Unsupervised Learning and Dimensionality Reduction Report

Lijun Huang  
lhuang349@gatech.edu

# Abstract

This unsupervised learning report will conduct analysis on two different datasets using K-Means and Expectation Maximization clustering algorithms. It will explore dimension reduction algorithms like PCA, ICA, Randomized Projection and Forward Selection. In addition, we will try to see how the dimension reduction output impact our clustering and Neural network results, as well as how clustering output can change our Neural network results.

# Dataset description

## Bank marketing dataset [1]

This is a dataset that used in the previous assignment, it’s a real-world dataset related with direct marketing campaigns of a Portuguese banking institution. The goal is to measure if the product (bank term deposit) would be 'yes' or 'no'.

There are 21 columns and 4119 examples in total. The data includes bank client data like their jobs, age, marital status, as well as some other attributes like monthly consumer price index and number of contacts performed.

## Digit dataset [2]

This is a dataset from machine learning repository that consist of digit data by collecting 250 samples from 44 writers. The input attributes are integers in the range of 0 – 100, and the last attribute is the class code of 0 – 9. There are total of 5620 observations and 64 features (8\*8 grid that is flatten to 1D space). The reason why I choose this dataset is because this is one of the “standard” datasets in the machine learning. I want to see how the behavior compared to a more complicated dataset like the bank dataset above. In Addition, this dataset has 10 optimal clusters in principle, compared to the binary results in the above. Hopefully this would give us more interesting results in the experiment.

# experiments

## Experiments: clustering algorithms

### Intro

This section will talk about two clustering algorithms, to be specific, K-Means clustering and Expectation Maximization. It will cover the methodology of choosing optimal number of clusters and show some interesting finding of clustering results.

### Methodology for choosing K

As this is an unsupervised learning problem, it will be primarily the model to work on its own to discover interesting patterns and information. Therefore, choosing the number of clusters, K, is one of the important step in the clustering algorithm. In this report, the silhouette method is used to determine the optimal K. This method measures how similar a point is to its own cluster compared to other clusters. The range of the score is between -1 to +1, with a higher score to be more desired.

### Clustering results

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

1. Silhouette score for EM and K-Means on two datasets

From the above graph, we can see EM and K-means both choose K = 2 for the bank dataset, and k = 10 for the digit dataset by using silhouette method, which all aligned with the actual dataset. This is a little surprising that the number of clusters align with actual label so well.

In order to get a better understanding of the results, I plotted the clusters by using t-SNE with 2 dimensions, below is a sample results for K-means clusters for the digit dataset. Each color represents a cluster, we can see overall, it works very well.

Chart, scatter chart, bubble chart

Description automatically generatedChart, scatter chart

Description automatically generated

1. Clustering plot on the original data for K-Means

In order to see how it corresponding to the labeled data, a confusion matrix for digit data is plotted, which can be seen from the confusion matrix below. Please be noted this can only serve as “proxy” as the clustering algorithm does not output the actual label. This confusion matrix is plotted by using mode of the predicted clustering to get a rough idea how the clustering is corresponding to the actual label. We can see overall the results looks promising, except 9 is easily confused with 3 and 1, and 5 is easy to be confused with 3, which can be expected by the way how the handwriting is.

A picture containing chart

Description automatically generated

1. Confusion matrix for digit data clustering

## Experiments: dimensionality reduction

### Intro

In this section, 4 different dimension reduction techniques will be applied, namely: PCA, ICA, Random Projection and Forward Selection. We will try to explore what is the best way to reduce the dimensionality of the data, hopefully this would help other algorithms to perform better in the later steps.

### Result Analysis

Chart, scatter chart

Description automatically generated

1. PCA explained variance plot

In the above graph, the distribution of eigenvalues for PCA was plotted for both dataset respect to different choose of number of components. We can see that there is no clear “Gap” for us to choose a good cut point. Therefore, a threshold of 95% of explained variance was chosen. The threshold was set in this way to ensures we don’t lose much information and reduce some dimension at the same time. Hopefully we could reduce the curse of dimensionality issue to improve our final model performance.

Chart, line chart

Description automatically generated

1. ICA average kurtotic distribution plot

ICA is a linear transformation method to make them linearly independent. In principle, ICA would allow us to discover fundamental features of the data, for example, edges or topics. In the above graph for distribution of the average kurtosis corresponding to different n of components. One interesting fact to notice is that for digit data, it’s almost monotonically increasing. This seems to suggest that every pixel in the digit data is “independent”, which is sort of make sense as there is no “individual object” in those handwriting, it is all those pixels that makes up the digit object.

Chart, line chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

1. Reconstruction error for Random Projection and PCA

For the reconstruction error of randomized projections as showed in the plot above, it seems it’s monotonically decreasing as the number of components increases. Even though the variation is not significant for both datasets, and we can see a clear trend in both datasets, the PCA reconstruction error does not have any variation at all! This makes sense because of PCA can be seen as just using a different “notation” of the same data, there is no information lose if we have all the components, that’s why we don’t have any variation for the reconstruction error.

Chart

Description automatically generated

1. Feature importance distribution plot

The last method was using forward selection to determine the important features. The reason why I choose this method is that this would give us more intuitive idea of what features contributes more information, compared to other methods where the data are projected to a new space. In the above plot was using the LassoCV to first plot the feature important for both datasets. We can see that, especially for the bank data set, there are many features that does not contribute to the predictive power at all! Using the forward selection method, first 10 features for bank data and first 30 features for digit data were chosen. As an example, the first 4 most important feature for bank dataset are duration, consumer confidence index, number of years employed and contact communication type. This makes intuitive sense of how it could possibly relate to if there is successful bank term deposit.

## Experiments: clustering on dimensionality reduced dataset

### Intro

Many interesting patterns arise after running clustering on the newly projected space by the dimensionality reduction algorithm. By looking at the results, most of the cases the optimal number of clusters changed, one possible explanation is that this new space reduced some noise and find some interesting patterns that were ignored before, even though it did not agree with our actual labeling. However, it does not need to be the case that it aligns with our label, as that was never part of the training. The cluster algorithm’s goal is to find patterns that are similar to each other, not to find the optimal labeling, even though it would be great if it happens to coincide with our labeling.

### Result Analysis

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

1. Clustering plot on the PCA reduced data for K-Means

Above is an example of new clustering by K-Means after PCA reduced data by using the 95% variance threshold. If we compare it to the original clustering, we can see that the digit data clusters are close to what we have originally. However, we have much more clusters for bank data. One hypothesis is that digit data is well structured and easily recognizable data that corresponding to handwriting, PCA removes some of the whitespaces that does not contains any useful information, thus the clustering we got are pretty much similar to what we have originally. However, this might not be true for the bank dataset, as it can be the case that those features itself have correlation among each other’s and get grouped together by the clustering algorithm, but not necessarily have any indictive power of final labeling. There is more to dig into to see what those clusters are representing.

## Experiments: Neural Network on dimensionality reduced bank dataset

### Intro

This section will run the neural network on the dimensionality reduced bank dataset by using the same parameters, to see how those algorithms would impact our final results. To be more specific, the following are the parameters for the neural network model:

* **Neural network model**: 'hidden\_layer\_sizes': (300, 300, 300, 300, 300,), 'learning\_rate\_init': 0.001, 'tol': 0.006, 'learning\_rate': 'invscaling'

### Result analysis

Chart, line chart

Description automatically generated

1. Learning curve for bank dataset after dimensionality reduction algorithms

From the above learning curve plot for the dimensionality reduced bank dataset running by the neural network model. We can see that PCA, Random Projection and the original model all has severe overfitting issue. This suggests we either need to reduce model complexity or gather more data. When looking back from now, it indeed was a bad choice of neural network model of hidden\_layer\_sizes of (300, 300, 300, 300, 300,). However, it seems our ICA and Forward Selection helped in a sense there is no severe overfitting, and the testing error rate is a little bit lower.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | PCA | ICA | Random Projection | Forward Selection | Original |
| Training Score | 0.9999 | 0.9686 | 1.0000 | 0.9595 | 1.0000 |
| Testing Score | 0.8578 | 0.8949 | 0.8662 | 0.9054 | 0.8737 |
| Runtime in sec | 6.2137 | 4.7403 | 4.2658 | 2.2571 | 4.1598 |

Above is the summary when compare the **ROC AUC** score between original dataset and dimensionality reduced dataset on an 80% training 20% testing split. The split was done before the dimension reduction, because we don’t want our test set to impact dimensionality reduction results. From the above learning curve plot and results table, forward selection seems to be the best algorithm in this, with reasonable fitting to the model and lowest runtime.

## Experiments: Neural Network on clustering results from bank dataset

### Intro

This section will run the same neural network model on the clustering results added bank dataset. The model has the same setting as above section.

### Result analysis

Chart, line chart

Description automatically generated

1. ROC AUC curve and running time for max\_depth parameter

From the above learning curve plot, both training error rate and testing error rate are pretty much a flat line, with almost 0 training error. This indicate the model overfits and we either need to reduce the complexity of the model or gather mode data. We can infer from the plot that there is not much meaningful impact of adding the clustering results into the input space. This might suggest that even though we coincide with 2 clusters in both K-Means and EM for the bank dataset, those clustering does not align with the label we had or provide much meaningful information.

|  |  |  |  |
| --- | --- | --- | --- |
|  | EM cluster | K-Means Cluster | Original Data |
| Training Score | 0.9998 | 1.0000 | 0.9996 |
| Testing Score | 0.8684 | 0.8757 | 0.8796 |
| Runtime in sec | 4.9876 | 3.7673 | 4.8180 |

From the above chart for the **ROC AUC** score and runtime comparison, their performance are very close, those results are kind of expected. Adding the clustering results does not have much meaningful impact to the model performance, nor the runtime results.

# summary

In this report, we conducted unsupervised learning report on two different datasets using K-Means and Expectation Maximization clustering algorithms. As well as explored dimension reduction algorithms like PCA, ICA, Randomized Projection and Forward Selection. There are more to dig into what the clustering output actually are, how the dimensionality algorithm could potentially help us to find a better clustering as well.

# References

1. Bank Data set: [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, <http://dx.doi.org/10.1016/j.dss.2014.03.001>. Available at: [bib] <http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt>
2. Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian. UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)]. Madison, WI: University of Wisconsin, Clinical Sciences Center.
3. Fabian Pedregosa, et al. "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research 12. 85(2011): 2825-2830.

# source code

Below is the link to the source code and data used for this report, more details of running the code can be found in file README.txt

<https://github.com/lijunujil/Gatech_ML_Course/tree/main/SupervisedLearningReport>