Estimating Wine Quality from a 12-Dimensional Data Set

Report of

CS212

Intelligent Data Analysis

MADE BY

Ziyuan Ye

11610203@mail.sustc.edu.cn

UNDER THE GUIDANCE OF

Peter Tiňo

AND

Guoji Fu

Department of Computer Science and Engineering

Southern University of Science and Technology

Shenzhen, China, July 2018

Abstract

This assignment focus on analsying the data set with the technique such as **K-means Clustering**, **Self Organizing Map (SOM)** and **Principal Component Analysis (PCA)**. I try to combine the above methods to work out some problems of a data set which is called **wine-quality**.)) These attributes of the data set seem not correlate with each other. However, there must exist some hidden relationship in the data set. **eg: wine quality must be influence by some attribute**. With the reduce of dimension, gradually the relationship between the attributes were disclosed. There are two data set in my folder, in these assignment I only focus on the red wine-quality data set.

Detail of Red Wine Quality Data Set

12 Dimensions

- Fixed acidity (g(tartaric acid)/dm3), range from 4.6 to 15.9 mean value is 8.3
- Volatile acidity (g(acetic acid)/dm3), range from 0.1 to 1.6 mean value is 0.5
- Citric acid (g/dm3), range from 0.0 to 1.0 mean value is 0.3
- Residual sugar (g/dm3), range from 0.9 to 15.5 mean value is 2.5
- Chlorides (g(sodium chloride)/dm3), range from 0.01 to 0.61 mean value is 0.08
- Free sulfur dioxide (mg/dm3), range from 1 to 72 mean value is 14
- Total sulfur dioxide (mg/dm3), range from 6 to 289 mean value is 46
- Density (g/cm3), range from 0.990 to 1.004 mean value is 0.996
- pH, range from 2.7 to 4.0 mean value is 3.3
- Sulphates (g(potassium sulphate)/dm3), range from 0.3 to 2.0 mean value is 0.7
- Alcohol (vol.%), range from 8.4 to 14.9 mean value is 10.4
- Quality, range from 3 to 8 mean value is 5.6

Data Set Source:

`https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/`

Two Main Topic

In my assignment, there are two question I want to analysis:

- **How does the quality of wine influenced by all the attribute in the data set?**
- **If there are any attribute that influence total sulfur dioxide?**

Preprocessing Data

Normalization

The **unit** and **scale** of the data is different, this method can make different characters have the same scale. Only in this way, the comparison can be valid.

$$\widehat{E[X_i]} \approx \frac{1}{N}\sum_{j=1}^{N} X_i^j \widehat{Var[X_i]} \approx \frac{1}{N}\sum_{j=1}^{N}(X_i^j - \widehat{E[X_i]})^2 = \sigma^2 \widehat{X_{i,nor}^j} \approx \frac{X_i^j - E[X_i]}{\sigma}$$

Chapter 1

Topic 1: How does the quality of wine influenced by all the attribute in the data set?

1.1 Labelling Strategy

When analysing the first problem, the category of quality is the main attribute we want to focus. By analysis the data set of red wine, I discover that quality is range from 3 to 8. So I come up with two strategies.

1.1.1 Labelling Strategy 1

Obviously, the quality of wine can be separated into two class: Acceptable, Recommended. According to the number of different quality of wine. I intuitively partition them as follow.

- Class 1: Quality in range [3, 5].
- Class 2: Quality in range (5, 8].

The final labelling result is:

- Class 1: 744 cases.
- Class 2: 855 cases.

Figure 1.1: Labeling Strategy 1

Histogram of Quality

1.1.2 Labelling Strategy 2

According to James Halliday from James Halliday Annual Wine Companion, he separate wine quality into 5 class which are:Outstanding wines, Highly recommended, Recommended, Acceptable and Others. Thus I separate the data into 5 class as he did.

- Class 1: Quality in range [3, 4].
- Class 2: Quality equals 5
- Class 3: Quality equals 6
- Class 4: Quality equals 7
- Class 5: Quality equals 8

The final labelling result is:

- Class 1: 63 cases.
- Class 2: 681 cases.
- Class 3: 638 cases.
- Class 4: 199 cases.
- Class 5: 199 cases.

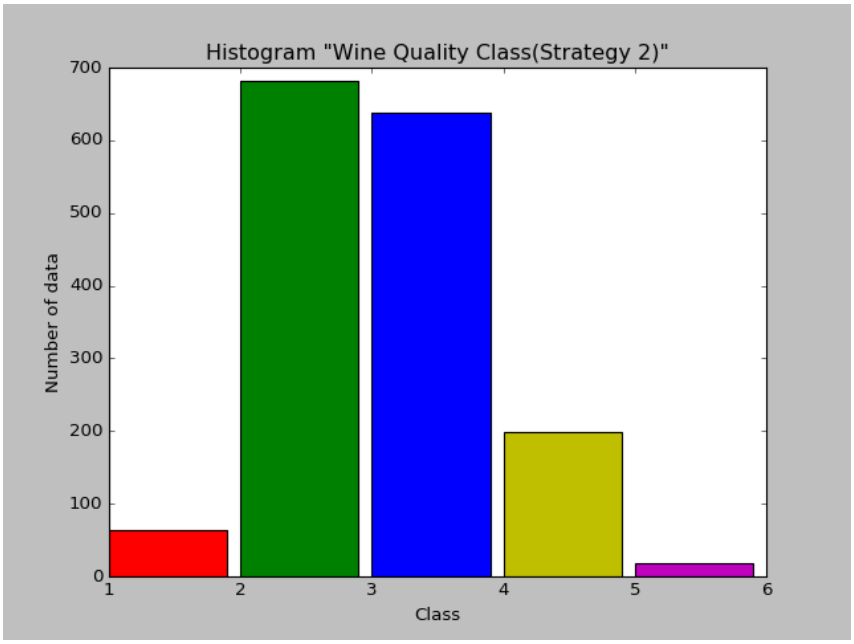

Figure 1.2: Labeling Strategy 2

Histogram of Wine Quality

1.2 Visualization

1.2.1 Principal Component Analysis

After normalization, the co-variance of matrix X can be estimated as:

$$\widehat{C_{j,k}} \approx \frac{1}{N}\sum_{i=1}^{N}(x_j^i - \mu_j)\cdot(x_k^i - \mu_k) = \frac{1}{N}\sum_{i=1}^{N} x_j^i \cdot x_k^i \implies \widehat{C} \approx \frac{1}{N}XX^T$$

If the covariance result is positive, that means that X,Y are positively correlated. The large the covariance result is, the dependency between X and Y will be.

After calculating covariance result, linalg.eig() function in numpy was called to generate the eigenvalues, eigenvectors of the co-variance matrix.

Once finish SVD decomposition, Cov[X] can be calculated as a matrix whose diagonal elements are exactly the eigenvalues of Cov[X].

Many of the features here are related to class tags, but there is noise or redundancy. In this case, a feature dimensionality reduction method is needed to reduce the number of features, reduce noise and redundancy, and reduce the possibility of over-fitting.

The idea of PCA is to map n-dimensional features to the k dimension (k<n), which is a new orthogonal feature.This k-dimensional feature is called the pivot element, and it is a reconstructed k-dimensional feature, rather than simply removing the k dimension features from the n-dimensional feature

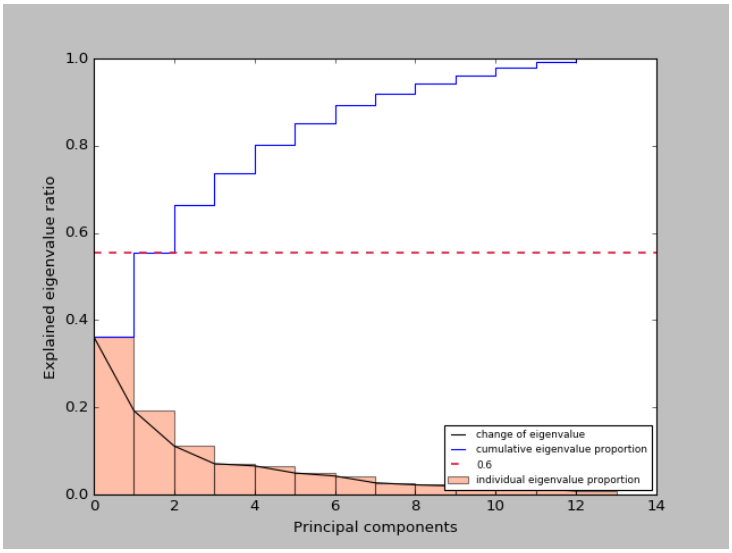

Figure 1.3.1: Cumulative Eigenvalues of Co-variance Matrix

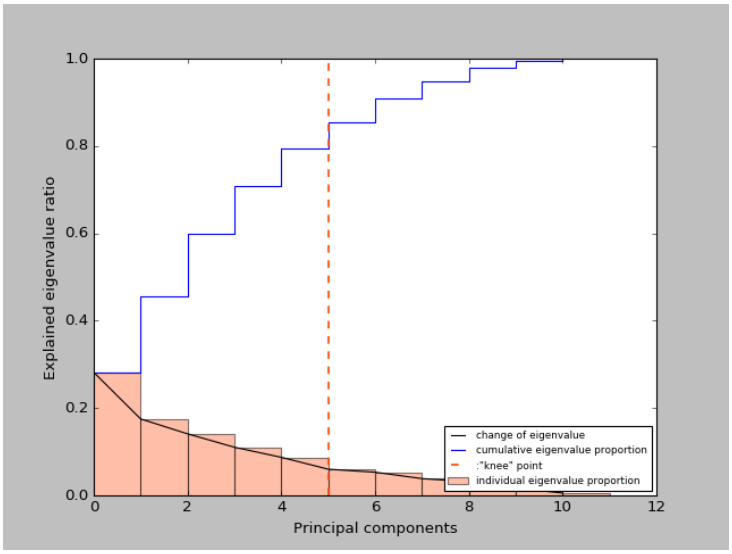

Figure 1.3.2: Knee Points

The above **Figure 1.3.1** shows the cumulative eigenvalues of the co-variance matrix. It illustrates that if we just take two principal components base on this data set, it would not represent the original data set roughly. Hence, in the **Figure 1.3.1**, the knee point can be found according to the decrease ratio as the figure present.

The most important and useful information from PCA is the ranked imporance for the attributes. It means how many separable or variance information is allocated in a certain attribute from the dataset. In the following table, I list the value of the first 2 eigenvectors of the covariance matrix. The first column here shows coefficients of linear combination that defines principal component 1, and the second column shows coefficients for principal component 2.

If the value is positive, there is a positive correlation between the value and the principal component it project to. While, if it is negative, things will opposite the positive one. However, whatever the sign of a value is, I just care about how large the absolute value it will be. The large the absolute value is, the large effect it will do to principal component.

Top 2 Largest Eigenvalues and Eigenvectors			
1st largest: 3.09913244		2nd largest: 1.92590969	
Dimensions	Values	Dimensions	Values
fixed acidity	0.48931421519678553	density	0.5694869591070468
citric acid	0.4636316563339583	total sulfur dioxide	-0.5236044991201001
sulphates	-0.43851962406530764	volatile acidity	0.3986719794517932
pH	0.3953530087692868	density	-0.37857017927744485
alcohol	0.242921330946993	'chlorides'	0.20275695801841312
volatile acidity	-0.23858436259606988	citric acid	-0.1691026594058847
chlorides	0.21224658194729165	sulphates	-0.1637023809766136
residual sugar	0.14610715358517834	residual sugar	0.10175002720609787
total sulfur dioxide	-0.11323206500149913	pH	-0.07645796788428783
free sulfur dioxide	-0.036157524410518845	free sulfur dioxide	0.055788721251921254
density	0.023574853564211597	fixed acidity	0.04560728999309162

Table 1.3.1: Top two eigenvectors

Analysing the **Table 1.3.1: Top two eigenvectors** In new axis 1,**fixed acidity** index is the most important part because the coeffcient is **0.48931421519678553**which do largest contribution to the **1st principal component**. Similarly, **density** is the most important part of **2 nd principal component** which is **0.5694869591070468**. Combine with two column as the table display. We can only find that just **residual sugar** appear in the **last four value** both in 1st and 2 nd column. Thus, probably **residual sugar** is the most useless feature in all data set.

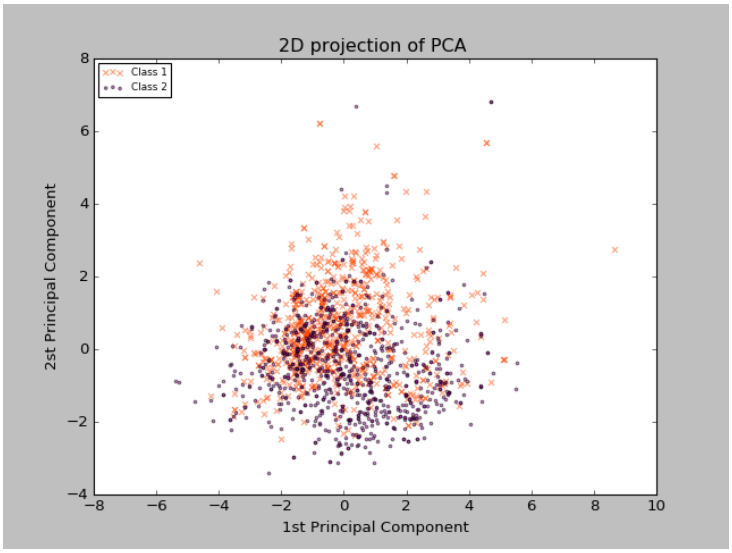

Figure 1.3.3: 2D PCA projection

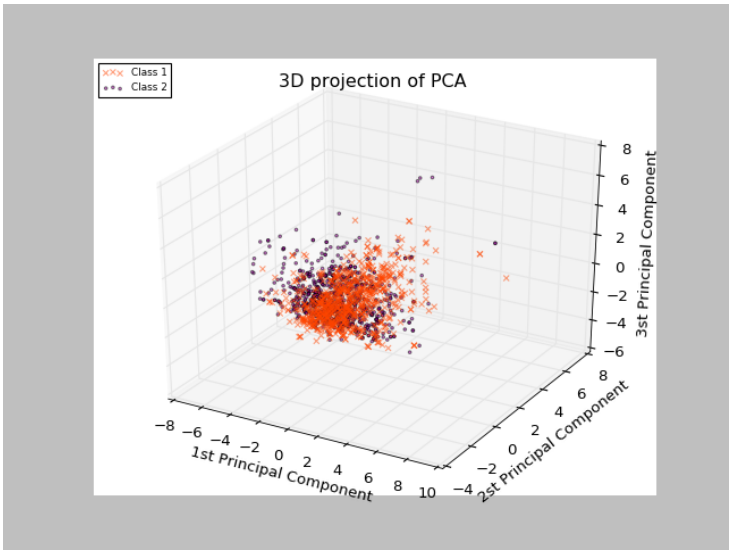

Figure 1.3.4: 3D PCA projection

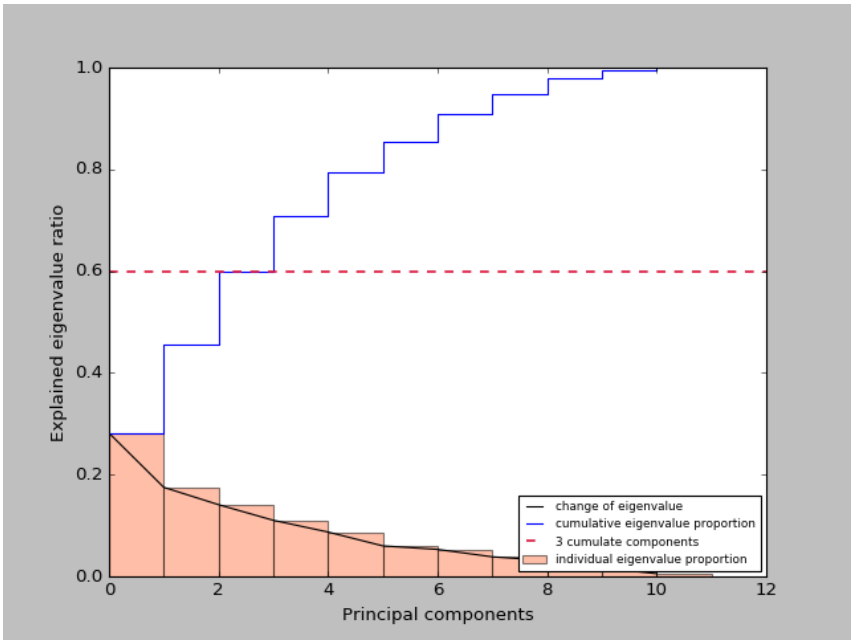

Figure 1.3.5: Ratio of 3 cumulative components

1.2.2 Self-organizing Map

To compare the result of directly putting original data set to SOM method with SOM data after PCA, I do the direct SOM first.

It is quite clear that the data be separated into two region, but actually we want to see the data can be analysis or observe in the lower dimension. Directly som seems to reduce the dimension, however it is a **fake**. With the check of normal vectors of the grids most of the cos value of the normal vectors are negative values or very small values, which means that the angle between each grid is pretty large, which confirm my judgement.

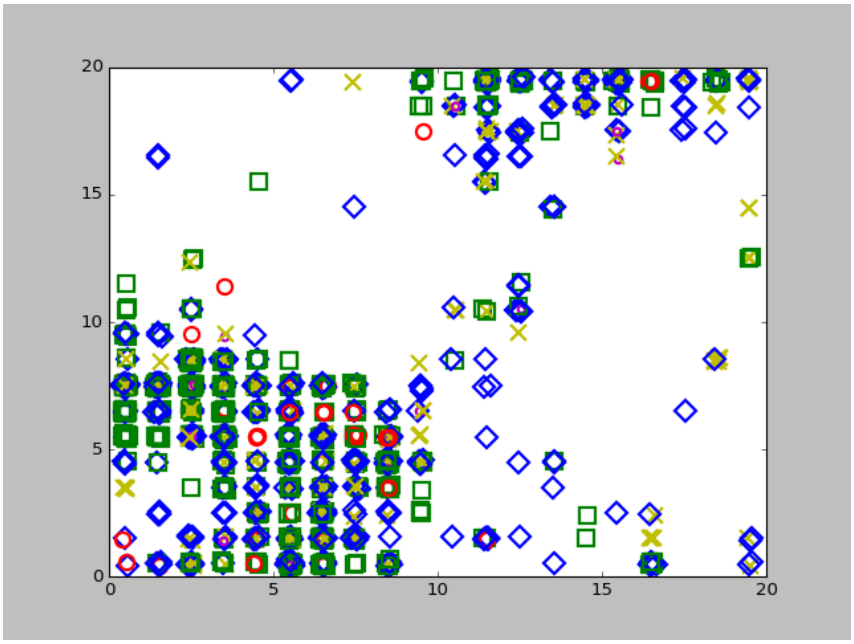

Figure 1.3.6: Direct SOM of 11 Dimensional data

Next, I used PCA to make the data reduce to **5 dimensions** with the use of label strategy 2. Here, the figure cover iteration 10times, 30 times, 50times.

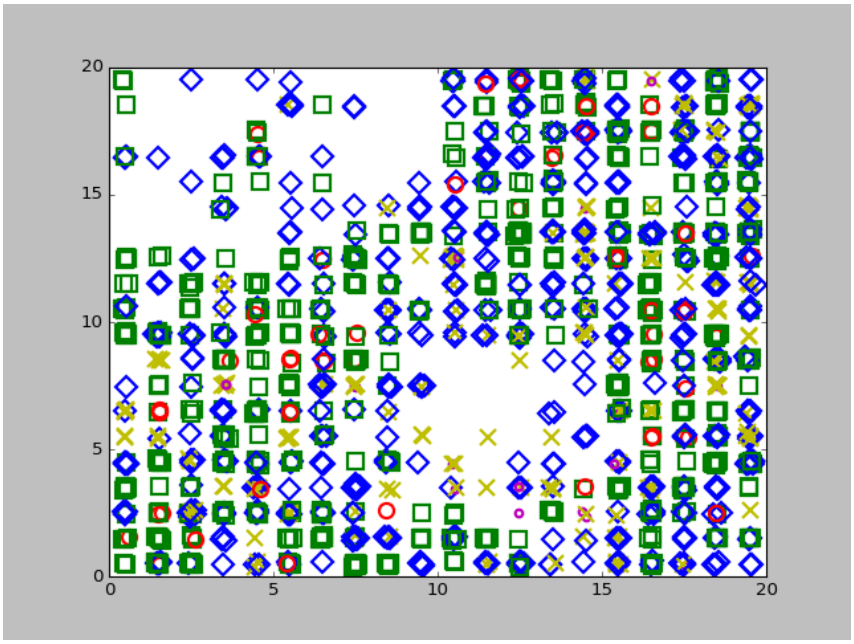

Figure 1.3.7: SOM (Iteration: 10 times)

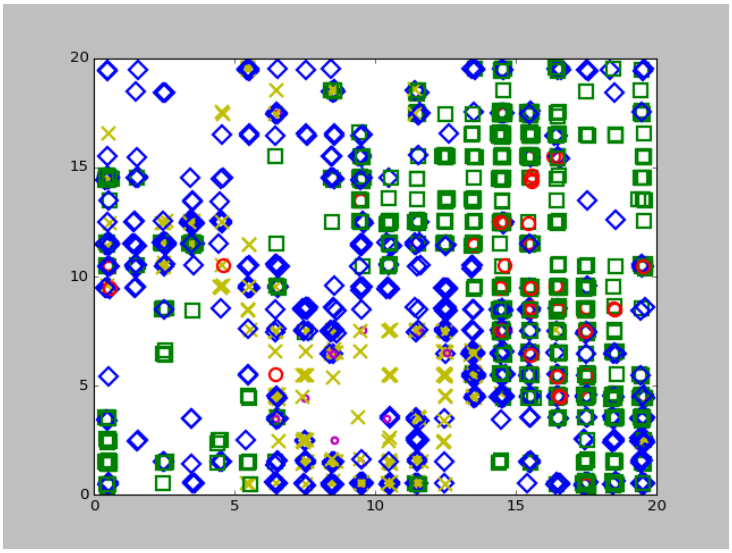

Figure 1.3.8: SOM(Iteration: 30 times)

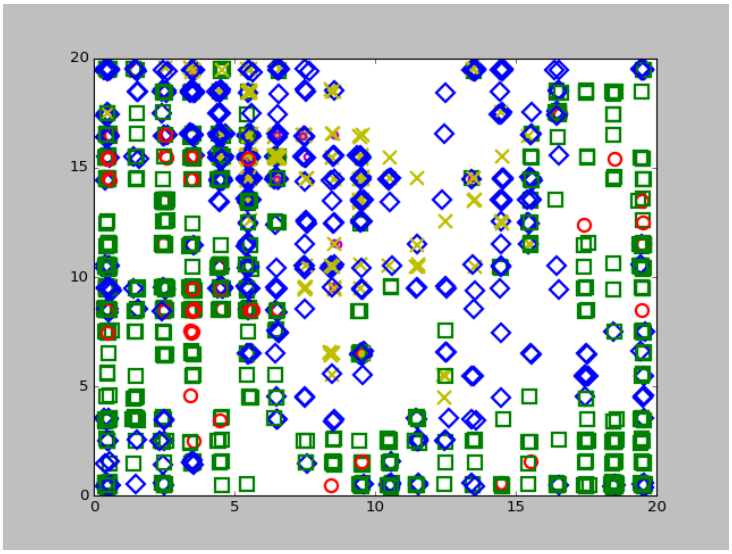

Figure 1.3.9: SOM(Iteration: 50 times)

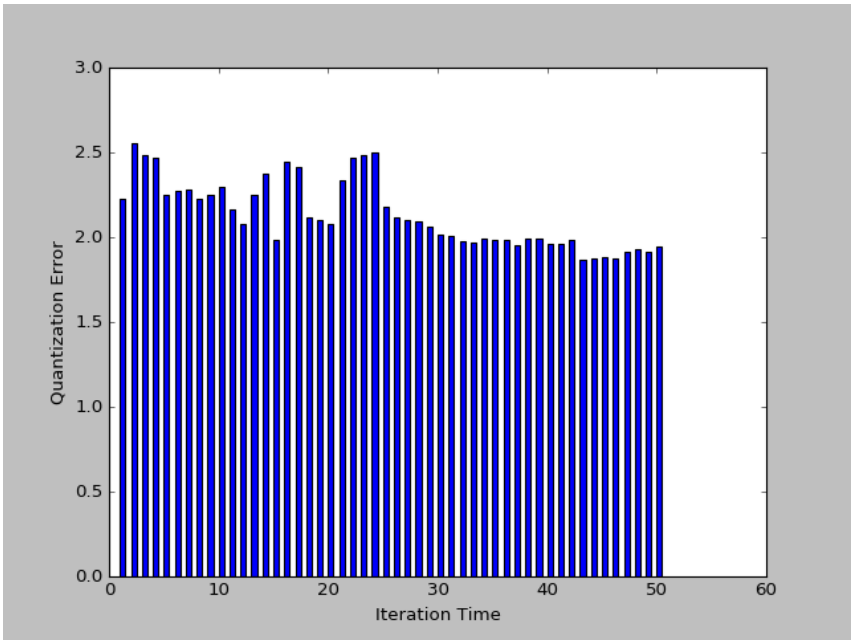

Figure 1.3.10: SOM Quantization Error)

The fluctuation of the **Figure 1.3.10** shown might resulted by high learning rate α and high τ

In conclusion, SOM can reduce the high level dimension data set to lower level dimension which will make us easier to see the data feature in the current lower coordinate. Although in some coordinate there exist some separate small cluster, we can find the link of nodes with same label dominate to whole graph. And we can simply find the the data be separated into **mainly 3 clusters: yellow, green, blue**.

1.2.3 The Comparison PCA and SOM

PCA and SOM are different technique to reduce the dimension of data. There exist some common and difference between them. My conclusion is as follow.

- When we focus a data set with high dimension, if we want to reduce the dimension, we should **not use SOM alone**. We should use PCA first to reduce the dimension as your wish, then do the SOM to make it easier to see the data feature in the current lower coordinate.
- Both of them can reduce the dimension of data. However SOM is more likely to use for compressing and stretching the data to suit the new coordinate. PCA can just combine the current related dimension of coordinate to form a new coordinate which is rather than a projection but a combination.

1.3 Clustering

According to the conclusion of SOM after PCA, we simply find that there exist 3 clusters in the data set.

In order to find the difference of clustering with and without PCA, here I made a contrast group as follow.

1.3.1 Clustering Without PCA

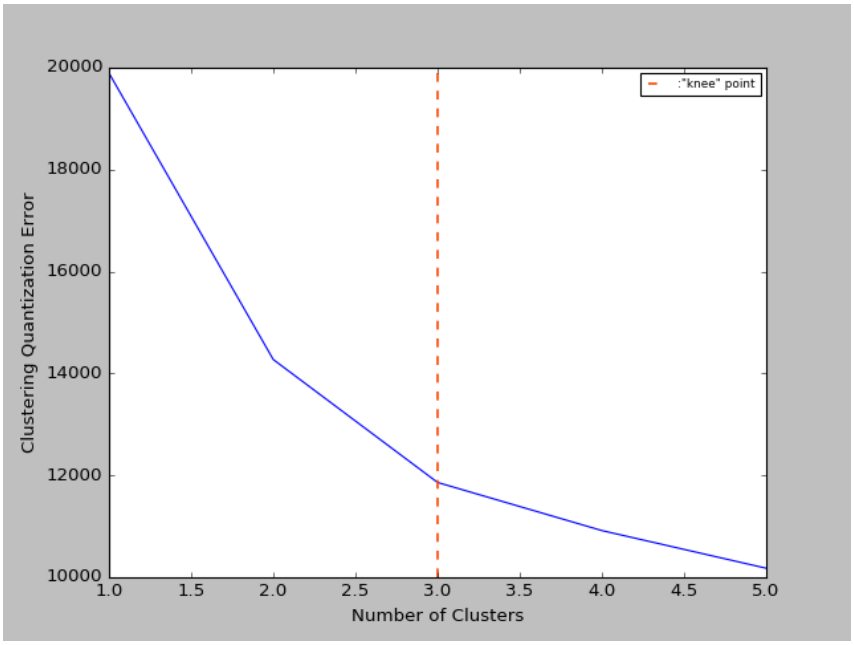

Figure 1.3.1:Clustering Quantization Error without PCA

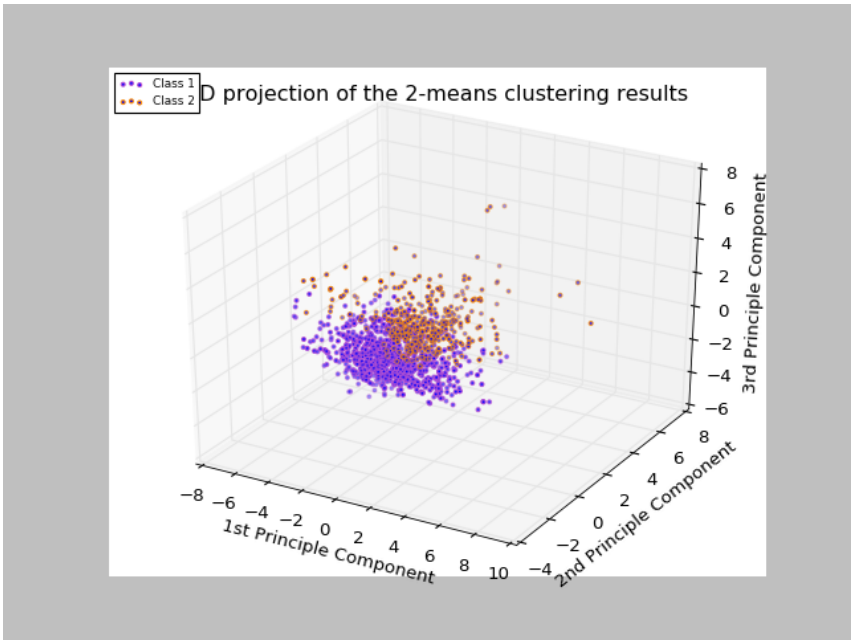

Figure 1.3.2: 2-Means Clustering

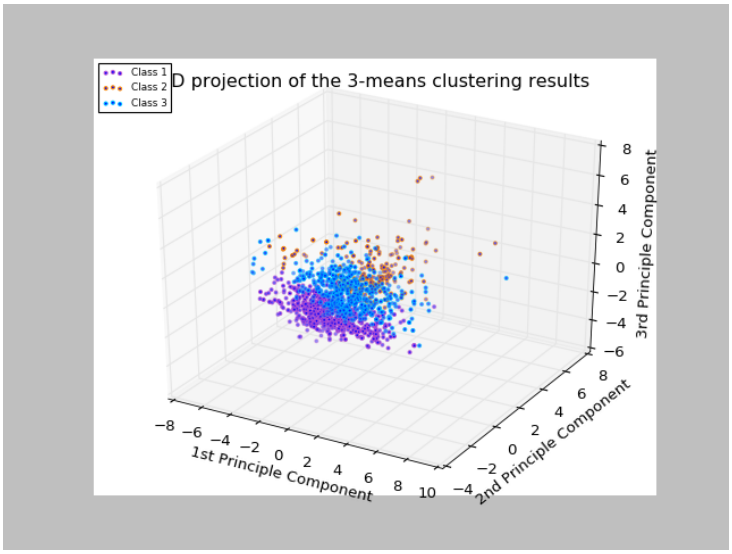

Figure 1.3.3: 3-Means Clustering

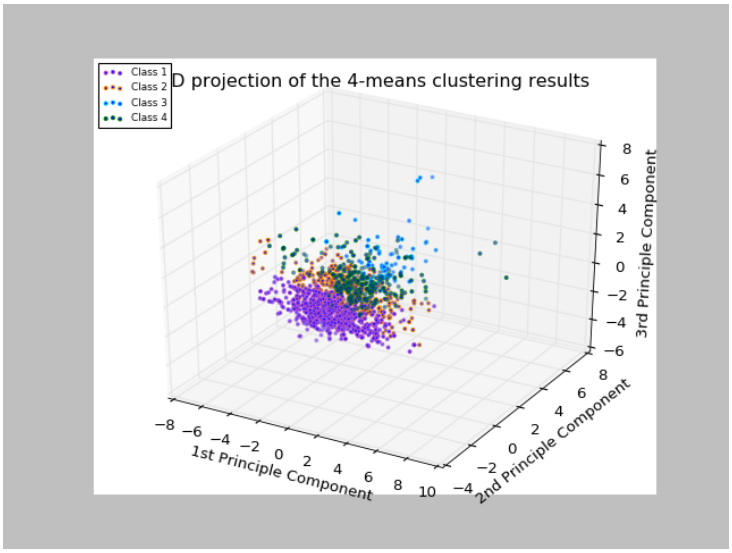

Figure 1.3.4: 4-Means Clustering

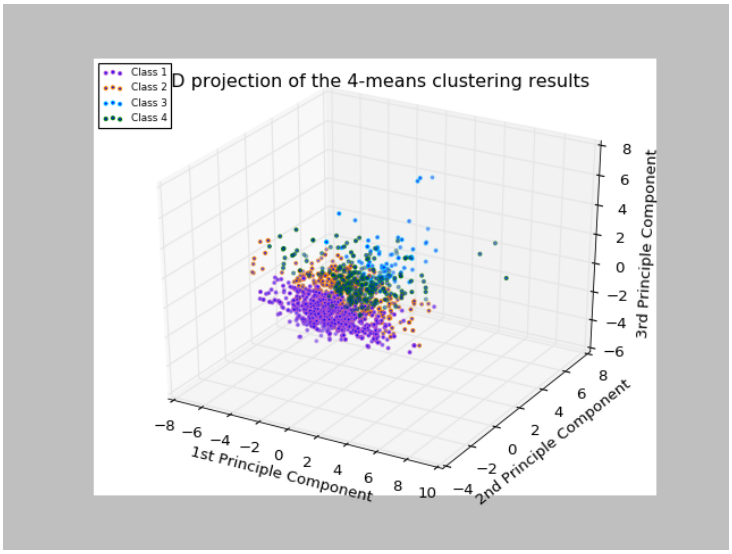

Figure 1.3.5: 5-Means Clustering

1.3.2 Conclusion

Without doing the PCA, the data we observe in the 3D level always aim to keep out each other. For example, in the Figure 1.3.5, the node with green color block out the shown of other node. If we can not observe the data straight forward, it might not a good method.

1.3.3 Clustering with PCA

We have see the result of clustering without PCA. It seems not be a good method to obsever. Hence, I try to do PCA first to reduce the dimension of data, to see if it can improve the observation.

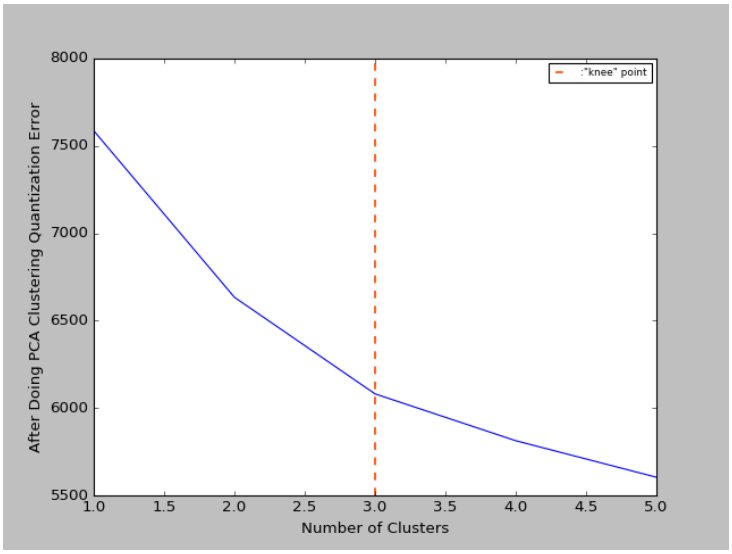

Figure 1.3.6: Clustering Quantization Error with PCA

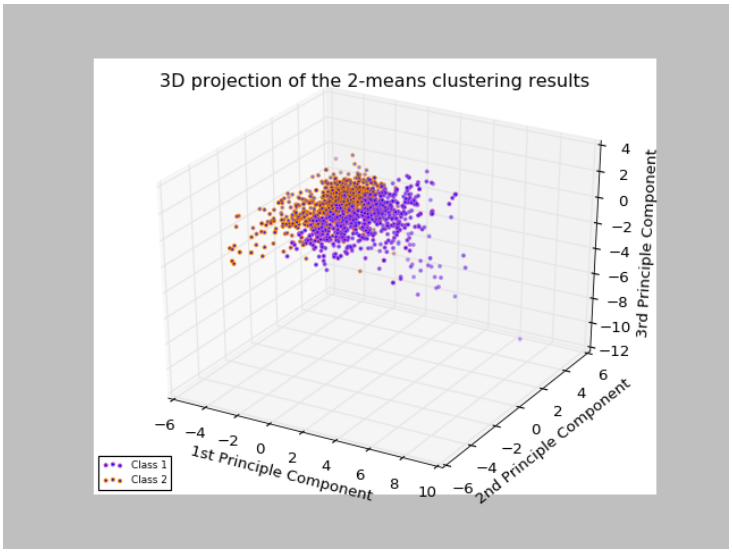

Figure 1.3.7: 2-Means Clustering

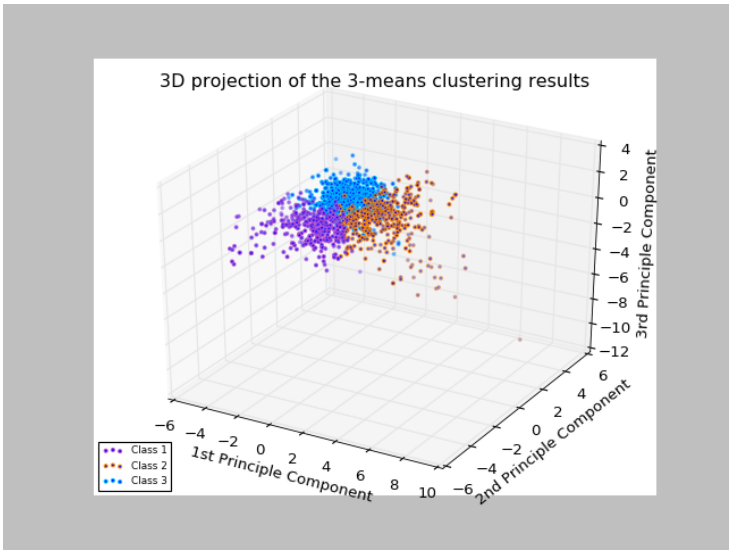

Figure 1.3.8: 3-Means Clustering

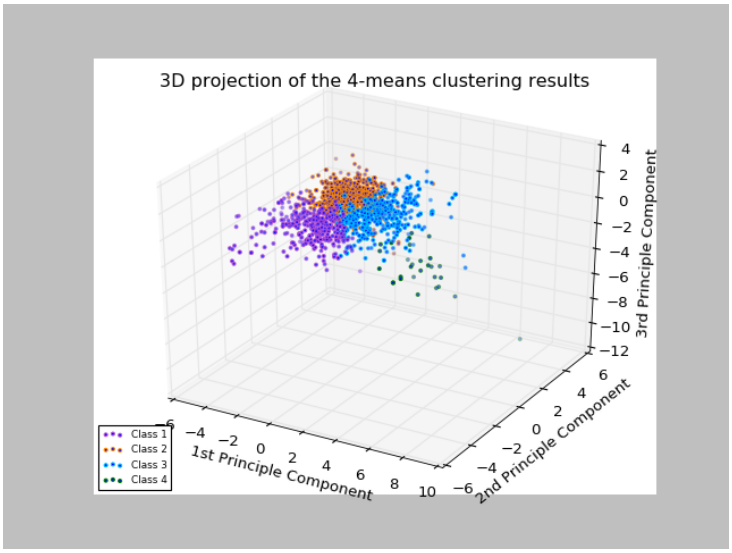

Figure 1.3.9: 4-Means Clustering

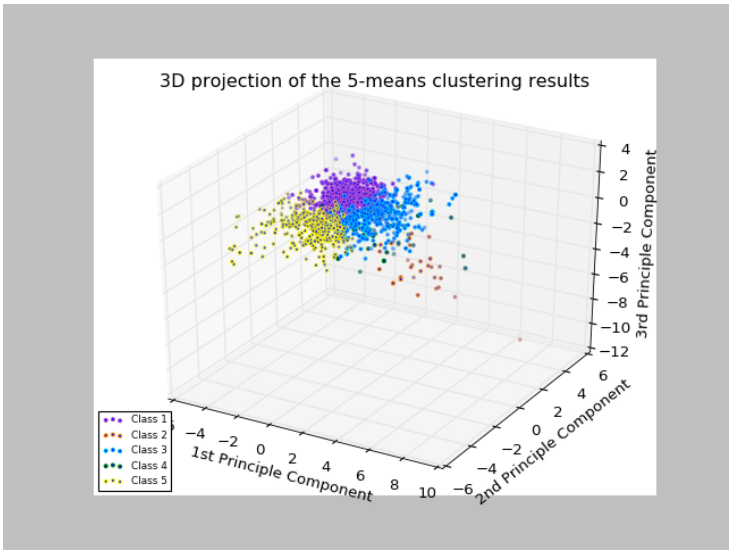

Figure 1.3.10: 5-Means Clustering

1.3.4 Conclusion

After doing the clustering with or without PCA, my conclusion is as follow. Comparing the knee point in Figure 1.4.1 and Figure 1.4.6. It is obvious that clustering with PCA reduce the quantization error. It might because of the reduction of dimension the PCA have simultaneously reduced the noise features.

The knee point in graph of quantization error can not be a judgement of result of clustering. So we see the Figure 1.4.8 to estimate if this clustering method is suitable to this data set. We can find that the data set be separated unambiguous. Thus, k-means clustering is suitable for this data set.

Chapter 2

If there are any attribute that influence total sulfur dioxide?

2.1 Labelling

Strategy of labelling the total sulfur dioxide (0,160], (160 440]

2.2 Visualization

2.2.1 Principal Component Analysis

Just like what I did in Section 1.3, I used the same technique to complete this part. The result is as follow. The first two feature affect to the data are fixed acidity and alcohol, which are 0.48809550387543194 and -0.5528903530300658

Top 2 Largest Eigenvalues and Eigenvectors			
1st largest: 3.12113839		2nd largest: 2.11175471	
Dimensions	Values	Dimensions	Values
fixed acidity	0.4880955038	alcohol	-0.552890353
citric acid	0.473432526	quality	-0.523604499
pH	-0.432604648	volatile acidity	0.398671979
density	0.369891526	density	-0.3785701792
volatile acidity	-0.265523183	'chlorides'	0.2027569580
sulphates	-0.254531216	citric acid	-0.169102659
chlorides	0.1971526382	sulphates	-0.163702380
residual sugar	0.1386021172	residual sugar	0.1017500272
quality	0.1131864476	pH	-0.0764579678
alcohol	-0.0724207390	free sulfur dioxide	0.05578872125
free sulfur dioxide	-0.04726047885	fixed acidity	0.0456072899

Table 2.3.1: Top two eigenvectors

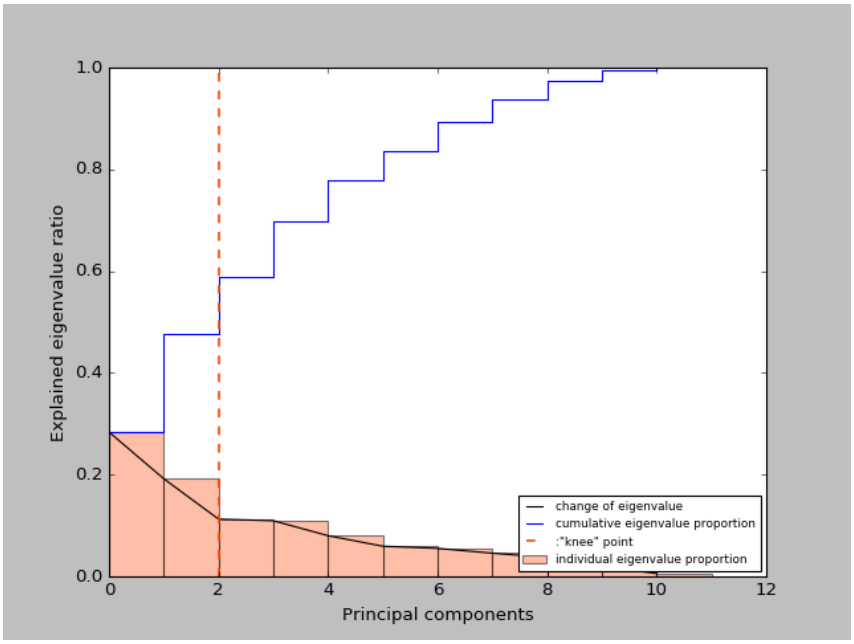

Figure 2.3.2: Cumulative eigenvalues of co-variance matrix

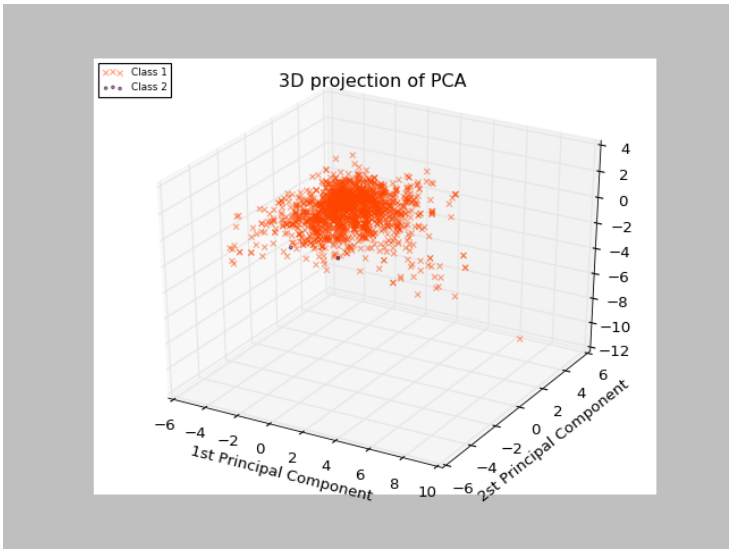

Figure 2.3.3: 3D PCA projection of total sulfur dioxide

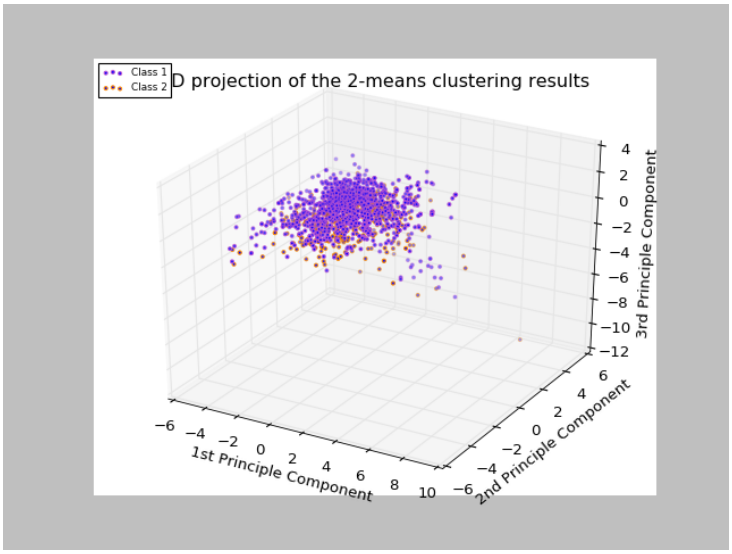

Figure 2.3.4: Clustering of total sulfur dioxide

2.3 Conclusion

As what I mention, The first two feature affect to the total sulfur dioxide are fixed acidity and alcohol, which are 0.48809550387543194 and -0.5528903530300658

Because of the limitation of working time and similarity of works, SOM and clustering only present one figure.

However, I cover all the requirements.

If it is possible, I am looking forward to be your students in the future.

Finally, I sincerely express my thanks here to GuoJi and Peter.

Code & Related Files:

`https://github.com/Voldet/data_analysis`