Machine Learning Report

Qiaotong Huang
College of Engineering
Northeastern University
Toronto, ON
huang.qiaot@northeastern.edu

Abstract

In this report, I explored the performance of unsupervised learning algorithms on two interesting datasets. The learning algorithms include k-means, Gaussian Mixture Model, PCA, ICA and UMAP. The two datasets are bank marketing and loan prediction.

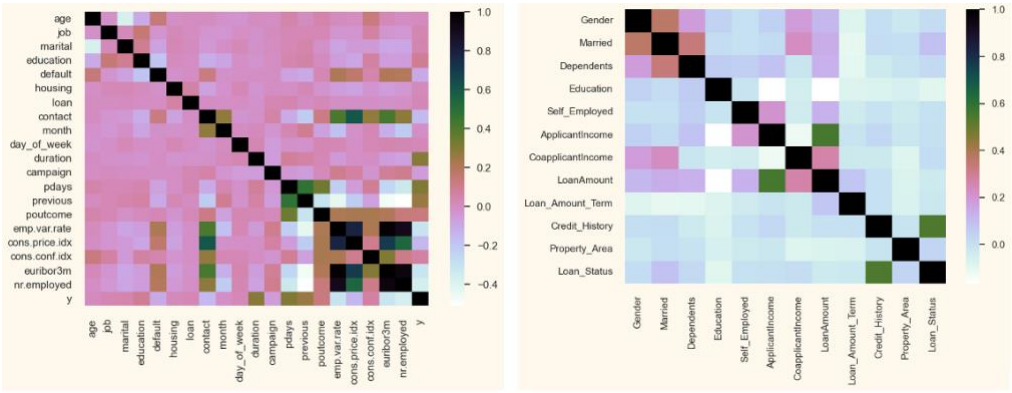
1 Datasets

I used two train datasets which are dividually saved in two CSV files. Dataset 1 is from assignment2. It's a bank marketing prediction feature dataset, and it has a column to label whether the marketing is successful. So we can discover the clustering result usage in supervised learning.

	Data Set Characteristics	Attribute Characteristics	Associated Tasks	Number of Instances	Number of Attributes
Bank Marketing Prediction	Multivariate	Real	Clustering	5000	20
Loan Prediction	Multivariate	Real	Clustering	614	12

1.1 Data characteristics

In the data preprocessing, I found that both datasets are relatively complete and only need to be processed for individual outliers. I repaired some data and visually displayed the data at the end. The correlation heat maps of two the datasets are as follows.



1.2 Why are these interesting datasets?

The practical applications of both datasets are interesting. In the financial field, how to judge whether to grant loans to customers is one of the most important tasks, because you do not want to turn away any good customers, nor do you want any potential loan default risks to affect your own safety. Therefore, we need a predictive model with the highest possible accuracy, which is crucial for banks.

How to judge whether banks are successful in recommending deposit products to customers through telemarketing is a more challenging problem. For this question, I used the data set from the previous assignment2.

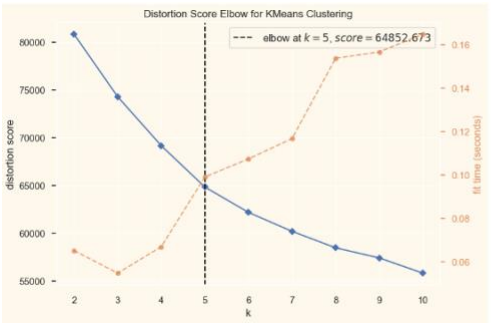
The two datasets have some similarities and some differences. They are all based on commercial real-world datasets with completely different sizes and properties, and their class frequencies are all unbalanced.

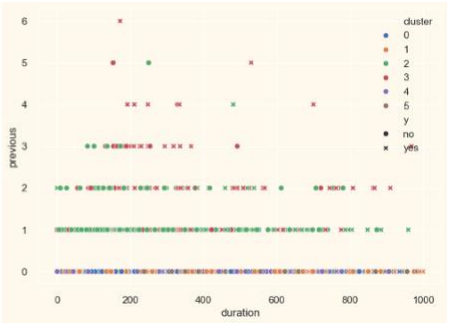
2 Model Implement

In this part, I will separately analyze two clustering model. Raw data have been standard scaled.

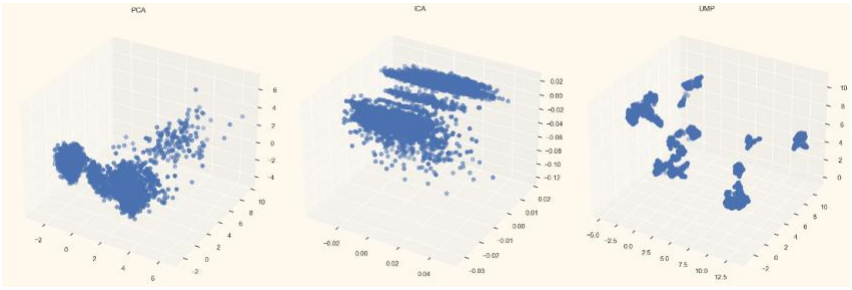
2.1 Bank marketing

Firstly, I tried to cluster without dimension reduction. Elbow curve given by yellowbrick shows the best k is 5

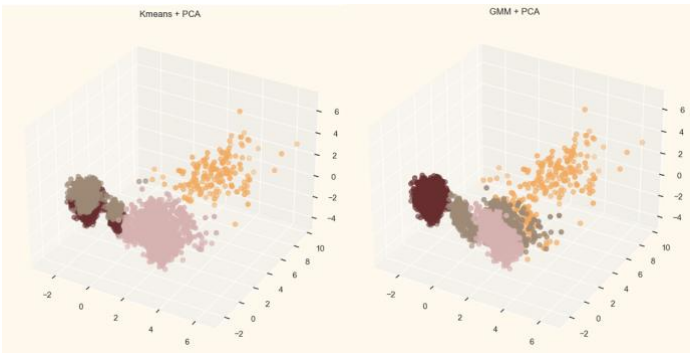




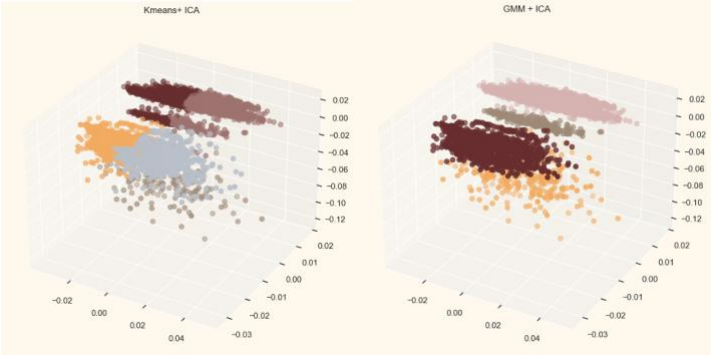
Reducing the dimension into 3dimensions the data in the reduced dimension using PCA, ICA, and UMAP is as follows:



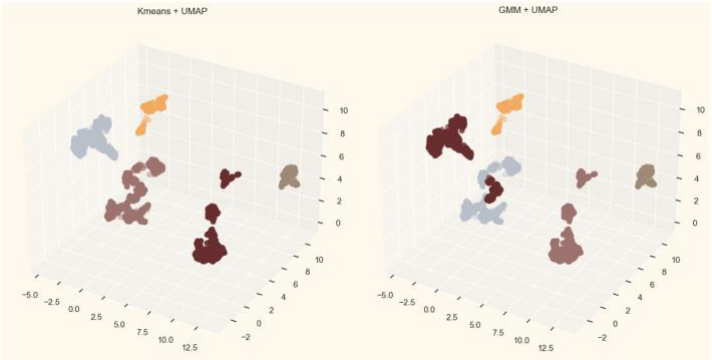
For PCA, the Clusters given by k-means and GMM distribution in 3D space are as follows (best k given by elbow is 4)



For ICA, the Clusters given by k-means and GMM distribution in 3D space are as follows (best k given by elbow is 4)

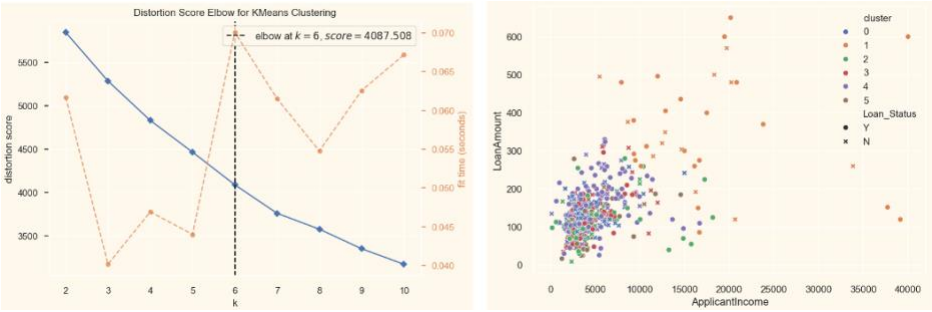


For UMAP, the Clusters given by k-means and GMM in 3D space are as follows (best k given by elbow is 5)

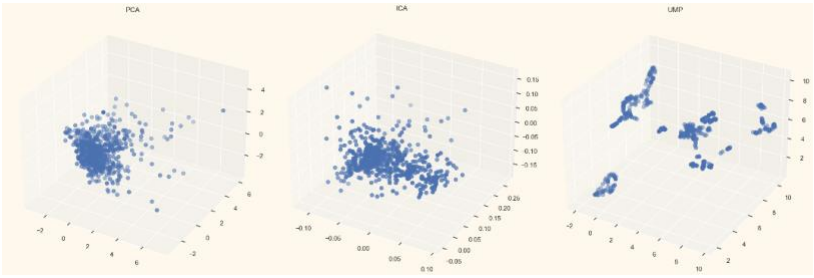


2.2 Loan Application

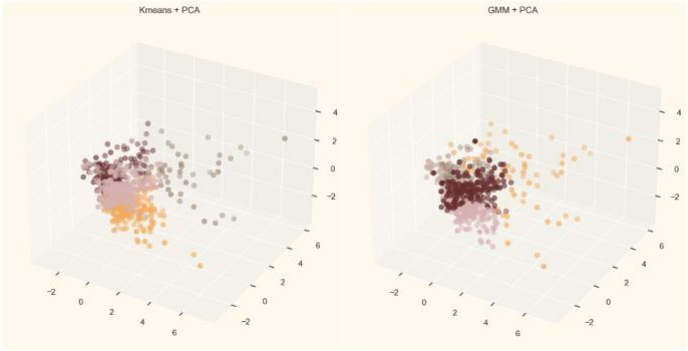
Firstly, I tried to cluster without dimension reduction. Elbow curve given by yellowbrick shows the best k is 5



Reducing the dimension into 3dimensions the data in the reduced dimension using PCA, ICA, and UMAP is as follows:

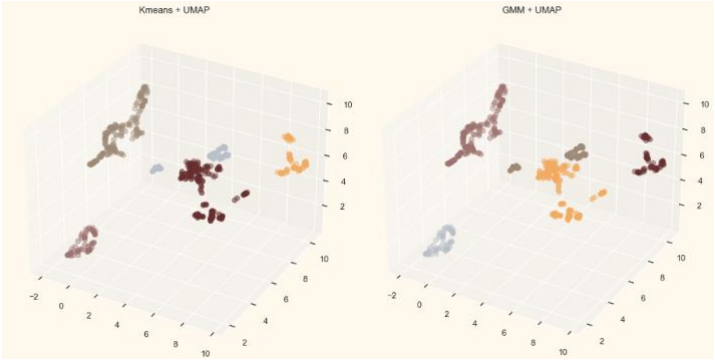


For PCA, the Clusters given by k-means and GMM distribution in 3D space are as follows (best k given by elbow is 4)



For ICA, the Clusters given by k-means and GMM distribution in 3D space are as follows (best k given by elbow is 4)

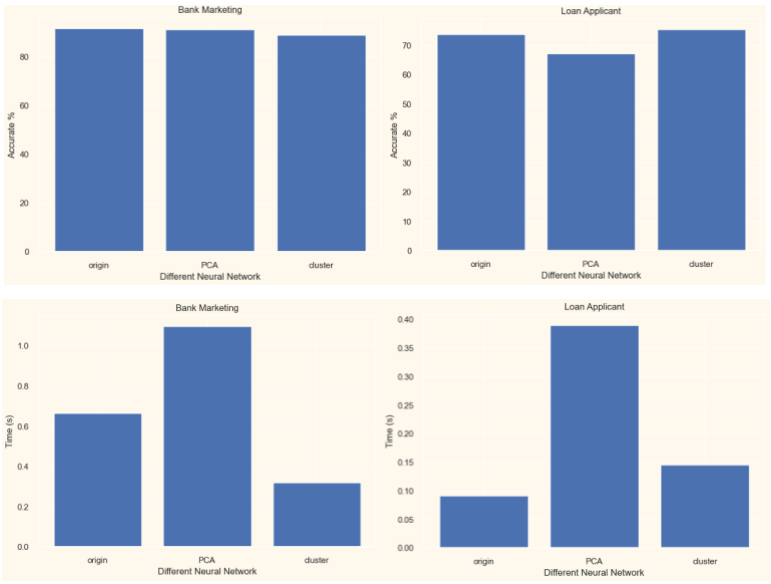
For UMAP, the Clusters given by k-means and GMM in 3D space are as follows (best k given by elbow is 5)



3 Neural Network Comparison

In this part, I used the neural network model to compare the original data and the dimensionally reduced data. The comparison results are shown in the figure below.

In terms of parameters, I chose greedy search to select the best parameters. In the end, there was a slight difference in accuracy between the two data sets, and it is not always the data after dimensionality reduction that can obtain higher accuracy. I think this is related to the neural network. It has something to do with the fact that the network is not sensitive to data latitude. As for the time consumption, it is obvious that the time taken by PCA has increased.

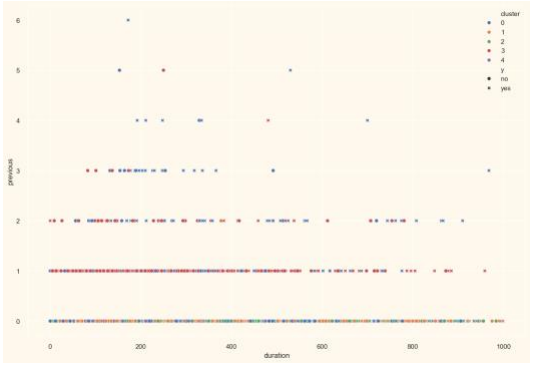


4 Analyze clustering results

From the classification chart in the previous part, we found that in the two data sets, the UMAP-processed data has the clearest clustering performance. Let's take a deeper look at the clustering results. Here I choose the combination of K-means and UMAP.

4.1 Bank marketing

The scatterplot of the location of each cluster is not very clear in terms of duration and previous.



So, in other dimensions of this dataset. We can see the split ratio of different clusters.

I summarize the characteristics of each cluster as much as possible

Cluster0: almost no default, almost no loan, nearly half single, success previous outcome.

Cluster1: almost no default, almost no loan, most married, previous outcome nonexistent.

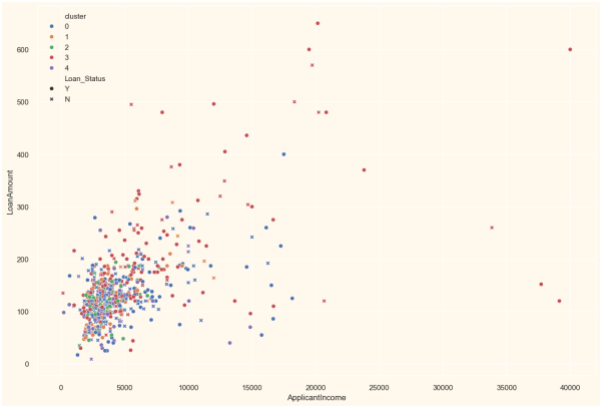
Cluster2: no default, almost no loan, most married, previous outcome nonexistent.

Cluster3: almost no default, almost no loan, most married, failure previous outcome.

Cluster4: almost no default, almost no loan, almost married, previous outcome nonexistent.

4.2 Loan Application

The scatterplot of the location of each cluster is not very clear in the loan amount and applicant income.



So, in other dimensions of this dataset. We can see the split ratio of different clusters.

I summarize the characteristics of each cluster as much as possible

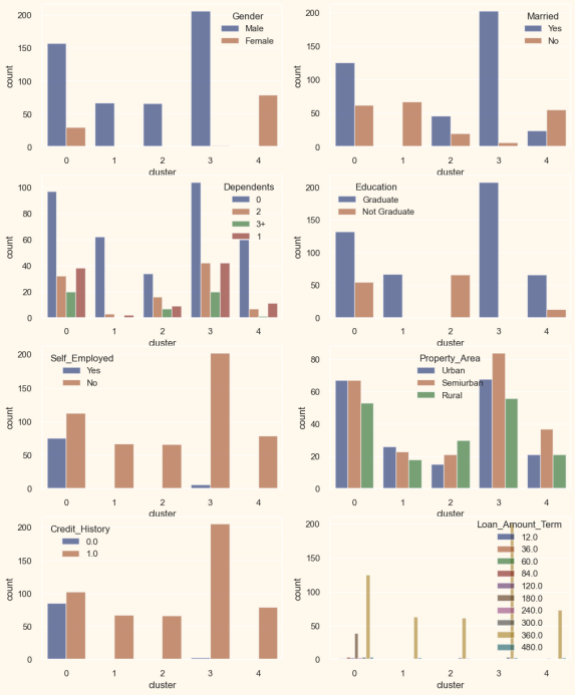
Cluster0:almost cellular, all not self-employed, most graduated, about half have bad credit, most no married, almost all men.

Cluster1:all cellular, all not self-employed, all graduate, all have good credit, no married, all men

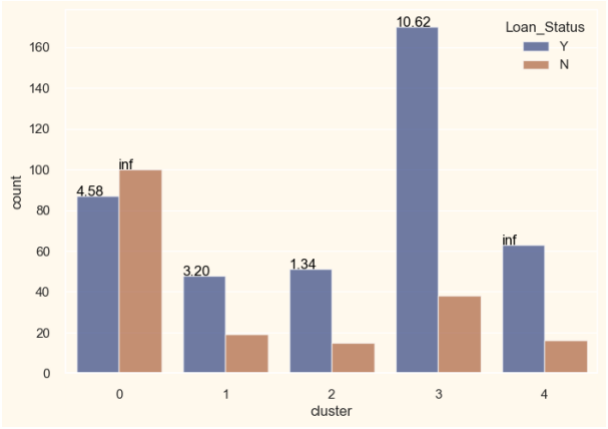
Cluster2: all telephone, all not self-employed, no graduated, all have good credit, most married, all men.

Cluster3: almost cellular, almost not self-employed, all graduated, almost all have good credit, most married, all men.

Cluster4: all telephone, all not self-employed, most graduated, all have good credit, most not married, all female.



Because this is a supervised learning dataset, it has a label: Loan_Status. Here we see from the figure below that except for cluster 0, other clusters are approved more than rejected.



5 Conclusions

In the analysis of the bank marketing dataset and the loan prediction dataset, the characteristics and category distribution of the data were discovered, and a variety of dimensionality reduction and clustering methods were used to explore the distribution of the data. The performance of the neural network model was observed by comparing data sets processed using different algorithms.

Analysis of clustering results includes interpretation of the characteristics and distribution of individual clusters. In the bank marketing data set, there are obvious differences in personal characteristics between different clusters; in the loan prediction data set, there are also obvious differences in the loan approval rates of different clusters.

The study used a variety of unsupervised learning algorithms to analyze two data sets: bank marketing and loan forecasting. Demonstrated deep understanding and insight into the data through the exploration of different algorithms and interpretation of the characteristics of the data set. The reported results provide useful references for further research and practice.

Acknowledgments

The learning code is adapted from: https://github.com/scikit-learn/scikit-learn and references therein.

References

[1] Sukanta Roy (2019) Machine Learning Case Study: A data-driven dpproach to predict the success of bank telemarketing. https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c

[2] Yonatan Rabinovich, Kaggle.com "Loan Prediction Dataset ML Project" https://www.kaggle.com/code/yonatanrabinovich/loan-prediction-dataset-ml-project

[3] Rodrigo Fragoso, Kaggle.com "Explained K-means+ PCA visualization"

https://www.kaggle.com/code/rodrigofragoso/explained-k-means-pca-visualization