Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data

Keren He* and Cuiwei He†
*Institute of Integrated Science and Technology
Hosei University, Japan
†School of Information Science
Japan Advanced Institute of Science and Technology, Japan
Emails: *keren.he.6r@stu.hosei.ac.jp, †cuiweihe@jaist.ac.jp,

2021 IEEE International Conference on Computing (ICOCO) | 978-1-6654-3689-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOCO53166.2021.9673533

***Abstract*—The use of machine learning aided techniques to analyze real estate data is emerging as a trending research topic and has attracted a lot of interests from both industry and academia. In this paper, both linear regression and logistic regression algorithms are comprehensively reviewed and used to analyze realistic housing price data from the Melbourne real estate market. The trials of the learning process show that the optimum model coefficients which minimize the error function can always be efficiently obtained. The results demonstrate that the trained linear regression model can very accurately predict the average housing price. Also, using the logistic regression algorithm, the houses sold at different councils are precisely classified into different categories.**

***Index Terms*—Data Science, Housing price model, Linear regression, Logistic regression**

I. Introduction

It is well recognized that changes in residential housing prices can dramatically affect many aspects of our modern economy. An accurate prediction model of the housing price as well as the related beneficial data analysis are in the interest of all real estate market participants such as home sellers, home buyers, investors, banks, builders and also the government. Melbourne has been recognized as the world's most liveable city according to the Economist Intelligence Unit and its population has been stably growing for many years by attracting new migrants from all over the world. As a result, the housing price in Melbourne is continuously increasing over the past few decades and buying houses is usually considered to be the most secure and profitable way of investment [1].

In recent years, thanks to the massive explosion in data generation, artificial intelligence (AI) driven decision making is increasingly becoming popular. Machine learning based algorithms have been widely used in data analysis [2] and their related powerful applications have been broadly recognized in both industry and academia [3], [4]. Using machine learning aided techniques to analyze and/or predict housing prices have naturally become a popular research and study field [5], [6].

*The work of Keren HE was performed when she was with Guangxi University of Science and Technology.

Linear regression and logistic regression are two well-known machine learning algorithms which are typically classified as supervised learning techniques [7], [8]. In the supervised learning process, a dataset that contains the desired outputs is first used for model training and the well-trained system is then capable of processing new inputs to generate accurate outputs. Linear regression is commonly used for solving regression problems whereas logistic regression is usually used for data classification [8]. Although housing price data has been very often used to explain the mechanisms behind linear prediction algorithms, partially the linear regression method, in many online data science forums or lectures [9], these examples are usually either lack of technical details or using unrealistic datasets.

In this paper, the technical details of both the linear regression and the logistic regression are explained thoroughly. Moreover, realistic datasets from the Melbourne real estate market are used to verify the performance of the trained models. In the analysis, the housing price dataset of 2017 is used for model training and the dataset of 2018 is used for testing purposes. In the first part of the paper, the linear regression algorithm is used to model the relationship between the price of the house and its distance to the Melbourne city center. It shows that a simple linear model can accurately predict the average sold price of the houses if the training dataset contains enough data. In the second part of the paper, the logistic regression model is used to classify the sold houses which belong to different councils based on both the price and the distance information. It shows that when the considered two councils are far away from each other, the classification accuracy is 100%. When the considered two councils are close to each other, the classification accuracy is still above 85%.

The rest of this paper is organized as follows. The dataset used in the analysis and its preprocessing are first discussed in Section-II. Section-III and Section-IV then describe the studied linear regression and logistic regression models in detail. Next, in Section-V and Section-VI, both the training process and the testing results are presented. Finally, Section-VII concludes this paper.

II. Dataset and its preprocessing

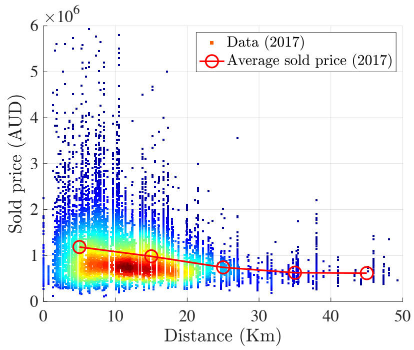

Fig. 1. The prices of the houses sold at Melbourne in 2017 (20261 examples) and the associate distance information.

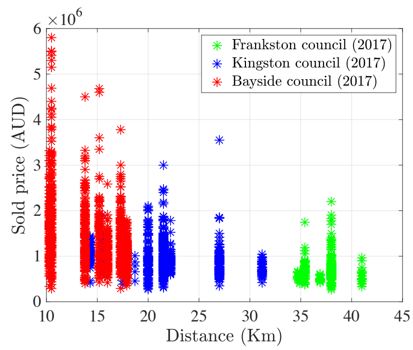

Fig. 2. The prices of the houses sold at Melbourne in 2017 in the Bayside city council (1030 examples), the Kingston city council (812 examples) and the Frankston city council (812 examples).

The housing dataset used in this paper is obtained from the online machine learning community Kaggle.com and it can be found in [10]. This dataset includes a wide range of features of all sold houses in Melbourne, Australia. These features include the sold price, the number of rooms, the sold date, the postcode, the distance to the center of the city, the suburb information and the council information. In this paper, without loss of generality, the data of the houses sold in 2017 is used for model training and the data of 2018 is used for testing purposes. In practice, although the price of one house depends on all these features, the dominating feature is usually its location which can be represented by 1. the house's distance to the center of the city and 2. the council (or suburb) which the house belongs to. The distance information is very important because it is usually related to the convenience of the transportation. The council is normally associated with the quality of public facilities (such as schools, hospitals) and therefore also crucially affects the price of the house. In this paper, the distance data and the council information are considered and their relationship to the price of the house is analyzed using both linear regression and logistic regression models. In the analysis, in order to establish a more accurate model, the dataset is first preprocessed by excluding the 'outlier' examples of which the sold prices

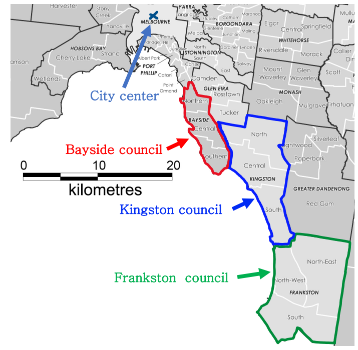

Fig. 3. The map of Melbourne and the locations of the Bayside council, the Kingston council and the Frankston council.

are above 6,000,000 dollars and the distances to the city are beyond 50 km. After the preprocessing, the dataset of 2017 contains 20261 examples and the dataset of 2018 contains 15066 examples.

Fig. 1 shows the prices of the houses sold in 2017 and their associated distance information. It can be seen that most of the sold houses are located at distances to the city center between 5km and 20km and their sold prices are between 500,000 dollars to 1,500,000 dollars. The average sold price is plotted as a function of the distance and also shown in Fig. 1. In the first part of this paper, a linear regression model is established and used to predict the average sold prices of houses located at different distances.

In the second part of this paper, the differences between the houses sold in different councils are analyzed. In the analyzed dataset, besides the distance and the price information, it also contains the council information of all sold houses. It is very interesting to notice that, when the houses are located at different councils, both the distance and the price information are very distinguishable. In this paper, it is shown that, by using a practical logistic regression model, the sold houses can be successfully classified into different categories with each category only contains the houses that belong to the same council. Fig. 2 shows the prices of the houses sold in the Bayside council, the Kingston council and the Frankston council in 2017. The locations of these three councils are shown on the map in Fig. 3. It can be seen that the Bayside council is relatively close to the city and thus the prices of many sold houses are much higher than those sold in the Kingston council and the Frankston council. It also can be seen that, because the Bayside council and the Kingston council are neighbours, the distance data examples are 'overlapped'. The Frankston council is located far away from the city center and therefore the housing price is much lower compared to the Bayside council and the Kingston council.

III. Linear regression model

In the first part of this paper, a linear regression based machine learning algorithm is used to predict the average price of the house based on its distance information. The detail steps of this method are described in this section. In this method, the relationship between the predicted housing price, $\bar{y}$, and the distance of the house to the center of the city, x, is modelled using

$$\bar{y} = \beta x + \alpha \tag{1}$$

where α is the bias of the linear model and β is the slope coefficient (or the weighting factor). In practical situations, based on the model described in (1), there are differences between the predicted housing price and individual sold prices. In this paper, the well-known mean squared error (MSE) is used as a performance metric to evaluate the accuracy of this linear regression model and it is calculated using

$$e_{\text{MSE}} = \frac{1}{2L}\sum_{l=1}^{L}(\bar{y}_l - y_l)^2 \tag{2}$$

where y_l is the lth element of the dataset vector, $\mathbf{y} = [y_1, y_2, ..., y_l, ..., y_L]$, which contains the actual sold prices of all L houses and $\bar{y}_l$ is the related predicted housing price. When this linear modelled is well trained, e_{MSE} should be minimized.

In the model training process, in order to efficiently obtain the values of α and β which can minimize e_{MSE}, a gradient descent based algorithm [11] is considered in this paper. In this approach, the gradients of α and β are calculated using partial derivatives as

$$\frac{\partial e_{\text{MSE}}}{\partial \alpha} = \frac{\partial e_{\text{MSE}}}{\partial \bar{y}_l}\frac{\partial \bar{y}_l}{\partial \alpha} = \frac{1}{L}\sum_{l=1}^{L}(\bar{y}_l - y_l) \tag{3}$$

and

$$\frac{\partial e_{\text{MSE}}}{\partial \beta} = \frac{\partial e_{\text{MSE}}}{\partial \bar{y}_l}\frac{\partial \bar{y}_l}{\partial \beta} = \frac{1}{L}\sum_{l=1}^{L}(\bar{y}_l - y_l)x_l \tag{4}$$

In this case, (3) and (4) indicate whether the values of α and β need to be increased or decreased in order to minimize the mean squared error. In the learning procedures, α and β are first initialized with random values and then updated iteratively. In each iteration, α and β are updated simultaneously using

$$\alpha_{\text{updated}} = \alpha - \eta\frac{\partial e_{\text{MSE}}}{\partial \alpha} = \alpha - \frac{\eta}{L}\sum_{l=1}^{L}(\bar{y}_l - y_l) \tag{5}$$

and

$$\beta_{\text{updated}} = \beta - \eta\frac{\partial e_{\text{MSE}}}{\partial \beta} = \beta - \frac{\eta}{L}\sum_{l=1}^{L}(\bar{y}_l - y_l)x_l \tag{6}$$

where η is the learning rate. The steps of the gradient descent based updating process can be summarized using the simplified pseudo code listed in the Table-I.

TABLE I
THE PSEUDO-CODE OF THE UPDATING PROCESS OF α AND β

Input: $\mathbf{x} = [x_1, x_2, ..., x_L]$, $\mathbf{y} = [y_1, y_2, ..., y_L]$
Output: α, β
Randomly generate initial α
Randomly generate initial β
Set η
while $i < I_{\text{iteration}}$ **do**
    $\bar{\mathbf{y}} = \beta\mathbf{x} + \alpha$
    $\alpha \leftarrow \alpha - \frac{\eta}{L}\sum_{l=1}^{L}(\bar{y}_l - y_l)$
    $\beta \leftarrow \beta - \frac{\eta}{L}\sum_{l=1}^{L}(\bar{y}_l - y_l)x$
end while

IV. LOGISTIC REGRESSION MODEL FOR DATA CLASSIFICATION

In the second part of this paper, a practical logistic regression model is studied to classify the information obtained from the dataset into different categories. As discussed in Section-II, the data of the sold houses located in three different councils is used in the analysis and the performance of the classification algorithm is validated by considering two scenarios. In the first scenario, the algorithm classifies the data between the Bayside council and the Frankston council and these two councils are far apart from each other as shown in Fig. 3. In the second scenario, the house data of the Bayside council and the Kingston council is analyzed. This scenario is more challenging since these two councils are neighbours and their data is more similar compared to the first scenario. In both scenarios, the data needs to be classified into two categories and $c \subset \{0, 1\}$ is used to indicate the data category. In the first scenario, $c = 1$ means the Bayside council and $c = 0$ means the Frankston council. In the second scenario, $c = 1$ indicates the Bayside council and $c = 0$ indicates the Kingston council. In this following, the details of this model are described.

First, similar to the linear regression model, the values of the input features (e.g. the distance and the price) are normalized and then combined using a linear equation by

$$z = \gamma y + \beta x + \alpha \tag{7}$$

where y is the normalized sold price, x is the normalized distance information. γ and β are the weighting coefficients and α is the bias factor. Note that, in the linear regression model, x is the only input and the predicted house price, $\bar{y}$, is the output. In the logistic regression analysis discussed in this section, both x and y are the input features and the final output is the estimated council information, $\bar{c}$. To estimate c, in the next step, z is input into a logistic function using

$$P = \frac{1}{1 + \exp(-z)} \tag{8}$$

which is plotted in Fig. 4. The objective of using the logistic function is that the value of P is constrained to be within 0 and 1 and it can be used to indicate the probability that the predicted council belongs to the first category ($c = 1$) [8]. Accordingly, $1 - P$ indicates the probability that the predicted

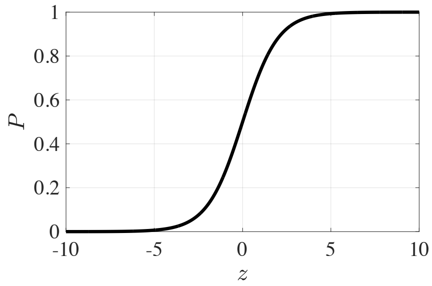

Fig. 4. The logistic function used in the data processing step

council is in the second category ($c = 0$). In the prediction stage when α, β and γ are already updated with the optimum values, the decision is made using

$$\bar{c} = \begin{cases} 1, & \text{if } P \geq 0.5 \\ 0, & \text{if } P < 0.5 \end{cases} \tag{9}$$

In the machine learning process, an error function need to be first defined and then minimized to find the best values of α, β and γ. In this case, unlike the linear regression model, the mean squared error can no longer form a convex problem and therefore the gradient descent method would not work [8]. In order to form a convex optimization problem, an error function of the predicted probability is defined as [8]

$$e_{\text{p}} = \begin{cases} -\dfrac{1}{L}\displaystyle\sum_{l=1}^{L} \log(P_l), & \text{if } c = 1 \\ -\dfrac{1}{L}\displaystyle\sum_{l=1}^{L} \log(1 - P_l), & \text{if } c = 0 \end{cases} \tag{10}$$

where L is the number of examples within the used dataset. For example, based on (10), when the data belongs to the first category, (e.g. $c = 1$) and the predicted probability that the data belong to the first category is one, (e.g. $P = 1$), the error coefficient, e_{p}, is zero. Similarly, when the data belongs to the second category, (e.g. $c = 0$) and the predicted probability that the data belong to the first category is zero, (e.g. $P = 0$), the error coefficient, e_{p}, is also zero based on (10). Therefore, e_{p} indicates the error of the predicted probability. To simply the analysis, (10) is rewritten as

$$e_{\text{p}} = -\frac{1}{L}\sum_{l=1}^{L} \big(c_l \log P_l + (1 - c_l)\log(1 - P_l)\big) \tag{11}$$

Next, similar to the calculations in Section-III, in order to efficiently obtain the values of α, β and γ which can minimize the error factor, e_{p}, a gradient descent based algorithm is used. In this method, the gradients of α, β and γ are determined using partial derivative calculations. The gradient of α is

$$\frac{\partial e_{\text{p}}}{\partial \alpha} = \frac{\partial e_{\text{p}}}{\partial P_l}\frac{\partial P_l}{\partial z}\frac{\partial z}{\partial \alpha} \tag{12}$$

where

$$\begin{aligned} \frac{\partial e_{\text{p}}}{\partial P} &= -\frac{1}{L}\sum_{l=1}^{L} \frac{\partial\big(c_l \log P_l + (1 - c_l)\log(1 - P_l)\big)}{\partial P_l} \\ &= -\frac{1}{L}\sum_{l=1}^{L}\Big(\frac{c_l}{P_l} - \frac{1 - c_l}{1 - P_l}\Big) \end{aligned} \tag{13}$$

$$\frac{\partial P_l}{\partial z} = P_l(1 - P_l) \tag{14}$$

$$\frac{\partial z}{\partial \alpha} = 1 \tag{15}$$

Substituting (13) - (15) into (12) gives

$$\frac{\partial e_{\text{p}}}{\partial \alpha} = \frac{1}{L}\sum_{l=1}^{L}(P_l - c_l) \tag{16}$$

Using similar calculations, the gradients of β and γ are

$$\frac{\partial e_{\text{p}}}{\partial \beta} = \frac{1}{L}\sum_{l=1}^{L}(P_l - c_l)x_l \tag{17}$$

and

$$\frac{\partial e_{\text{p}}}{\partial \gamma} = \frac{1}{L}\sum_{l=1}^{L}(P_l - c_l)y_l \tag{18}$$

Finally, α, β and γ are updated simultaneously using

$$\alpha_{\text{updated}} = \alpha - \eta\frac{\partial e_{\text{p}}}{\partial \alpha} = \alpha - \frac{\eta}{L}\sum_{l=1}^{L}(P_l - c_l) \tag{19}$$

$$\beta_{\text{updated}} = \beta - \eta\frac{\partial e_{\text{p}}}{\partial \beta} = \beta - \frac{\eta}{L}\sum_{l=1}^{L}(P_l - c_l)x_l \tag{20}$$

$$\gamma_{\text{updated}} = \gamma - \eta\frac{\partial e_{\text{p}}}{\partial \gamma} = \gamma - \frac{\eta}{L}\sum_{l=1}^{L}(P_l - c_l)y_l \tag{21}$$

where η also denotes the learning rate. It is very interesting to notice that (19)-(21) have identical form with (5) and (6) used in the linear regression model. The procedures of this logistic regression method is summarized in the Table-II.

V. Linear regression results

In this section, the performance of the described linear regression method is studied. First, multiple trials of system training process are used to explain how the gradient descent based algorithm can effectively find the optimum coefficients of the linear model. Then, the testing results are used verify the accuracy of the prediction housing price based on the distance information.

A. The trained model

Fig. 5 shows the mean squared error, e_{MSE}, plotted as a function of α and β based on the data of 2017 shown in Fig. 1. The black curves are the associated contour plots. On top of this 2-D plot, the red line indicates the learning process based

TABLE II
THE PSEUDO-CODE OF THE UPDATING PROCESS OF α, β AND γ

Input: $\mathbf{x} = [x_1, x_2, ..., x_L]$, $\mathbf{y} = [y_1, y_2, ..., y_L]$, $c = [c_1, c_2, ..., c_L]$ $(c \subset \{0, 1\})$
Output: α, β, γ

Randomly generate initial α
Randomly generate initial β
Randomly generate initial γ
Set η
while $i < I_{\text{iteration}}$ **do**
    $\mathbf{P} = \dfrac{1}{1 + \exp(-(\gamma\mathbf{y} + \beta\mathbf{x} + \alpha))}$
    $\alpha \leftarrow \alpha - \frac{\eta}{L}\sum_{l=1}^{L}(P_l - c_l)$
    $\beta \leftarrow \beta - \frac{\eta}{L}\sum_{l=1}^{L}(P_l - c_l)x_l$
    $\gamma \leftarrow \gamma - \frac{\eta}{L}\sum_{l=1}^{L}(P_l - c_l)y_l$
end while

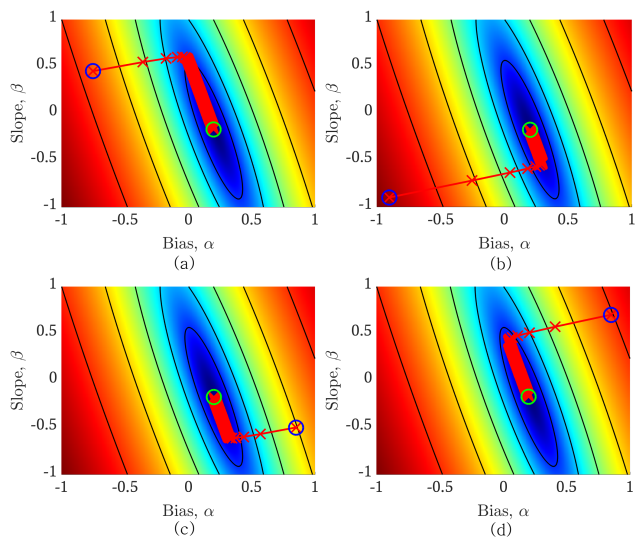

Fig. 5. The mean squared error, e_{MSE}, plotted as a function of α and β and the associated contour plot. The learning procedures based on the studied gradient descent algorithm in four different learning trials when the initial values of α and β were set randomly and $\eta = 0.5$ are indicated using red lines. The blue circle indicates the 'start' point of the learning process and the green circle is the 'end' point of the learning process.

on the studied gradient descent algorithm. It can be seen from the four different learning trails that, regardless of the initial values of α and β, the algorithm can always efficiently locate the values of α and β which are associated with the minimum mean squared error. It also can be seen that the learning curve is always orthogonal to the contour lines. When the learning rate is fixed at a certain value (e.g. 0.5 in this figure), the initial 'steps' in the learning process is relatively large. Then, when the gradient of the e_{MSE} becomes small, the learning steps also become small. It shows in Fig. 5 that, based on the data of 2017, α becomes 0.2012 and β becomes -0.1736 when the training process is completed.

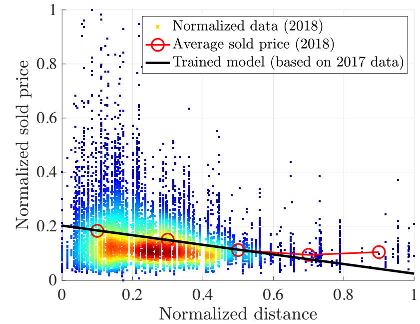

Fig. 6. The predicted average prices of the house using the studied linear regression model.

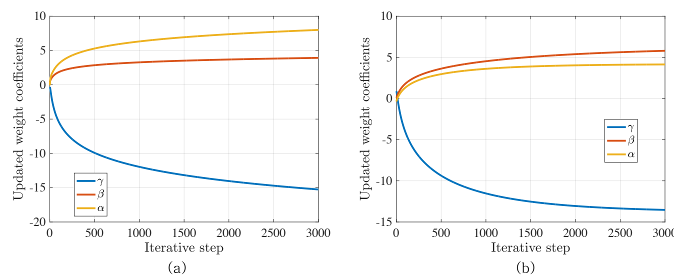

Fig. 7. The updated weight coefficients at different iteration steps with a learning rate of 1, (a) the first scenario, (b) the second scenario.

B. The testing results

In this section, the accuracy of this trained model is tested by predicting the average price of the house based on its distance information. Using on the updated values of α and β obtained in Section-V-A, the trained linear model is shown in Fig. 6. In order to verify the accuracy of this model, based on the data of 2018, the normalized average sold price of the houses is plotted as a function of the normalized distance on the same figure. It can be seen that the trained linear model highly matches the average sold price curve. Only when the normalized distance is greater than 0.7, the model becomes slightly inaccurate.

VI. LOGISTIC REGRESSION RESULTS

In this section, the performance of the studied logistic regression algorithm in Section-IV is examined. Two different scenarios are considered. In the first scenario, the data of the Bayside council and the Frankston council is used in the analysis. In the second scenario, the data is classified between the Bayside council and the Kingston council.

A. The trained model

In the training process of the logistic regression model, the data of 2017 is used. Both the price and the distance information are inputs and their weighting coefficients, γ, β, as well as the bias, α, are updated iteratively based on the desired output, $c \subset \{0, 1\}$, which is the council information. In both

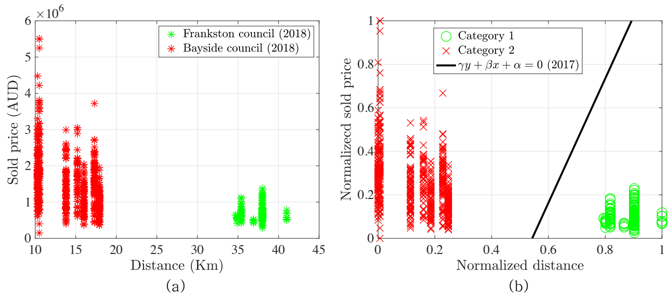

Fig. 8. The data classification results between the Frankston council and the Bayside council, (a) the original data of 2018 with the correct council information, (b) the classification results based on the trained model using the data of 2017.

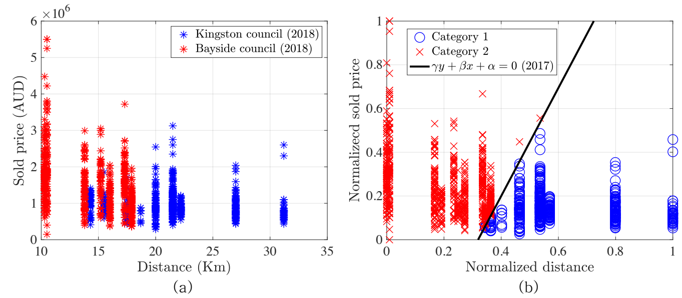

Fig. 9. The data classification results between the Kingston council and the Bayside council, (a) the original data of 2018 with the correct council information, (b) the classification results based on the trained model using the data of 2017.

scenarios, $c=1$ means that the house belongs to the Bayside council. $c=0$ denotes the Frankston council in the first scenario but the Kingston council in the second scenario. Fig. 7 shows the updated weight coefficients at different iteration steps when the learning rate, η, is fixed at 1 for both scenarios. It can be seen that, when the number of iteration step is above 2500, the trained coefficients become steady. In the first scenario, γ, β, α are -15.26, 3.92 and 8, respectively. In the second scenario, γ, β, α become -13.53, 5.8 and 4.13.

B. The testing results

In this section, the performance of the trained logistic model is tested using the data of 2018. Fig. 8 (a) shows the original data used for the classification and its correct council information. Fig. 8 (b) shows the classified results based on the trained linear boundary, $\gamma y+\beta x+\alpha=0$. The data examples marked using red crosses belong to one category. In this category, all data examples satisfy that condition that $\gamma y+\beta x+\alpha>0$ and thus $P>0.5$. The decision is made based on the threshold defined in (9) and therefore $\bar{c}=1$ which means the data belongs to the Bayside council. Similarly, data examples marked using green circles belong to one category which is related to $\gamma y+\beta x+\alpha<0$ and thus $P<0.5$. In this case, $\bar{c}=0$ and the sold houses are estimated to be within the Frankston council. Compared to the correct council information shown in Fig. 8 (a), it can be seen that no classification error is made in the results shown in Fig. 8 (b).

Next, the classification results between the Kingston council and the Bayside council are shown in Fig. 9. Fig. 9 (a) shows the original data to be classified and its correct council information. It can be seen that the distance information of the houses between these two councils is ‘overlapped’ and the price information is also very similar. Although this data classification becomes more challenging compared to the first scenario, it can be seen from Fig. 9 (b) that, most of the data can still be categorized into the correct council and the accuracy rate is above 85%.

VII. Conclusion

In this paper, the housing price data obtained from the Melbourne real estate market is analyzed using linear regression and logistic regression based machine learning algorithms. The key technical points of both algorithms are first explained in detail and their performance is then verified using realistic data. The exact steps in the model training processes are summarized systematically and multiple learning trails are utilized to demonstrate the effectiveness of the learning procedures. It shows that the trained model based on the data of 2017 using the linear regression algorithm can very accurately predict the average housing price in 2018. Also, using the logistic regression method, the sold houses can be classified into different categories based on their belonging councils. The results show that, when the two considered councils are far away apart (e.g. Frankston council and Bayside council), the accuracy of the classification results using the trained model is 100%. When two councils are neighbours (e.g. the Kingston council and the Bayside council), the classification accuracy is still above 85%.

References

[1] P. Abelson and D. Chung, “The real story of housing prices in australia from 1970 to 2003,” *Australian Economic Review*, vol. 38, no. 3, pp. 265–281, 2005.

[2] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[3] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, “Supervised machine learning: A review of classification techniques,” *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[4] P. Langley and H. A. Simon, “Applications of machine learning and rule induction,” *Communications of the ACM*, vol. 38, no. 11, pp. 54–64, 1995.

[5] B. Park and J. K. Bae, “Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data,” *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934, 2015.

[6] N. Nghiep and C. Al, “Predicting housing value: A comparison of multiple regression analysis and artificial neural networks,” *Journal of real estate research*, vol. 22, no. 3, pp. 313–336, 2001.

[7] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2012, vol. 821.

[8] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.

[9] A. Ng, “[online] coursera, machine learning course,” *https://www.coursera.org/learn/machine-learning*.

[10] Kaggle.com, “[online] melbourne housing market,” *https://www.kaggle.com/anthonypino/melbourne-housing-market*.

[11] S. Ruder, “An overview of gradient descent optimization algorithms,” *arXiv preprint arXiv:1609.04747*, 2016.