Int. J. Math. And Appl., 9(3)(2021), 125–128
ISSN: 2347-1557
Available Online: http://ijmaa.in/

International Journal of *Mathematics And its Applications*

Application of Linear Regression to Real Estate

Archit Kumar[1,*]

1 Rocking Horse Ct, Dublin, California, United States.

Abstract: Linear regression is a technique to create a model from given data, where the data is a set of examples and the information from each example consists of a set of features and a target variable. The model can then be used to predict the target variables of new examples from their features. We review this technique and apply it to a problem in business to illustrate how it can help making good business decisions.

Keywords: Machine learning, Supervised learning, Linear regression.

1. Introduction

Machine Learning is the field of computer science that uses data to develop computational models to make predictions and help make decisions [1, 4]. There are several classes of type of problems within machine learning. The problem we consider in this article belongs to the class of problems known as supervised learning.

Square footage	Number of bedrooms	Price
2104	3	399900
1600	3	329900
2400	3	369000
1416	2	232000
3000	4	539900

Table 1. **Example of a data set**

Feature 1	Feature 2	Label
x_{11}	x_{12}	y_1
x_{21}	x_{22}	y_2
$\dots$	$\ddots$	$\dots$
x_{n1}	x_{n2}	y_n

Table 2. **Data set of a supervised learning problem with two numerical features and one label**

Consider Table 1. That table consists of information about several houses. This is part of the data set we consider in this article. While the complete data set consists of information about 47 houses, we show only the data of 5 houses in Table 1.

* *E-mail: roundmonk-research@yahoo.com*

In the machine learning language, each house is called an example. In this case, the information about each house is its square footage, its number of bedrooms, and the price at which it was sold. The goal of applying machine learning techniques to a data set such as the one on Table 1 is to develop a computational model that can later be used to predict the price at which other houses are expected to be sold. We know the square footage and the number of bedrooms of these other houses (not in our data), but we do not know their price, in fact, they have not been sold. In other words, the goal of machine learning is to create a function (that is called model in the machine learning language) that takes as input the square footage and the number of bedrooms of a house and gives as output the price the house is expected to be sold at.

The data problem of predicting the prices of houses as just described, belongs to the class of problems known as supervised learning. In this class of problems, the information about each example consists of features and a label. The features are properties of the examples that are used to predict the label. Thus, in the data set of this article, each example (house) has two features, the square footage and number of bedrooms of the house, and the label is its price.

There are several techniques to analyze supervised learning problems. The particular machine learning technique that we will use is linear regression [3, 2]. This method is explained in Section 2. The analysis of our data set is carried out in Section 3. This article ends with a small discussion in Section 3.

2. Linear Regression

In general, the data of a supervised problem with two numerical features and a numerical label is as displayed in Table 2. We denote with n the number of examples. In the real estate data set we consider in this article $n = 47$. We denote with x_{i1} and x_{i2} the first and second features of the i^{th} example, respectively. In our real estate data set, x_{i1} and x_{i2} are the square footage and the number of bedrooms, respectively. For example, from Table 1, we have that $x_{11} = 2104$ and $x_{12} = 3$, as the first example is a house with 2104 square feet and 3 bedrooms. The label of the i^{th} example is y_i. This means that in our real estate data set, y_i is the price of the i^{th} house. For example, $y_1 = 399900$, as this is the price at which the first house was sold.

We will denote by $\widehat{y} = \widehat{y}(x_1, x_2)$ the prediction from linear regression on an example with features x_1 and x_2. For example, in our real estate problem, having $\widehat{y}(2000, 3) = 350000$ means that our model predicts the price of a house of 2000 square feet and 3 bedrooms to be \$350000. Linear regression assumes that there are some constants w_1, w_2 and b such that $\widehat{y}(x_1, x_2) = w_1x_1 + w_2x_2 + b$. The numbers w_1, w_2 and b are called parameters, and are computed from the given data as is explained later in this section. Before diving into this explanation, and to illustrate this concepts, consider again our real estate data set. If $w_1 = 150$, $w_2 = 20000$ and $b = -10000$, we have that the model predict the price of a 2000 square feet and 3 bedrooms house to be $\widehat{y}(2000, 3) = 150(2000) + 20000(3) - 10000 = 350000$.

2.1. The mean square error

Before we explain how the parameters w_1, w_2 and b are computed, we need to explain the notion of the mean square error. We denote by $\widehat{y}_i$ the prediction of the model on the i^{th} example of the given data set, i.e the data set of Table 2. In other words, $\widehat{y}_i = \widehat{y}(x_{i1}, x_{i2})$. The absolute error the model makes on this prediction is the absolute value of the difference between the prediction of the label and the actual value label of the i^{th} example, in a mathematical formula, the absolute error on the i^{th} example is $|\widehat{y}_i - y_i|$. For example, going back to our data set of Table 1, if $\widehat{y}_1 = 360000$, we have that the absolute error on this example is $|\widehat{y}_1 - y_1| = |360000 - 399900| = 39900$.

Similarly, the square error of the i^{th} example is $(\widehat{y}_i - y_i)^2$. For our data In For example, in our data set of Table 1, if $\widehat{y}_1 = 360000$, we have that the square error on this example is $(\widehat{y}_1 - y_1)^2 = (360000 - 399900)^2 = 1592010000$. The mean

square error on a data set is the average of the square errors on the examples of the data set. We denote this error by J. The equation for J is

$$J = \frac{(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \ldots + (\hat{y}_n - y_n)^2}{n}. \tag{1}$$

Note that the smaller the mean square error, the better the predictions of the model are on the data set. This observation motivates the strategy that the technique of linear regression uses to select the parameters w_1, w_2 and b.

2.2. Minimization of the error to select the parameters

Note that $\widehat{y}_i = w_1 x_{i1} + w_2 x_{i2} + b$. Thus, replacing this expression into the formula for the mean square error leads to

$$J = \frac{(w_1 x_{11} + w_2 x_{12} + b - y_1)^2 + (w_1 x_{21} + w_2 x_{22} + b - y_2)^2 + \ldots + (w_1 x_{n1} + w_2 x_{n2} + b - y_n)^2}{n}. \tag{2}$$

The numbers x_{ij} and y_i are given by the data, we do not have the freedom to change them. However, the parameters w_1, w_2 and b have not been determined yet. Thus, we can regard the mean square error j as a function of the parameters, i.e. $J = J(w_1, w_2, b)$. Different values of the parameters w_1, w_2 and b give different values of the error J. This observation leads to the linear regression technique. Namely, the parameters selected by linear regression are the parameters that make the error J smallest.

2.3. Normal equations

In the formulas below, we use the standard notation $\sum_{i=1}^{n} g(i) = g(1) + g(2) + \ldots + g(n)$ for any function g. For example, $P_{3_{i=1}} i^2 = 1^2 + 2^2 + 3^2 = 14$. Known facts from linear algebra lead to the following result. The parameters w_1, w_2 and b that minimize the mean square error are the parameters that solve the following three equations:

$$\left(\sum_{i=1}^{n} x_{i1}^2\right) w_1 + \left(\sum_{i=1}^{n} x_{i1} x_{i2}\right) w_2 + \left(\sum_{i=1}^{n} x_{i1}\right) b = \sum_{i=1}^{n} x_{i1} y_i \tag{3}$$

$$\left(\sum_{i=1}^{n} x_{i1} x_{i2}\right) w_1 + \left(\sum_{i=1}^{n} x_{i2}^2\right) w_2 + \left(\sum_{i=1}^{n} x_{i2}\right) b = \sum_{i=1}^{n} x_{i2} y_i \tag{4}$$

$$\left(\sum_{i=1}^{n} x_{i1}\right) w_1 + \left(\sum_{i=1}^{n} x_{i2}\right) w_2 + 2b = \sum_{i=1}^{n} y_i \tag{5}$$

Thus, given data, such as the one in Table 2, we first need to construct the equations above and then solve for the parameters w_1, w_2 and b to have our linear regression model $\widehat{y} = \widehat{y}(x_1, x_2) = w_1 x_1 + w_2 x_2 + b$. These calculations are in practice carried out with the help of libraries for scientific computing, such as Numpy in Python. We have carried out these calculations in our data set. The results are discussed in the next section.

3. Linear regression on our data set

We carry out linear regression, as explained in the last section, on our real estate data set. We obtain:

$$w_1 = 139, \quad w_2 = -8050, \quad b = 87800. \tag{6}$$

This means that the model is

$$\widehat{y} = 139 x_1 - 8050 x_2 + 87800, \tag{7}$$

where x_1 is the square footage of the house, x_2 is its number of bedrooms and $\widehat{y}$ is the price. We could also write

$$\text{price} = 139 \text{ square feet} - 8050 \text{ number of bedrooms} + 87800. \tag{8}$$

For example, if a house has 2000 square feet and 3 bedrooms, the model predicts its price to be

$$\text{price} = 139(2000) - 8050(3) + 87800 = 341650. \tag{9}$$

4. Conclusions

In this article we review the concepts of machine learning, supervised learning, mean square error and it minimization, and we explained how these concepts leads to the method of linear regression to solve supervised learning problems when the label is a numerical variable. We illustrated this technique with a simple example from real estate.

References

[1] Andriy Burkov, *The hundred-page machine learning book*, Volume 1, Andriy Burkov Canada, (2019).

[2] George AF Seber and Alan J Lee, *Linear regression analysis*, Volume 329, John Wiley & Sons, (2012).

[3] Xin Yan and Xiaogang Su, *Linear regression analysis: theory and computing*, World Scientific, (2009).

[4] Cha Zhang and Yunqian Ma, *Ensemble machine learning: methods and applications*, Springer, (2012).