BUAN 6341 - Machine Learning

Assignment 4

Yung-Kuei Chen

yxc177030

Data Description

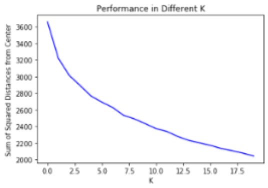
Nowadays, smart phones become an essential part of our life. To have a great phone with reasonable price is the first consideration when people are trying to buy a new phone. Therefore, it is very important to understand the relationship between price and performance about mobile phones to have the highest cost-performance ratio. This is the reason why I chose this mobile price classification dataset. The problem for this dataset is to correctly classify the price range of each cell phone from their features. There are 21 variables in the dataset. The target variable price_range is a categorical variable contained the integer number from 0 to 3. The price_range is 0 means that the lowest price group and the price_range is 3 means the highest price group. The independent variables are the features about the mobile phone. There are some numeric variables such as the size of RAM, the length of the talk time, the speed of CPU and so on. Besides, there are some dummy variables in this dataset, for example, the dual_sim presents whether the phone supports dual sim card or not.

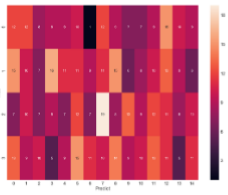
Another dataset is the online news popularity dataset from previous assignment which we want to classify the article as a high share (1) or low share (0) from the attributes of article. This time, I will use unsupervised clustering and dimension reduction skill to it performs

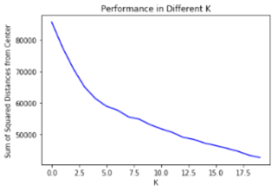
I will apply Kmeans clustering and EM algorithm to cluster the original data. Then apply the dimension reduction skill such as feature selection, principal components analysis, independent components analysis and random projection analysis to reduce the dimension of original dataset.

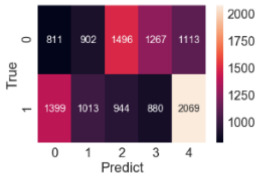
Moreover, I will build neural network to predict the result from the data which has been reduced dimension to see how the performance of dimension reduction algorithm.

Kmean



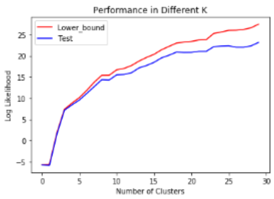


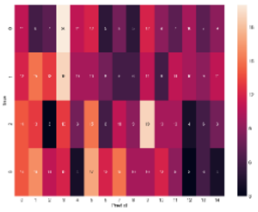


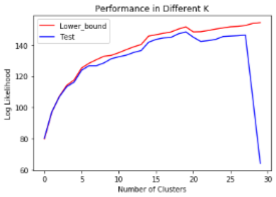


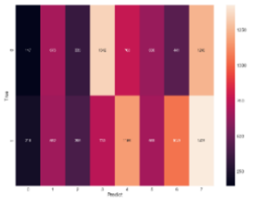
After running the Kmean algorithm, I can identify the performance from all the models with different number of clustering. We can see that the sum of squared distance from the center is decreasing when the number of clustering K is increasing. In the mobile price dataset, I chose K=15 to do clustering. For the online news dataset, I chose K=4 based on the performance plot. To get the better understanding, I plotted the confusion matrix to identify the clustering result.

EM



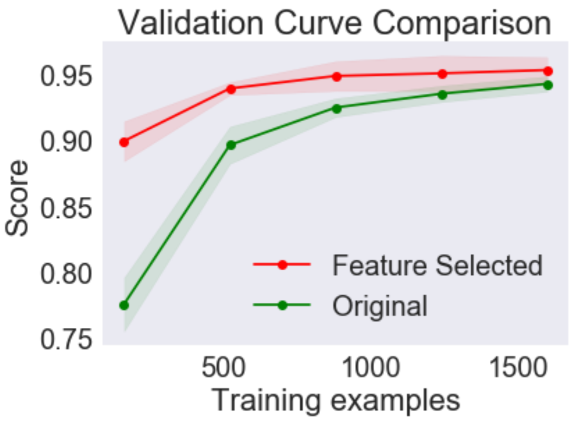


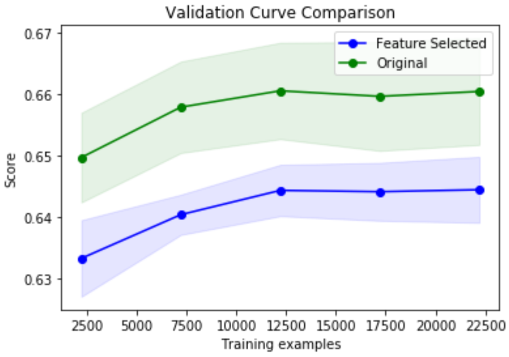




The EM algorithm works the equivalent way that Kmean does. However, unlike the kmean clustering, EM algorithm use log likelihood to identify the wellness of the clustering result. The higher log likelihood is, the better the clustering algorithm is. For the mobile price dataset, the suitable K is 15. For the online news dataset, we will choose K = 20 based on the plot.

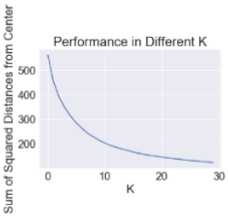
Feature Selection

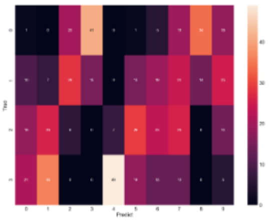


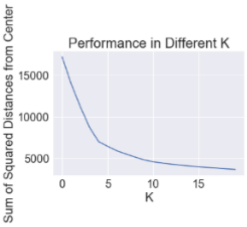


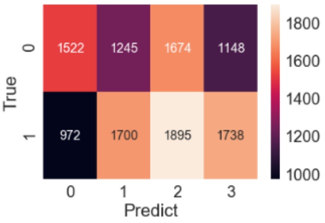
Feature Selection – KMEAN

I applied decision tree algorithm to select the important feature in the dataset. In mobile price dataset, the score result is showing the score base on linear SVM model which performs best among all the other models that I tried in previous assignment. The Feature Selected result is more consistent than the original dataset which reach 90% accurate in the small sample training. However, on the other hand, the feature selected method is not working well in the online news dataset. The selected variable cannot well present the original data.



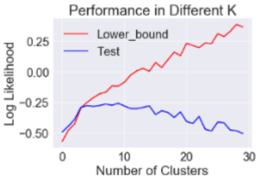


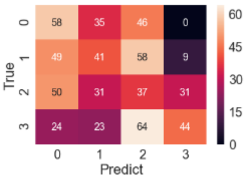


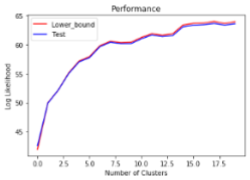


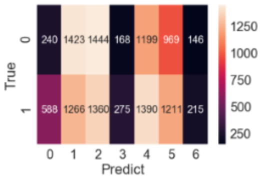
Then, I applied the clustering algorithm again to categorize the selected variables. For selected features in mobile dataset, the chosen K is 10 and for the online news data, the chosen K is 4 based on performance curve. The confusion matrix here is more clearly than the confusion matrix based on the original mobile dataset. That means the feature selection works well.

Feature Selection – EM



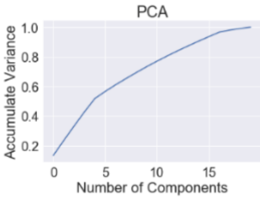


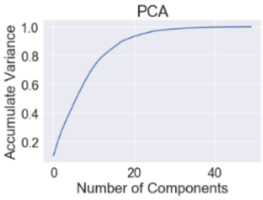


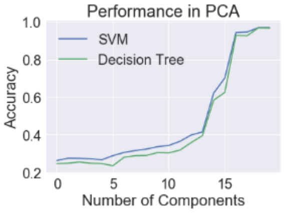


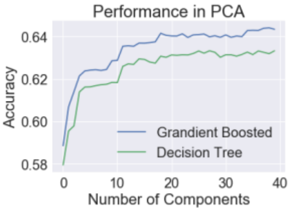
For the EM algorithm, mobile data presents worse after the number of cluster greater than 4. Therefore, I chose K=4 for the mobile dataset. For online news data, the gain of log likelihood decreases a lot after K>7. Thus, I chose K=7. From the confusion matrix we can see that the cluster methods did not perform well. We cannot identify the result clearly.

PCA



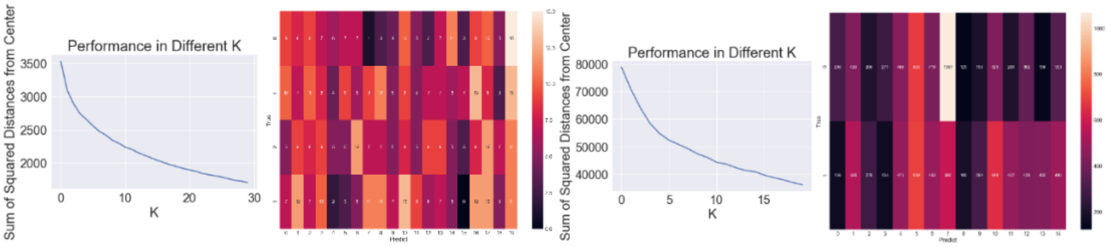






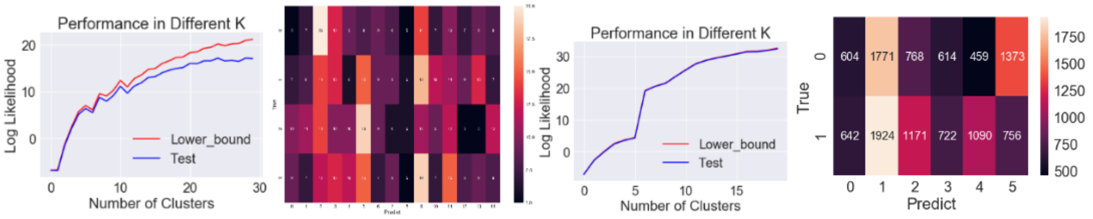
The number of components that we should choose is based on the variance. If a component has the largest variance, it will be the most valuable component. The left 2 plots are talking about the accumulation of variance ratio by different number of components. Combine this information with the accuracy from the model to decide the final number of component. The number of components for mobile data is 17 and for news data is 20.

PCA – KMEAN



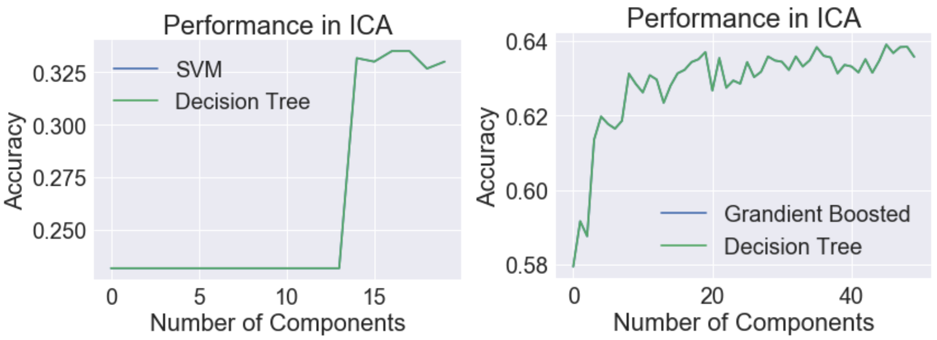
For the Kmean algorithm, I chose K=20 for mobile data and K=15 for online news from elbow test. The confusion matrix in the online news dataset is with lots of black blocks. That indicates that in certain clustering, the cluster does good job because it can separate the data well.

PCA – EM



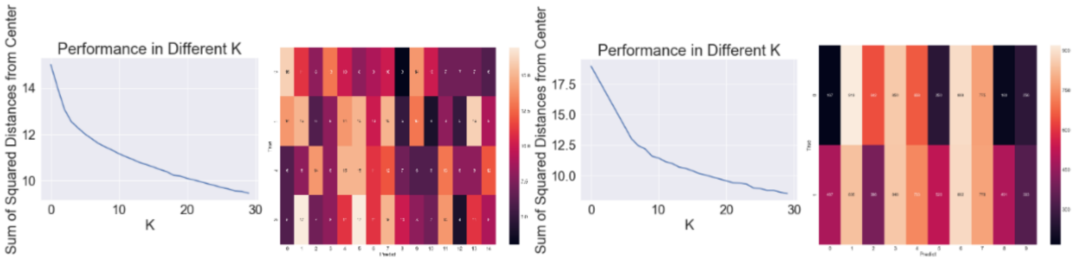
For the EM algorithm, I chose K=15 for mobile data and K=7 for online news data. Because these two number are the most efficient number of component in each dataset. The confusion matrix plot for the mobile data is still messy we cannot indicate any result from it.

ICA



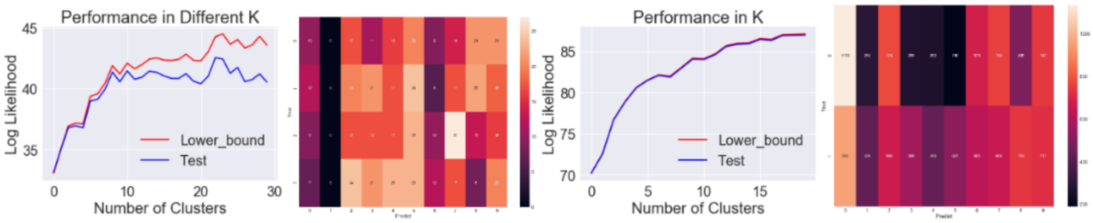
I experienced different number of components in the ICA model with other machine learning algorithm to identify the right number of component in ICA. For the mobile price data, the number of components should at least choose more than 14 because the SVM and DT models perform much better. Here, I chose 15. For the online news dataset, we should choose the number of component around 19 which will have a most efficient result.

ICA – KMEAN



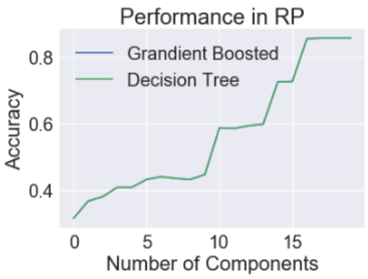
For the Kmean algorithm, I chose K=15 for mobile data and K=10 for online news from elbow test. The confusion matrix in the online news data has some ambiguous clustering such as clustering 1,2 and 7 which are have almost same color in the matrix. That means these clustering have hard time to predict the correct result.

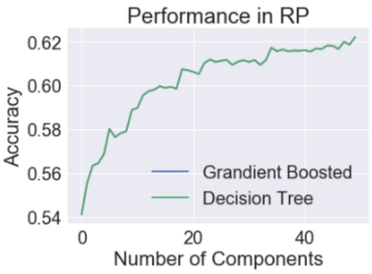
ICA – EM



The log likelihood in the mobile dataset differs a lot from test and training after having more than 10 components. I chose K=10 for mobile data and K=10 for online news data. The confusion matrix plot tells us that there are some clusters are not efficient because there are having the same color which means for different true labels, they all predict the same amount of observations.

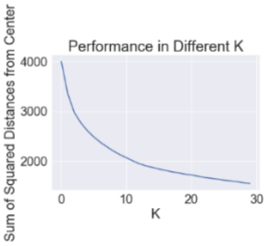
Random Projection

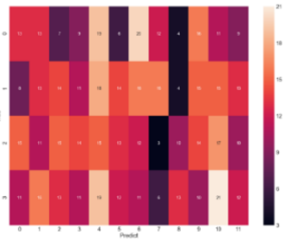


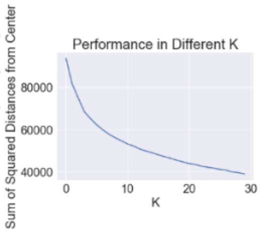


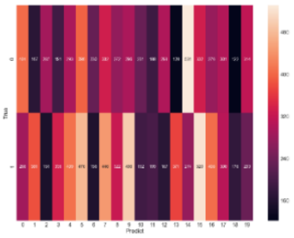
In order to choose the correct number of components in Random Projection method. I applied different number of components in the random projection model with other machine learning algorithm. For the mobile price data, the number of components should at least choose more than 15 because the SVM and DT models perform up to 80% accuracy. Here, I chose 15. For the online news dataset, from the model result, we should choose the number of component around 35 to reach the highest accuracy.

Random Projection – KMEAN



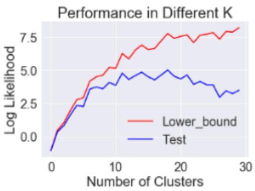


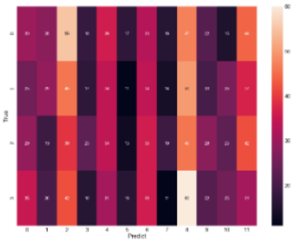


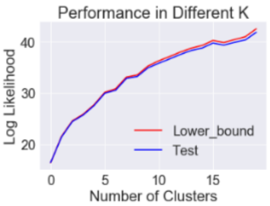


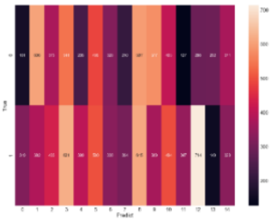
For mobile price dataset, the curve of sum of squared distance rapidly decreased in the first 12 number of clusters. Therefore, I chose K=12. On the other hand, I chose K=20 for the online news dataset because in that situation, the sum of squared distance will reduce at least half when compared to K=1. The confusion matrix in the mobile dataset is holding almost same color in the predicted label. That means the reduction method and clustering are not doing well.

Random Projection – EM



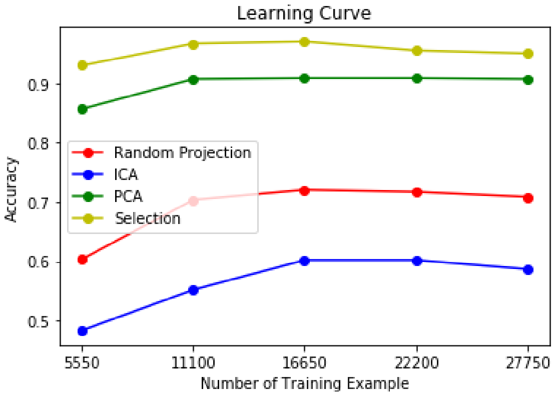


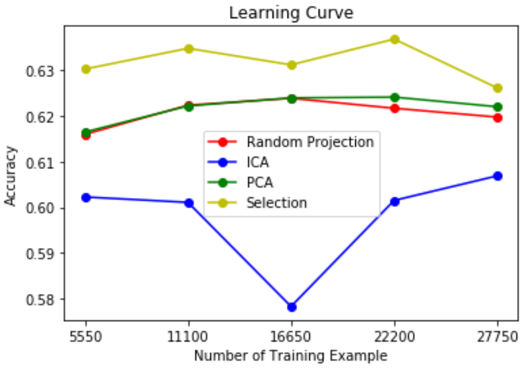




By the elbow test, I chose k = 12 which has the max log likelihood with test data in the mobile price dataset. For the online news dataset, I chose k = 15 which has the most efficient result from the log likelihood plot. With the same situation, the EM algorithm does not predict the label well for both datasets.

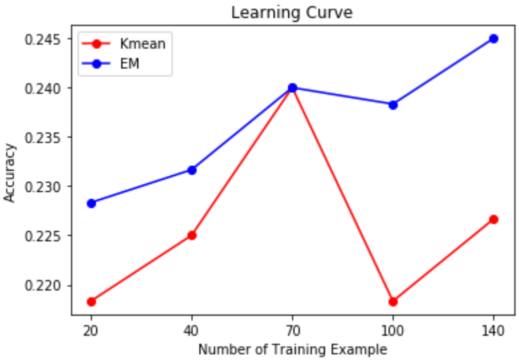
Neural Network

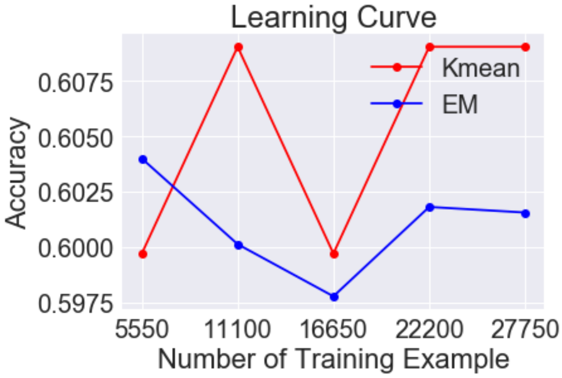




After applying the dimension reduction methods, we can conclude that not all the dimension reduction skills are useful for certain dataset. In the mobile price dataset (left hand), the feature selection method did a great job. The accuracy increased to 0.958 is the highest among other reduction algorithm. In the online news dataset, the feature selection method is still a good method for reducing dimension which reaches 64% accurate. The training time for the neural network is significantly decrease. It is because after applying the dimension reduction method. The size of the data is smaller. Therefore, it does not need as much as computer power as putting the original dataset into neural network.

Label prediction





The last section is about using the predicted label as input and using the true label as output to build a neural network. I used two different clustering algorithm prediction – Kmean and EM as input. For the mobile price dataset, the EM algorithm clustering is a better clustering method, but this method is not a good model for prediction. Although the accuracy is increasing with respect to more training examples, the prediction result is still below 0.25 which is worse than guessing the result. On the other hand, the Kmean clustering performs well in online news dataset, but still does not improve much.

Conclusion

The clustering methods are not suitable for these two datasets. Even if applying the dimension reduction skill, the predicted clustering labels do not really present the true cluster at all. The clusters line up like randomly. They are nonsense. I think the reason behind this is that the data is very similar or compact in the higher dimensional space that make the unsupervised clustering methods hard to separate the data clearly.

After applying the dimension reduction methods, I found out that there is no significant relation about the clustering results between different dimension reduction skills. This problem might be caused by the natural of the dataset. Since the data is not easy to separate into clusters in the original datasets, there is high probability that the transformed data will be hard to separate as well.

The distribution of data in the transformation dimension vary. For example, for the mobile price dataset. After applying the decision tree feature selection. The SVM model performs pretty well. That means the data becomes more linear separable which helps the SVM model can reach higher score in small amount of training example. However, one the other hand, the data transformed by ICA is not a good linear separable data. This is because that the SVM model performs worse after applying ICA.

As a result, it is very important to choose the right dimension reduction method to certain dataset. It will greatly impact the data distribution and makes result worse.