Information Engineering and Technology Faculty
German University in Cairo

Machine Learning Assignment 1 & 2 Redo

Author: Ahmed Osama Saleh
Supervisor: Dr. Maggie Mashaly
Submission Date: 30 July, 2020

Contents

Assignment 1

Linear Regression Overview

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

The first requirement was to use linear regression techniques on a single feature and multiple features on a small dataset in order to estimate the house prices. The second one was to also use the linear regression technique with multiple variables to a much larger dataset with a different perspective in order to enhance the accuracy of the machine learning model.

Handling Dataset

The dataset contained 21 features for describing the house. The date and id columns were dropped from the dataset as they were found to be redundant and don't contribute in predicting the house price. Feature reduction was performed using the heat map correlation matrix such that features that had a correlation coefficient of less than 0.5 with the price feature were dropped. This resulted in having five features which were bedrooms, sqft_living, grade, sqft_above and sqft_living15.

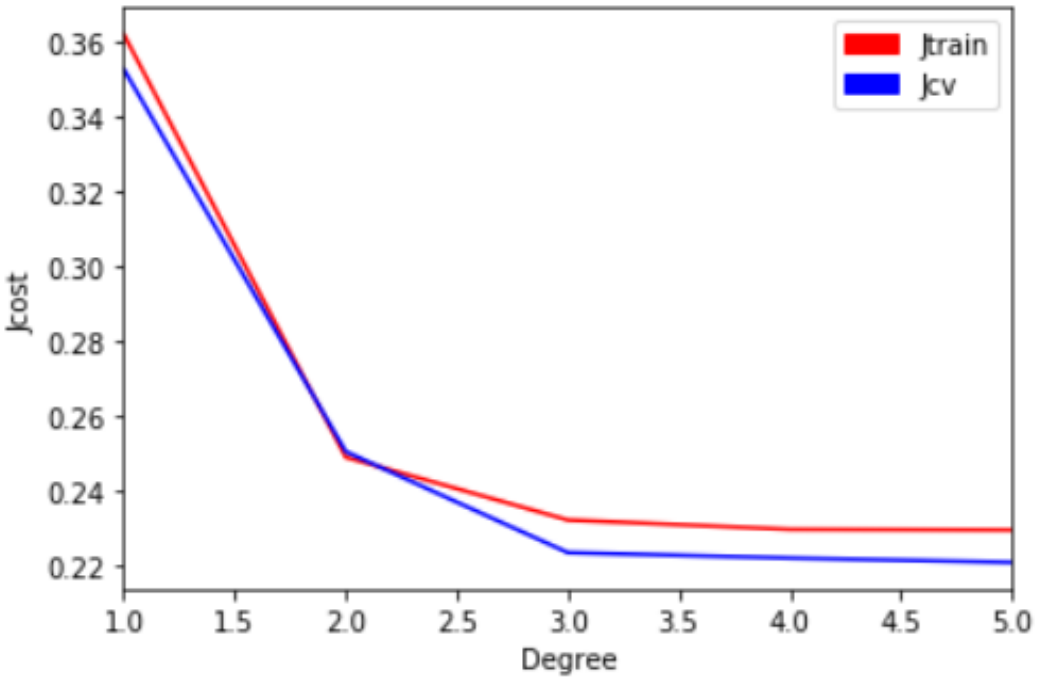
The approach done in Assignment 1 Redo was to use the cross-validation technique so the dataset was divided to 60% as training data, 20% as validating data and 20% as testing data. Random seed is used in the splitting of data into training, validation and testing. Normalization was applied to the dataset as it provides more accuracy for machine learning model to help in the prediction.

Training Data

Training data was used to determine which polynomial degree provides the least cost in order to use the corresponding thetas calculated from the gradient descent in the hypothesis function so all the degrees were tested from 1 till 5. The 5th degree provided the least cost which was equal to 0.2296.

Validating Data

Validating data was used to validate that the 5th degree provided the least cost. It proved such claim as it provided a cost of 0.221 which was the least among all degrees.



Testing Phase

Testing data was used to compare between the prediction of the linear regression model and the actual prices of the houses. The testing data was converted back from the normalization to see compare the efficiency of the machine learning model. The RMSE of the linear regression model is 249595.

Conclusion

The splitting of the dataset into a training, validating and testing made it easier to use five features only to have a polynomial of the 5^{th} degree in order to estimate the price of the houses. This leads to an effective decrease in the complexity of the model rather than using the 18 features which would have been inefficient.

Assignment 2

Logistic Regression Overview

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.

The first requirement was to use logistic regression techniques on a multiple feature dataset in order to estimate the exam score where admitted or not. The

second one was to use the regularized logistic regression technique with multiple variables to the same dataset with a different perspective in order to enhance the accuracy of the machine learning model.

Handling Dataset

The dataset contained 2 features for describing whether being admitted or not.

The approach done in Assignment 2 Redo was to use the cross-validation technique so the dataset was divided to 60% as training data, 20% as validating data and 20% as testing data. Random seed is used in the splitting of data into training, validation and testing.

Training Data

Training data was used to determine which lambda provides the least cost in order to use the corresponding thetas calculated from the gradient descent in the hypothesis function so ten values for lambda were tested. All lambdas provided the same cost which was equal to 0.588.

Validating Data

Validating data was used to validate which lambda provided the least cost. All lambdas provided a cost of 1.112 which was the same among all lambdas.

Testing Data

Testing data was used to calculate the accuracy of the prediction of the regularized logistic regression model to determine whether the exam score was admitted or not admitted. The machine learning model provided an accuracy of 70%.

Conclusion

The splitting of the dataset into a training, validating and testing made it more difficult to use such small dataset which lead to having less accurate prediction of the exam score.