Real Estate Price Prediction Taiwan

Juan David Torres A01702686

Abstract – The following document shows the explanation of the implementation of a linear regression algorithm to predict the price of a house in Taiwan depending on the features of the house.

I. INTRODUCTION

Real estate is a word used to describe the land with any building and natural resource in it. However, is used as well to describe the market of buildings and houses and when someone sell them is called a realtor.
The price of a house depends in many aspects, on the size, the location, how near it is from stores, train stations, how many rooms it has, the age and many other aspects.

The prediction of a house price could be very useful when trying to buy or sell a house and with machine learning this can be possible using dataset of the price with some characteristics or attributes of the house that affect in the price.

For the linear regression prediction, I will be using gradient descent by hand and SKLearn of python to compare each of the methods.

II. GRADIENT DESCENT

Gradient descent is one of the algorithms used in machine learning to predict a linear regression.
It uses the following formula to update the parameters of the model.

$$\Theta_j := \Theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\Theta(x_i) - y)x_i]$$

Tetha j will be in this case the parameter that is being constantly updated through the subtraction of itself minus the product of the learning rate divided by the size of the dataset, with the summatory of the difference between the hypothesis of the x and the real y multiplied by the value of Xij.

This process is done manually through a series of calculations in python using NumPy arrays and data frames.

III. SKLearn Python

The framework that will be used is the SKLearn which has its own way to perform linear regression using the functions on the library. The 10% of the dataset is splitted for test while the 90% is used for training.

IV. DATASET

The data set that was used was taken from the University of California at Irvine from the following source (Cheng, 2018). This data set is historical data of real estate valuation in New Taipei City in Taiwan. Which means this model will only work in Taiwan given that the real estate is differs a lot in each country and it is also because the latitude and longitude to give the location must be from the range of the country.

The data set has several columns of features and each one of them have an impact on the result of the price of the house. The first and only attribute that wasn't used for the training of the model is the transaction date because using a date could affect the model and its training due to its structure. The second attribute is the house age, which is given in years, it is given as a float to represent the months. The third one is the distance from the house to a Mass Rapid Transit station given in meters, this is very influential because a lot of the people in Taiwan use this medium of transportation. The fourth attribute is the the number of convenience stores in the living circle on foot. The fifth and sixth are the latitude and longitude that defines the

location of the home, and it is given in degrees.

The independent variable is the house price of unit per area, and it is given in 10000 New Taiwan Dollar/ Ping, Ping is a local unit to define 3.3 squared meters.

The trained then only can be used for real estate properties in Taiwan.

V. RESULTS AND CONCLUSIONS

The gradient descent approach had a much better accuracy because its mean squared error was lower. However, this may also mean that the model tended to overfit when the number of epochs were very high given that for the test results it wasn't performing well while in the training it did.

The framework had a bigger mean squared error, but it is probable that it was because the training was done with fewer epochs which means it didn't overfit.

In conclusion, having more records could be useful when performing the training and the use of regularization could play a great role in the mitigation of overfitting in the model. The number of epochs could affect significantly to the performance of the model as well. But in my opinion the use of gradient descent by hand is much more useful because with it you can learn more and understand how it works and it also can be more customizable in the case you want to change any of the training aspects.

The prediction of a house price can be very useful and with enough data it could be a great project to deploy in the future.

VI. References

Cheng, I. (2018, 18 agosto). *UCI Machine Learning Repository: Real estate valuation data set Data Set* [The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan.]. https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

Chen, J. (2021, 14 enero). *Real Estate Definition*. Investopedia. https://www.investopedia.com/terms/r/realestate.asp

GeeksforGeeks. (2021, 20 agosto). *Gradient Descent in Linear Regression*. https://www.geeksforgeeks.org/gradient-descent-in-linear-regression/