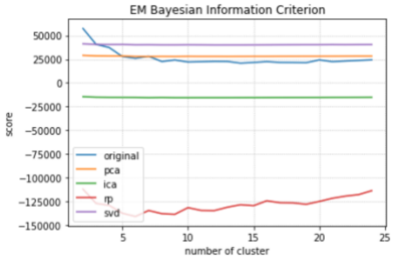
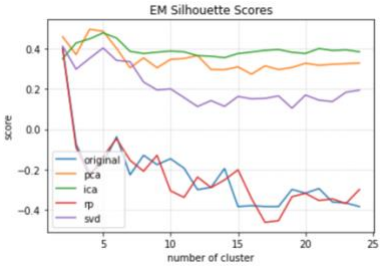
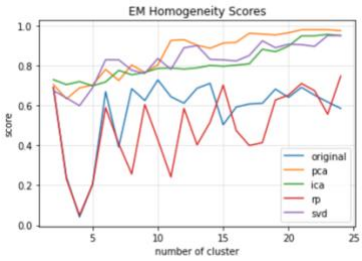


Figure 36: k means measurements(c)

Figure 37: EM AIC

Figure 38: EM BIC





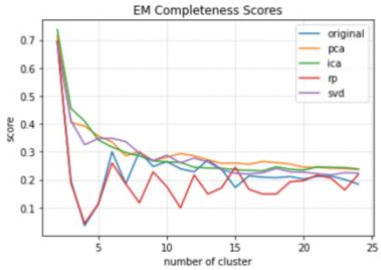


Figure 39: EM measurements(s)

Figure 40: EM measurements(h)

Figure 41: EM measurements(c)

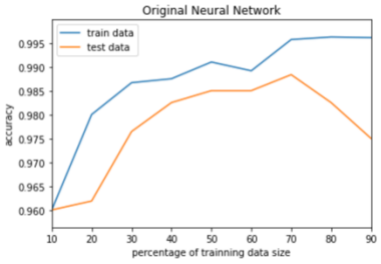
5 Neural Network

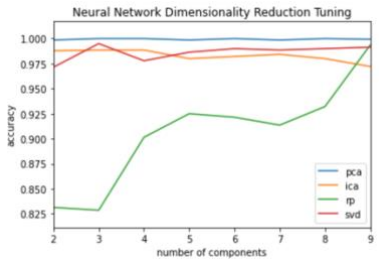
Neural Network will apply to dimensionality reduction data with 'relu' activation and 0.1 alpha. When providing a little larger size of data, the NN of training and testing is starting to convex. In Figure 42, it clearly shows that 70 sample size provides the highest testing score (not overfitting). In Figure 43, all four dimensionality reduction algorithms provide great scores but PCA, ICA and SVD are steady when the number of clusters grows. When providing a larger number of clusters, RP could catch a better score. Once provided a small number of clusters, RP struggles. Thus, cluster 2 is for PCA, cluster 2 is for ICA, cluster 3 is for SVD, and cluster 9 is for RP.

From an accuracy perspective, PCA shows the highest while ICA is the lowest. RP has very unique depending on the number of components. Considering the running time, ICA wins in a second while PCA loses in a second. It will become a running time and accuracy tradeoff overall.

After clarifying the right number of components for each algorithm, K-means and expectation maximization will jump here. Overall, they all perform very well. In the worst situation, training and testing scores are higher than 93%. The dimension reduction algorithms do not help improve the model accuracy. But the good thing here is that training results and testing results are far more close after applying dimension reduction algorithms. In Figure 44, NN with PCA shows stable training and testing results compared to original ones without losing many accuracies. In Figure 45, the EM model gets even better results in NN with ICA compared with the original model, which is very close to 100% accuracy. Even though K-means does not do better than the original model, its performance is stable. Just as mentioned before, RP performs very struggling in this problem if providing a small number of components. Since SVD is very similar to PCA, their overall model performances are very close, which is stable without losing many accuracies in Figure 47. All in all, clustering with dimensionality reduction is a good way to make model performance consistent and help improve the accuracy on some level. In other words, it could help predict a more correct group of different demands for bikes. Considering stableness and overall accuracies, NN with EM and PCA wins the game.

Finally, NN will be rerun on the new projected data.





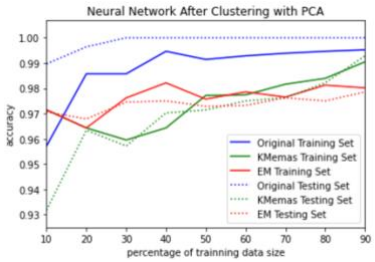


Figure 42: Train and test in ONN

Figure 43: NN dimension reduction

Figure 44: NN with PCA

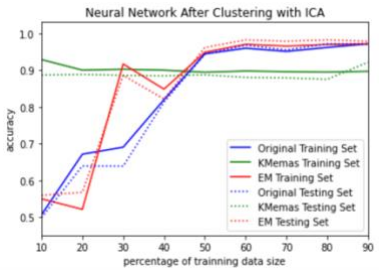


Figure 45: NN with ICA

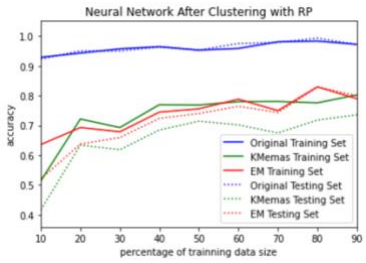


Figure 46: NN with RP

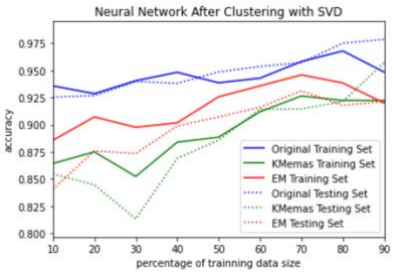


Figure 47: NN with SVD

6 Summary

The number of components for K means and EM are determined by BIC, AIC, silhouette, homogeneity, completeness and NMI. The results of K means and the EM could help improve NN model results and model stableness. Overall, model performance has been increase slightly by 0.55% and running time reduces by 3s.

Not all dimension reduction algorithms work well in this problem. RP struggles. PCA could show better results but running time is longer while ICA could running faster but accuracy is not as good as PCA. K means always show very stable results but it does not improve the original model sometimes while the EM is not very stable but it will give some surprise in results.

Original label does not align my cluster because the number of clusters is greater than the number of original label. Orignal label only considered small and large but clustering alogirthm thinks some of them in between should be label as “medium”. But this group is fairly small (less than 100). There are still 78% of data share the same label with the original one.