Sangyoon Kim

CPSC393: Machine Learning

Dr. Parlett

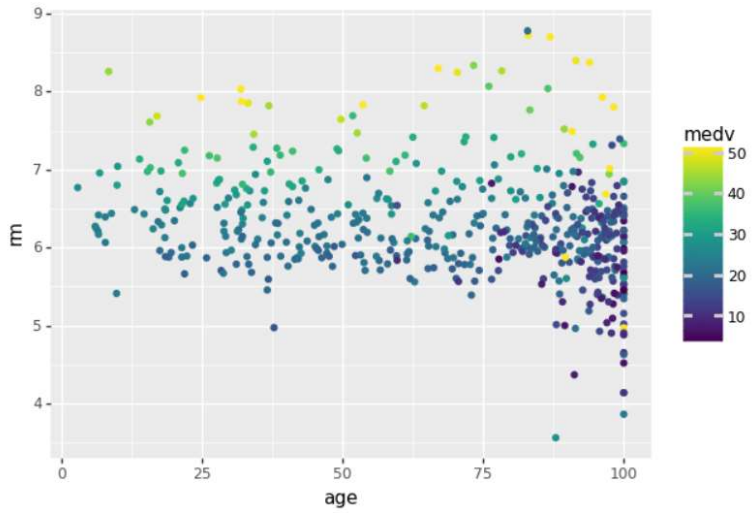
March 17, 2023

Introduction

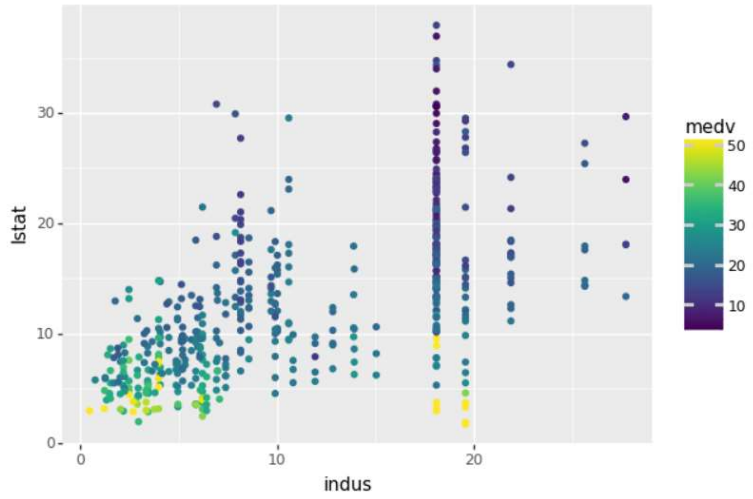
The dataset about Boston Housing Data has been used for this assignment. The dataset includes per capita crime rate by town, proportion of non-retail business acres per town, nitric oxides concentration, average number of rooms per dwelling, proportion of owner-occupied units built prior to 1940, weighted distances to five Boston employment centers, index of accessibility to radial highways, full-value property-tax rate per $10,000, pupil-teacher ratio by town, 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town, and % lower status of the population. Median value of owner-occupied homes in $1000's. The problem being solved in this assignment is predicting the median value of owner-occupied homes in $1000's, given the information of the other features.

Analysis

The dataset has imported as the CSV file, BostonHousing.csv, into a data frame format. Next, all the null data in the data set has dropped by the .dropna() function, which allows to have a complete data without incomplete data. delving into what the file looked like by using the .head() function, which displays the first 5 lines of the CSV file. The dataset has total 14 columns and total 506 rows. Each row represented a separate entry, and the columns were for crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, b, lstat, medv. According to using .info() function, all 506 rows has non-null. The data type is float, except chas, rad, and tax which is int. There were no null values in the dataset, and it was dropped by the .dropna() function.



The graph shows age on x-axis, and rm on y-axis, with medv. The age represents the proportion of owner-occupied units built prior to 1940. The rm average number of rooms per dwelling.



The graph shows indus on x-axis, and lstat on y-x axis, with medv. The indus represents proportion of non-retail business acres per town. The lstat represents % lower status of the population.

Methods

Homework 2 utilized Deep feed forward neural networks to classify the given dataset. The neural network has three hidden layers, and the layers used the ReLU activation. The input layer

has 13 input shape value since the dataset has three columns. For each input layer, dropout regularization by the number of 0 to 0.5, and the batch normalization after the dropout. The output layer has dense as 1. The model has trained by the optimizer the Adam. After that, the model has fitted with the epochs value 250, and validated with the X_test and the y_test. Lastly, the mean squared error and mean squared absolute error for both train and testing has printed out.

Results

By the .summary() function, the total params were 536, the trainable params were 488, and the non-trainable params were 48. The mean squared error of the model for train was 33.46, and the mean squared error of the model for test came out to be 40.02. This means due to the nature of the deep feed neural network, there is a chance in certain cases that the neural network is not accurate. Dropout regularization and the batch normalization has used to build this model. Dropout means to the practice of disregarding certain nodes in a layer at random during training. A dropout is a regularization approach that prevents overfitting by ensuring that no units are codependent with one another. The batch normalization is a method used to make training of artificial neural networks faster and more stable through normalization of the layers' inputs by re-centering and re-scaling. Although there is a lot of debate as to which order the layers should go. For this model, the batch normalization has used followed by dropout. The mean absolute error means to the magnitude of difference between the prediction of an observation and the true value of that observation. The closer mean squared error value is to 0, the more accurate the model is. However, it is an absolute value which is unique to each dataset and can only be used to say whether the model has become more or less accurate than a previous run.

Reflection

I found if the dataset has more full values that without drop the null values would be an excellent way to do data analysis. However, using deep feed neural networks allowed to get better results.