Danmarks Tekniske Universitet

DTU

02450 Introduction of machine learning and data mining–Project 1

Authors

Tianyi Ma - s210316
Hui Wang - s210331
Yun Ma - s202707

April 6, 2021

Contents

1 Description

Our data set is collected from UCI Machine Learning Repository [1], this data set used to propose an innovative real estate valuation approach to assume that the price per unit area of real estate is the average price per unit area of the particular circle of housing supply and demand multiplied by the product of several dimensionless adjustment coefficients of factors. The paper by Cheng et al [2] created a new approach called Quantitative Comparative Approach. They used four similar databases in that study, and we chose one of them to try to construct a similar relationship between house prices of unit area and influencing factors described in the paper.

This database is derived from the housing price market survey for Sindian Dist, New Taipei City, Taiwan area. The data set was randomly split into the training data set (2/3 samples) and the testing data set (1/3 samples). The database contains six attributes affecting house prices, transaction date, house age, distance from the surrounding public transportation network, number of surrounding shops, latitude and longitude location, etc. The last category is the transaction price per unit area of the house, which is what we are trying to predict.

We hope to perform supervised learning on the first six attributes in this database. Since the units of the six items are different, we will standardize them first, then perform prediction on the last attribute. And we are using observation value to predict a continuous response, it should be a regression problem. To transform it to be a classification problem, we are going to classify three different types of house price of unit area, it is basic on the 1/4 and 3/4 point, when the price is lower than the other three quarters we call it low price, as same, when the price is higher than three quarters we call it high price, the others is medium price. Using this way we can have an discrete result.

2 Explanation of the attributes

There are seven attributes in our data set, including six inputs and one output.

1. X1 (transaction date) is a Discrete and Interval attribute. For example, 2013.250 presents 2013 March, 2013.500 presents 2013 June, etc.
2. X2 (house age in year) is a Discrete and Ratio attribute.
3. X3 (the distance to the nearest MRT station in meter) is a Continuous and Ratio attribute.
4. X4 (the number of convenience stores in the living circle on foot) is a Discrete and Ratio attribute.
5. X5 (latitude of geographic coordinate in degree) is a Continuous and Interval attribute.
6. X6 (longitude of geographic coordinate in degree) is a Continuous and Interval attribute.
7. Y (the house price per area[1]) is a Continuous and Ratio attribute.

X2, X3, X4 and Y are Ratio attributes while the others are Interval. We said the attributes are Ratio because number 0 has a specific physical meaning for the variables. Taking House age as an example, a 0-year-old house is a new one, and the age of a house is 10 years means it is twice as older as the house with 5 years age. On the contrary, attributes like latitude and longitude are Interval, because the "Greenwich" or the equator are just made by convention, it makes no sense that 20 degree east longitude is twice as far east as 10 degree east longitude.

Above is a summary description of the attributes. Our data set has no missing values or corrupted data, so we won't cover this. Some summary statistics on the attributes has been conducted and the results can be seen in table 1

Table 1: Table of summary statistics

	X1	X2	X3	X4	X5	X6
count	414	414	414	414	414	414
mean	2013.148953	17.71256	1083.885689	4.094203	24.96903	121.533361
std	0.281995	11.392485	1262.109595	2.945562	0.01241	0.015347
min	2012.666667	0	23.38284	0	24.93207	121.47353
25%	2012.916667	9.025	289.3248	1	24.963	121.528085
50%	2013.166667	16.1	492.2313	4	24.9711	121.53863
75%	2013.416667	28.15	1454.279	6	24.977455	121.543305
max	2013.583333	43.8	6488.021	10	25.01459	121.56627

[1]The unit is 10000 New Taiwan Dollar/Ping, where Ping is a local area unit, 1 Ping = 3.3 square meter

In order to visualise the correlation of the variables in the data set, a correlation heat map has been generated, as shown in figure 1, where lighter colours means correlation and dark colours means no correlation. We can see that there are strong negative correlation between X3 (distance to the nearest MRT station) and X6 (longitude of geographic coordinate)

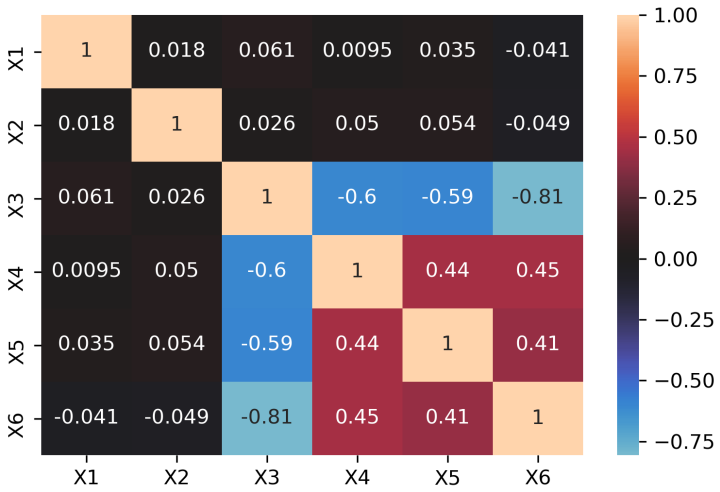

Figure 1: Heatmap of correlation

3 Data visualization and PCA

In order to make our data more visualized, we drew a box plot for seven attributes in the data set. Our box plot uses a standard definition. The red line in the middle of the box is the median of the data, and the triangle is the average. The left side is 1/4 point, and the right side is 3/4 point. Upper and lower whiskers are defined as the equation 1 and 2, the length of the whiskers is 1.5 times of the length of box. We judge that our data does not have missing data through the box plot, but we can see some outliers which is out of the whisker in the X3, X5 and X6 graphs. Since they are part of the truly data but not wrong data, we will not delete them.

$$Upperwhisker : min\left(x_{75} + \frac{3}{2}\left(x_{75} - x_{25}\right), v_n\right) \tag{1}$$

$$Lowerwhisker : max\left(x_{25} - \frac{3}{2}\left(x_{75} - x_{25}\right), v_1\right) \tag{2}$$

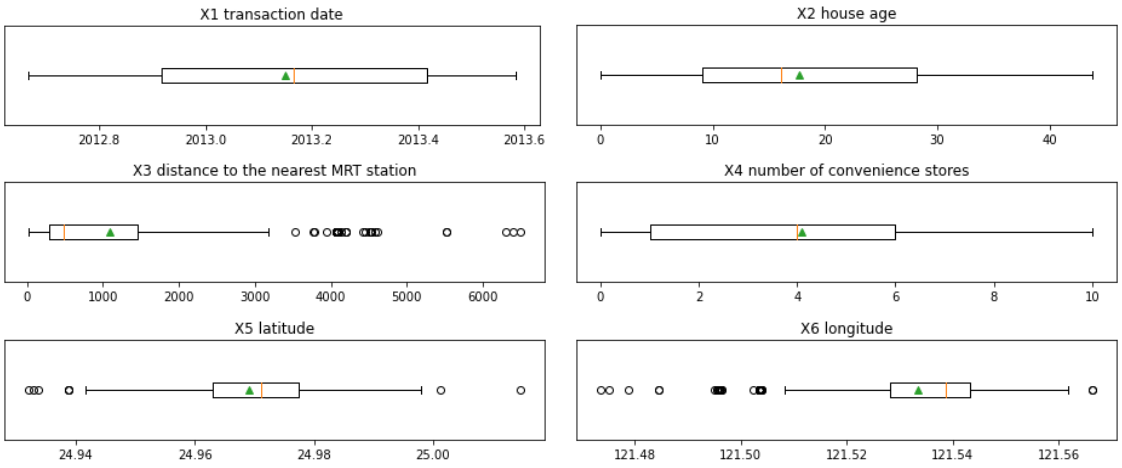

Figure 2: Boxplots of attributes

We create the density plots to check if our six attributes appear to be normal distributed, as the below figure 3. For X1 transaction date, it basically correspond to normal distribution, except for a relatively larger ratio on both ends. For X5 latitude, it can be regarded as a perfect normal distribution, and X6 longitude is very close to normal distribution. As for X2 house age, X3 distance to the nearest MRT station and X4 number for convenience stores, they are not very close to normal distribution, because they have a too large ratio for lower values, especially X3 distance to the nearest MRT station. This may because people prefer the houses nearby the station, so most of the house are built within lkm around the station. Same reason for X2 house age, 'younger' houses are more welcomed by the market, so the data is mostly concentrated in the front part.

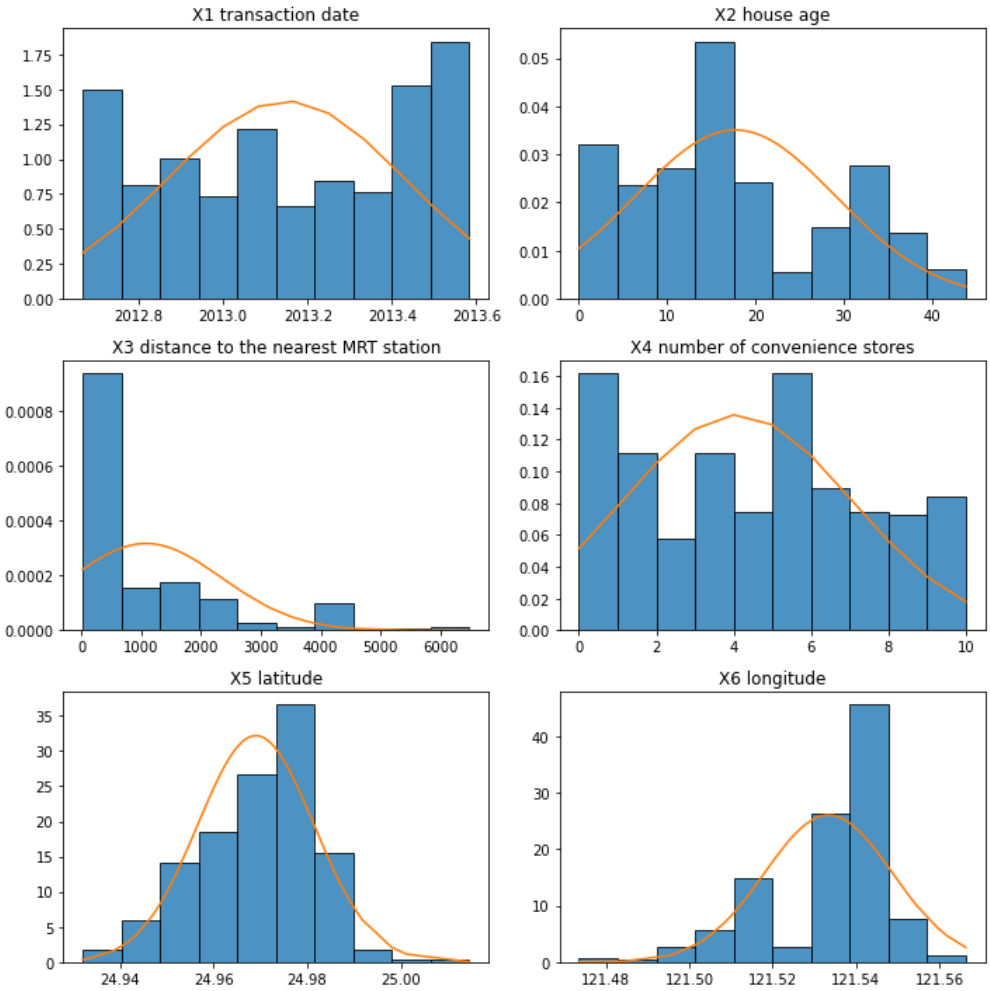

Figure 3: Density Histogram of all the attributes

Principal component analysis (PCA) is a data analysis tool to find a lower-dimensional representation of a high-dimensional data set. In this project, the singular value decomposition (SVD) was applied to find the principal component of our data set. To conduct PCA, the data was firstly centered by equation 3 [3].

$$\tilde{x_i} = x_i - m, \quad m = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3}$$

According to PCA theory, the attribute with large variance than others will have a great impact on first principle component. Meanwhile, the variance of an attribute is closely related to its scale. According to the statistic information of our data set in table 1, obviously, the attributes have very different scales, thus we need to normalize each attribute by equation 4 [3].

$$\tilde{x_{ij}} = \frac{x_{ij}}{s_k}, \quad s_k = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \tilde{x_{ik}}^2} \tag{4}$$

Lastly, the SVD was applied by equation 5 [3]. The columns of V are the principal components of the data set $\tilde{X}$. The element Σ_{ii} on the diagonal of Σ is the singular value of $\tilde{X}$.

$$\tilde{X} = U\Sigma V^T \tag{5}$$

The cumulative and individual variance explained by each principle component are shown in figure 4.

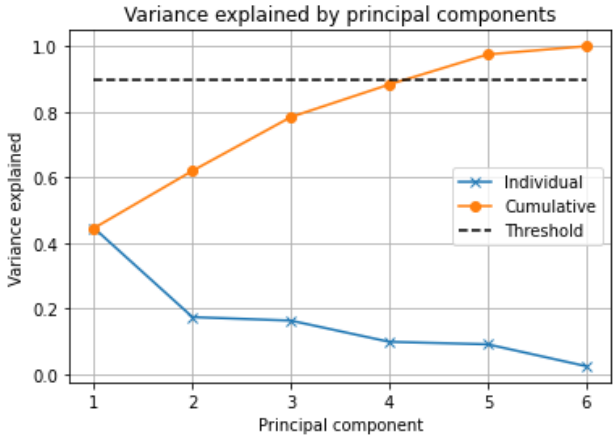

Figure 4: Cumulative and individual variance explained by each principle component

From figure 4, it is clear that the first four principal components explain nearly 90% of the data, and more than 90% of the variation is explained by the first five principal components. This means the data can be projected on to lower dimension and still keep most of the variance.

In the project, the first four of the principal components were chosen for further analysis and their coefficients for all attributes are shown in figure 5. For PC1, it can be found that it has a large coefficient for attributes X3 (distance to the nearest MRT station), which means X3 has a positive projection onto PC1. This indicates that PC1 mostly describes the variance from distance to the nearest MRT station. For PC2, it is clear that X1 (transaction date) and X2 (house age) have a large negative projection onto it. The PC3 also describes the attributes the same as PC2 but with different direction. And PC4 has more information about geographic coordinate, i.e. X5 (latitude) and X6 (longitude).

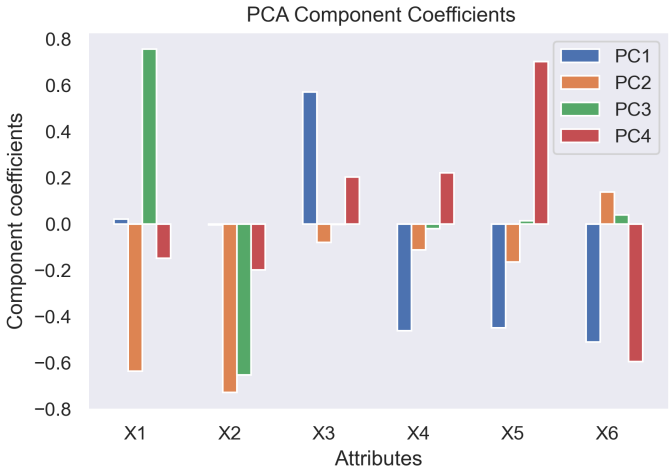

Figure 5: PCA Component Coefficients

Then let's consider the projections onto the first and second principal components. The vectors $\tilde{X}v1$ and $\tilde{X}v2$ was calculated and plotted as a scatter plot, as shown in figure 6). In the figure, some clusters of similar house price can be seen, which show that it is possible to classify the house price based on the data set. However, it is also clear that the separation between clusters is not significant, because the first two principal components may not enough to representing much variance. Thus, further analysis considering more principal components should be conducted.

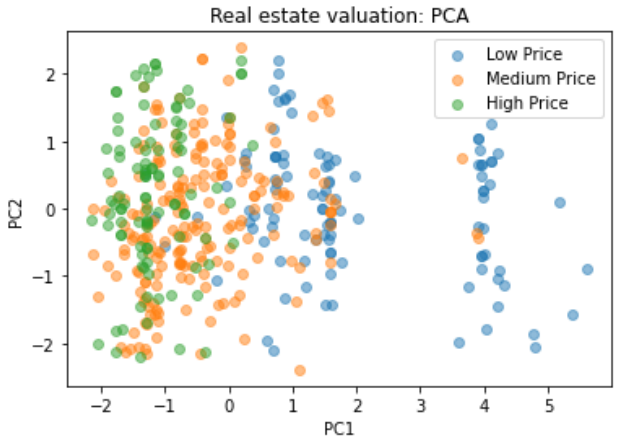

Figure 6: Data projected on PC1 and PC2

As known from figure 4, the first three principle components will explain about 20% more information than the first two principle components. In order to visualise more of the variance, the third principle component was included and a three dimensional scatter plot was shown as figure 7. As expected, there are similar groups as we found in figure 6. It seems that the extra dimension separates the Low Price (blue points) better, while the others are

still remain entwined. Therefore, as the first three principal components cannot fully explain all the data, we still cannot completely separate the data set.

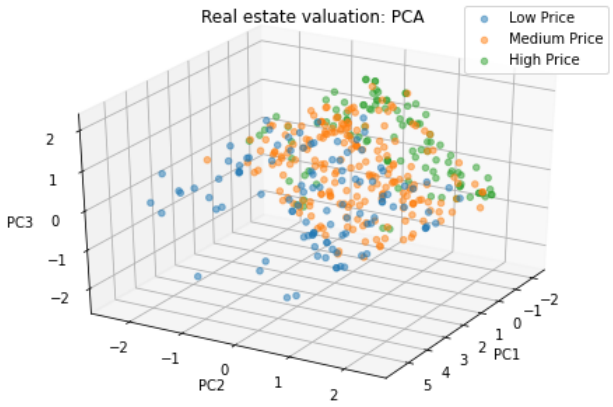

Figure 7: Data projected on PC1, PC2 and PC3

4 Conclusion

Firstly, after visualizing our data set, there are some outliers in the boxplot, however we believe the outlier is reasonable data and keep it according to the feature of the data set. Secondly, by analysing our data set we can find out the variance explained by the first four PCs sum up to 90%, it can be explained as each of the principal components carries only part of the information of the original data and can not explain most of the features of our data set. Based on PCA we can eliminate redundant information and keep most of the features. At the end, based on the projection information on each principal components, we can see there are some clusters of similar house price, it is possible to find a machine learning model to roughly distinguish between the different house price.

References

[1] "Real estate valuation data set Data Set.." `https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set#`. Accessed: Aug. 18, 2018.

[2] I.-C. Yeh and T.-K. Hsu, "Building real estate valuation models with comparative approach through case-based reasoning," *Applied Soft Computing*, vol. 65, pp. 260–271, 2018.

[3] T. Herlau, M. N. Schmidt, and M. Mørup, "Introduction to machine learning and data mining," *Lecture notes of the course of the same name given at DTU (Technical University of Denmark)*, p. 39, 2016.

5 Appendix

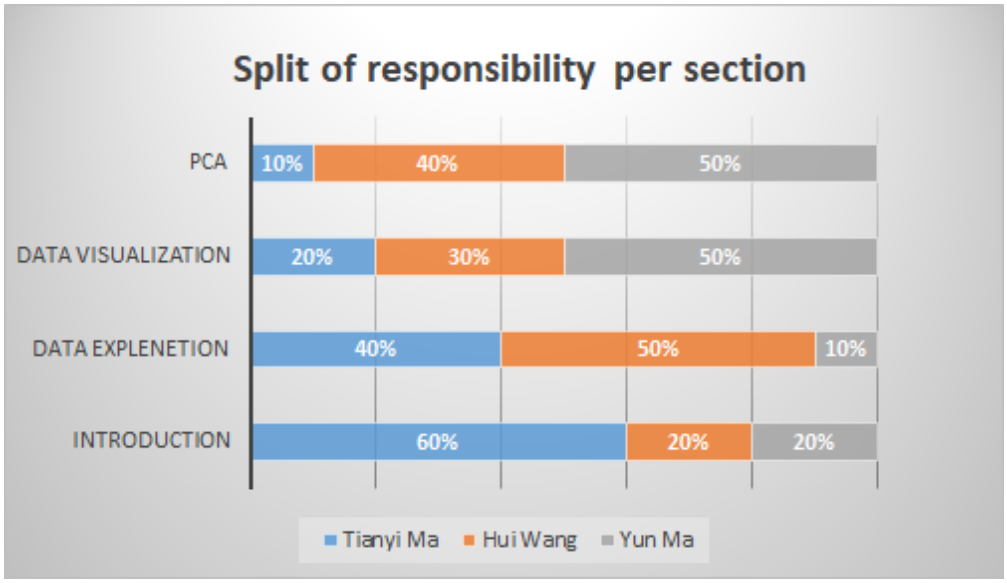

Figure 8: Split of responsibility per section