House Price Prediction in Boston Based on BP Neural Network

Zuxin Chen *

Department of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, China

* Corresponding Author Email: 2022270205@email.szu.edu.cn

Abstract. House is a basic demand for living. It is directly related to people's happiness. Affected by many factors such as the economy condition and the policy, house prices are always dynamic. House price prediction can provide decision support for investors, so it is important to optimize the prediction model. Many machine learning methods are applied to house price prediction, such as support vector machine and random forest. This paper predicts the house price of Boston. To solve the multicollinearity problem in the dataset, this paper builds a BP neural network model using Bayesian regularization. The fitting analysis of the data samples shows the total R-value is 0.9824. The loss function converges to the global minimum within 100 epochs. The experiment indicates that the model is both efficient and reliable in forecasting house prices. This research provides a method to deal with the multicollinearity problem in the dataset, which improves the generalization ability and predictive accuracy of the model.

Keywords: BP neural network; house price; prediction.

1. Introduction

House prices have been a hot topic which the public constantly pay a close attention to. With people's pursuit of better live and the reform of the housing system, the real estate industry has developed rapidly. The policy of reducing or exempting property tax attracts many investors to choose house property for investment. However, it is both risky and of great opportunities [1]. Once an investment decision is made incorrectly, it can lead to extensive damage. Therefore, feasibility studies of real estate investment are an indispensable task before executing investment decisions [2]. Real estate market forecasting is an important part of feasibility studies [3]. The background of Boston's housing price forecast can be dated back to the 1970s. House prices in Boston fluctuated greatly at that time, which brought great challenges to real estate investment. To solve this problem, researchers all over the world had started using machine learning algorithms to predict Boston house price. For instance, Zekun Chen and Xiaorong Cheng used gradient descent algorithm for regression analysis of datasets [4], Jia Sheng and Dongdong Pan used enhanced regression tree method to study the main factors affecting house prices [5], Adetunji AB, Akande ON, Ajala FA et al used random forests to predict house prices [6]. Rather than linear regression models, the BP model can handle many nonlinear relationships in house price prediction and have more accurate forecast results. In this paper, the Bayesian regularization is employed in the BP neural network. The performance of the model is discussed through the analysis of the loss function curve chart and line graph of forecast results.

2. Proposed method

2.1. Data source

The Boston housing data used in this experiment was collected by Harrison and Rubenfeld in 1978. The data shows the median housing prices in suburban Boston [7]. There are a total of 506 observations. Each data point has 14 attributes, including multiple metrics such as urban per capita crime rate and residential land ratio.

2.2. Data normalization

The z-score method can be used to normalize the dataset. The processed data conforms to a normal distribution [8]. It makes the comparison between different variables more intuitive. The normalization formula is shown:

$$\hat{x}=\frac{x-\mu}{\sigma}. \tag{1}$$

Where μ and σ represent the mean and the standard deviation of the data respectively.

2.3. Data visualization and correlation analysis

It is necessary to carry on correlation testing before constructing the neural network, as the correlation between variables can affect their weights [9]. It results in an increase in the estimated variance of model parameters, thus reducing the stability of the model. The presence of multicollinearity between variables can be measured by using heatmaps to represent the correlation matrix of variables or calculating the variance inflation factor (VIF). Specifically, VIF is obtained by calculating the ratio of the variance of each variable in the model to the variance of that variable in a linear combination of other variables. The formula is as follows:

$$VIF_i=\frac{1}{1-R_i^2} \quad (i=1,2,\ldots,k) \tag{2}$$

Where R_i is the complex correlation coefficient between the i-th variable X_i and all other variables X_j ($i=1,2,\ldots,k;\ i\neq j$).

When the VIF of a variable is greater than 10, it can be considered highly correlated with other variables [10]. When multicollinearity occurs, algorithms such as L-M optimization algorithm (trainlm) or Bayesian regularization algorithm (trainbr) can be used as neural network training methods. By adding regularization terms to the target function to limit the fluctuation of model coefficients, the problem of increased variance in model parameter estimation caused by collinearity can be effectively solved [11].

2.4. Model building

BP neural network is a method of monitoring learning [12]. The topology of the BP neural network is shown in Fig. 1. The training process of the network includes forward propagation and backpropagation [12]. The input signal is received and processed during the transport process among these three layers. The activation functions perform nonlinear processing on the input signal, which enhance the ability of the network to handle nonlinear tasks. If the output differs significantly from the expected value, the error will be returned along the original path using backpropagation [12]. The neurons in the neural network will adjust their weights based on the returned error and the error is reduced [12]. When the output is close enough to the expected value, the neural network completes one iteration. When all samples in the dataset have been trained once by the neural network, the neural network completes one epoch. This is the difference between iteration and epoch in the concept of neural networks.

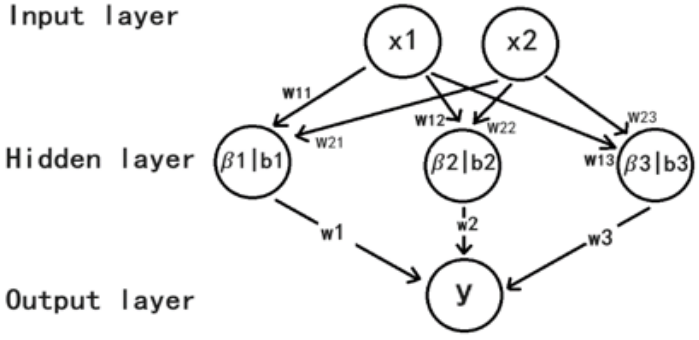

Fig 1. The Structure of BP Neural Networks.

This paper adopts the four-layer BP network structure based on data sample features. In the input layer and output layer, there are 13 neurons and 1 neuron respectively. The activation function of hidden layer nodes and output layer nodes are tansig function and linear purelin function respectively. The number of hidden layer nodes should be taken into serious consideration. If it is too small, it results in insufficient training of the neural network and low accuracy of the training results. If it is too large, it can reduce efficiency and result in overfitting. According to the empirical formula:

$$n=\sqrt{n1+n2}+C. \tag{3}$$

Where n, n1, n2 respectively stand for the number of neurons in the input layer, hidden layer, and output layer. C is an integer between 1 and 10.

By continuously adjusting the size of C and comparing the predictive performance of the model, n is recorded as 12 in this experiment.

After setting up the model, it can be used for training datasets.

3. Discussion and analysis of the results

This experiment divided the dataset randomly. 426 data samples were divided into the train set and 80 data samples were divided into the test set. The network was set up using the tool of Matlab. The maximum of epochs was 1000. The Bayesian regularization was selected as the learning method for the network. R value, mean square error (MSE) and predicted result line chart were used as indicators to evaluate model performance. The R value is used to measure the linearity between the predicted output and the target output. The more the R value is close to 1 or -1, the stronger the linearity between the predicted output and the objective output is. Conversely, the more the R value is close to 0, the stronger randomness between the two variables is.

The heatmap of correlation matrix reflects the relationship between variables. As shown in Fig. 2:

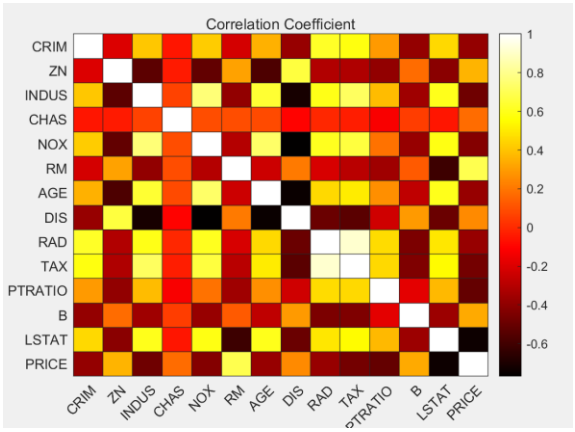

Fig 2. Heat map of correlation matrix.

In Fig. 2, the closer the color is to reddish brown or light yellow, the stronger the correlation is. It can be clearly observed that the correlation between the variables is strong.

The computed results of VIF for each variable are shown in Table 1:

Table 1. VIF values of each variable.

feature	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
VIF	2.100373	2.844013	14.48576	1.152952	73.89495	77.94828	21.38685
feature	DIS	RAD	TAX	PTRATIO	B	LSTAT	
VIF	14.69965	15.16773	61.22727	85.02955	20.10494	11.10203	

Since there are 10 features with VIF values greater than 10, there are 10 features with multicollinearity problem.

The results of the R coefficient, MSE and line graph of forecast results are shown in Fig. 3, Fig. 4 and Fig. 5.

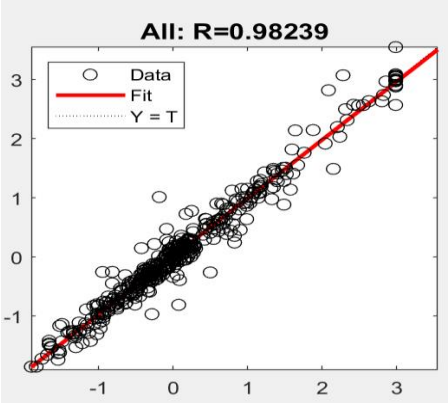

Fig 1. Overall R values.

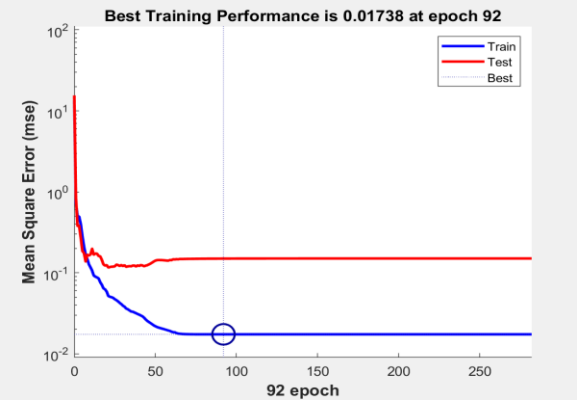

Fig 2. MSE chart.

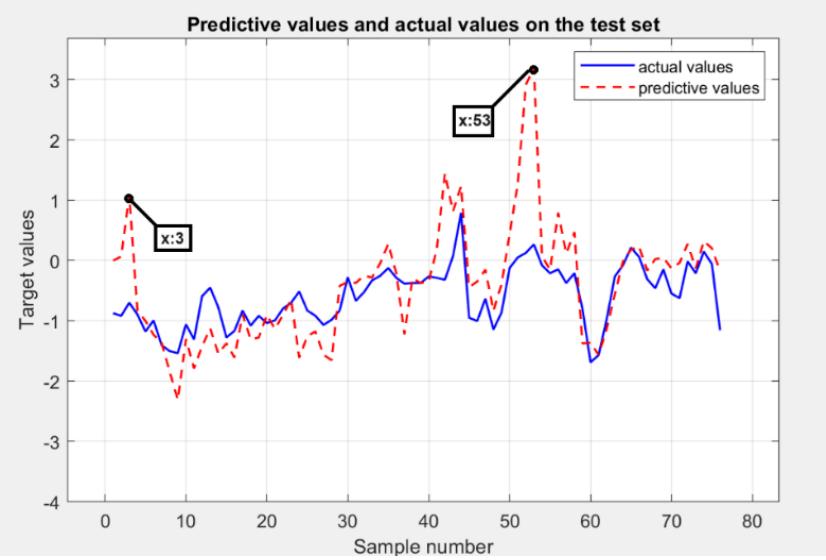

Fig 3. Line graph of forecast results.

According to the fitting analysis in Fig. 3, the total R-value is 0.98239. It indicates good overall prediction performance.

From Fig. 4, it can be observed that the MSE of both the test set and the training set rapidly decreases within 50 epochs and reaches optimal performance within 100 epochs. It demonstrates that the training efficiency of the model is considerable.

From Fig. 5, it can be seen that the predicted values roughly match the actual values. Although some data points have significant errors such as the third point and the 53rd point. Basically, the prediction accuracy of this model is good.

4. Conclusion

This experiment uses a Bayesian optimized BP neural network to predict housing prices. Although the dataset has multicollinearity issues, the results of R value, MSE and line graph of forecast results can verify that the model has addressed this problem and has good training efficiency and stability. The application of machine learning in housing prices is widespread and practical. However, it still has some deficiency such as overfitting and underfitting. The optimization process of the model is very complex, which requires lots of experiments and testing. Further optimization of the model can be considered from neural network structure design, neural network training methods and parameter settings such as learning rate. This experiment uses regularization methods to handle multicollinearity problem, which can also be addressed by using principal component analysis. Further research can apply principal component analysis to improve this experiment. Perhaps the combination of regularization methods and principal component analysis will make a difference.

References

[1] Tu Dongxiao. Risks and opportunities coexist in the real estate industry. New Finance, 2004, 12: 84-85.

[2] Yu Lijun. On the feasibility study of real estate investment projects. Journal of Changchun Institute of Technology (Social Science Edition), 2004, 5(1): 44-46.

[3] Qu Fuqiang, Qu Yingfang. Comparison and analysis of feasibility studies of real estate development projects and general construction projects. Infrastructure Optimization, 2003, 24(6): 38-39.

[4] Zekun Chen, Xiaorong Cheng. House price regression analysis and prediction based on gradient descent algorithm. Information Technology and Informatization, 2020, 4(5): 10-13.

[5] Sheng Jia, Pan Dongdong. Analysis of factors affecting housing prices based on boosted regression tree: A case study of Boston area. Statistics and Application, 2016, 5(3): 299-304.

[6] Adetunji A B, Akande O N, Ajala F A, et al. House price prediction using random forest machine learning technique. Procedia Computer Science, 2022, 199: 806-813.

[7] Zhao Ran. Analysis of the Correlation between Housing Price Data in Boston Based on the Regression Method. Statistics and Application, 2020, 9(3): 335-344.

[8] LI Z W, ZHENG W, FANG J, et al. Optimizing suitability area of underwater gravity matching navigation based on a new principal component weighted average normalization method. Chinese Journal of Geophysics, 2019, 62(9): 3269-3278.

[9] Bai Haiqing, Chen Xiaoyong. House Price Forecasting in Ames Based on Bayesian Regularized BP Neural Network. Automation and Machine Learning, 2023, 4(1): 17-23.

[10] Stine R A. Graphical interpretation of variance inflation factors. The American Statistician, 1995, 49(1): 53-56.

[11] GU YW G, ZHANG L. Regularization Based on Diagnosis of Multicollinearity. Journal of Information Engineering University, 2007, 8(8): 497-500.

[12] Li J, Cheng J, Shi J, et al. Brief introduction of back propagation (BP) neural network algorithm and its improvement. Advances in Computer Science and Information Engineering, 2012, 2: 553-558.