2019 Summer First Homework

Wei Guo

***Abstract*—This article is divided into two parts. In the first part, linear regression and polynomial regression models were built to fit the Boston house price dataset. The least squares method was used to solve the problem. The performance evaluation was carried out by MSE and R^2, and the results were compared and visualized. In the second part, the hand-drawn data set was reduced by PCA, and it was reduced from 64-dimensional to 3-dimensional and visualized. At this time, only 40% of the information was retained. Further, in order to retain 99% of useful information, we used PCA to reduce it to 41 dimensions, thus achieving the goal.**

***Index Terms*—linear regression, polynomial regression, PCA**

I. First Part

IN this section, 25% of the data in the Boston house price data set is selected as the sample, and the RM column (the number of rooms in each house) is used as the independent variable, and the house price is used as the dependent variable. Linear regression models, quadratic polynomials, fifth-order polynomials, and ten-order polynomials are used to fit them, respectively.

A. Fitting model

The formula for the model fit is as follows.

$$y = 8.27x - 29.52 \tag{1}$$

$$y = 2.5x^2 - 23.58x + 70.67 \tag{2}$$

$$y = -0.78x^3 + 17.08x^2 - 112.65x + 248.08 \tag{3}$$

$$\begin{aligned} y = & -0.48x^5 + 14.47x^4 - 169.94x^3 + \\ & 984.96x^2 - 2818.45x + 3199.97 \end{aligned} \tag{4}$$

$$\begin{aligned} y = & -0.0042x^{10} + 0.23^9 - 5.53x^8 + 76.18x^7 - \\ & 658.89x^6 + 3673.26x^5 - 12909.45x^4 + \\ & 25789.73x^3 - 1823396x^2 - 24984.63x + 41672.47 \end{aligned} \tag{5}$$

B. Evaluation index

This section uses MSE and R^2 to evaluate different fitting models. The results are as follows.

1) MSE: The mean square error in mathematical statistics refers to the expected value of the square of the difference between the parameter estimate and the true value of the parameter. It is recorded as MSE.MSE is a convenient method to measure the "average error". MSE can evaluate the degree of change of data, MSE The smaller the value, the better the accuracy of the predictive model describing the experimental data.

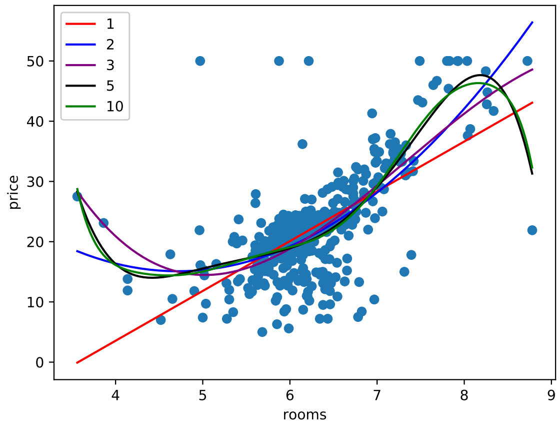

Fig. 1: House price fitting model map

2) R^2: The coefficient of determination (R^2) is used in statistics to measure the proportion of the variation of the dependent variable that can be interpreted by the independent variable, so as to judge the explanatory power of the statistical model.

C. Conclusion

TABLE I: Comparison of evaluation indicators

Module	MSE	R^2
degree = 1	45.24	41.6%
degree = 2	39.58	48.91%
degree = 3	38.33	50.52%
degree = 5	35.17	54.59%
degree = 10	35.06	54.75%

It can be seen that as the highest number of polynomials increases in sequence, the MSE gradually decreases, and the amount of information containing the original data is increasing. Since the Boston dataset used has 13 factors that affect house prices, using only one column of data as an independent variable does not work well.

II. Second Part

This section uses a handwritten data set, each image is 8 * 8 size, the corresponding data dimension is 64 dimensions. In order to be able to visualize it, we reduced it to 3 dimensions.

The analysis of the 3D data after PCA only contains about 40% of the original data, indicating that the PCA dimension reduction process has lost too much information. Therefore, in order to retain 99% of useful information, the PCA automatically selects which dimensions of information to retain, and finally finds that when retaining 41-dimensional information,

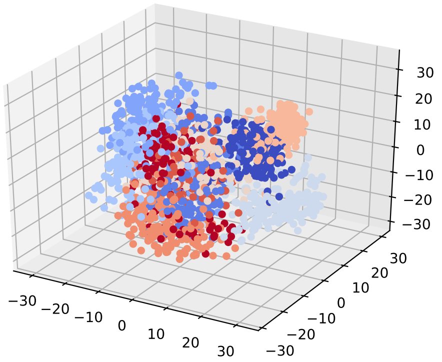

Fig. 2: Data points after dimensionality reduction

more useful information can be retained, and the dimension of the data amount is reduced to some extent. It is helpful to speed up the training process and prevent the occurrence of over-fitting.

References

[1] https://www.researchgate.net/publication/243771074
[2] https://blog.csdn.net/u013096666/article/details/72627001
[3] https://www.cnblogs.com/yifdu25/p/8330652.html
[4] https://liam.page/2014/09/08/latex-introduction/