CH 5650: Molecular Data Science
Assignment:1

Shashwat Patel (MM19B053)

27 February 2022

1 Objective

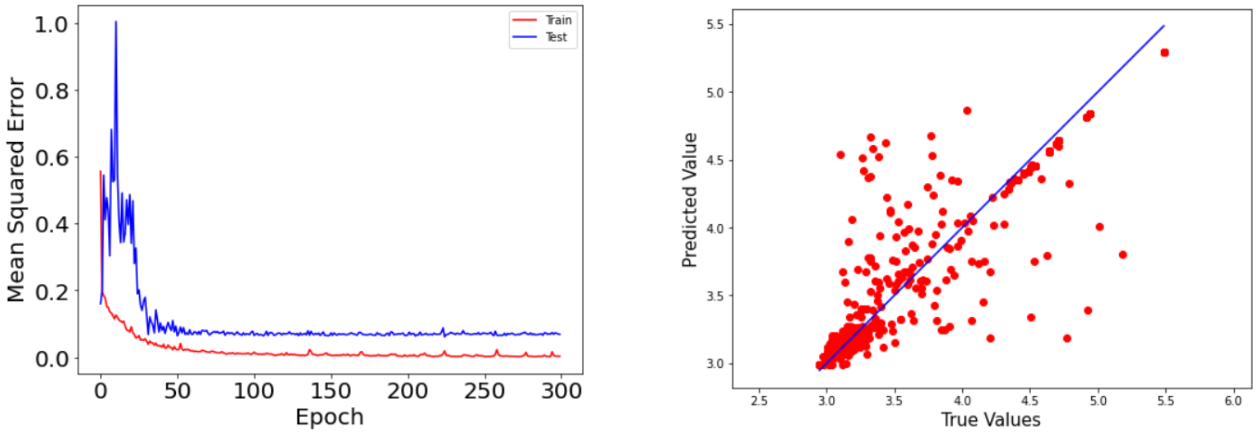
The objective of this assignment is to learn the basic concepts of Keras, a deep learning framework. Using keras, the task is to predict the radius of gyration of given sequence of polymers.

2 Machine Learning Models

3 models are built, 2 of them are neural network models and another is random forest model. 20% of the dataset was used for testing the model.

2.1 Neural Network Model(No tuning)

The 1st neural network was built without any hyperparameter tuning. This neural network consists of 3 hidden layers. The input layer consists of 400 neurons. The 1st hidden layer has 300 neurons, 2nd hidden layer has 150 neurons, 3rd hidden layer has 100 neurons. The model gave an R^2 **value of 0.8219** and **test mean squared error value of 0.0687**, overall showing good results.



(a) Mean Squared Error Vs No. of epochs

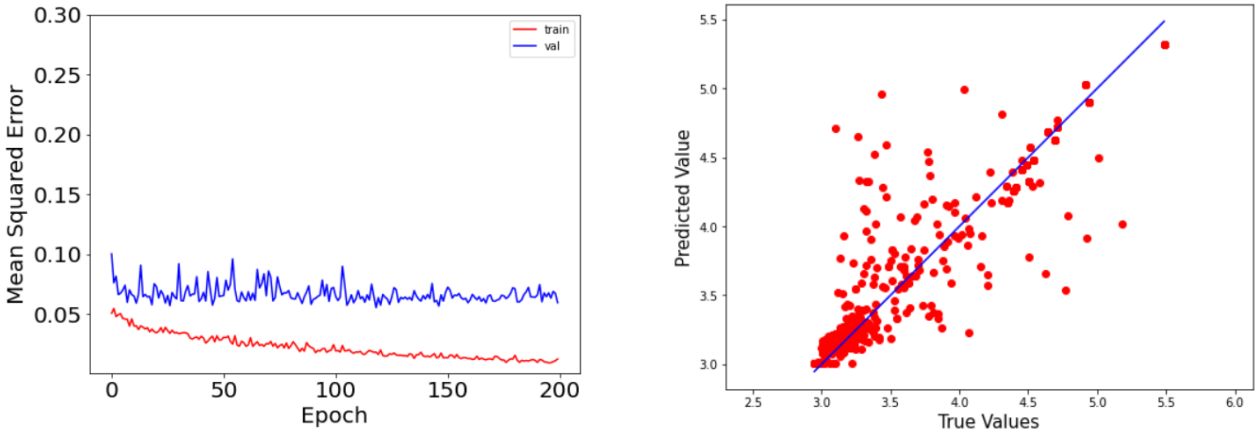
(b) Actual Vs Predicted values

Figure 1: Neural Network without tuning

From figure 1(a), it can be seen that around 100 epochs, there was not much change in error observed in both training and test data. A good fit can be observed in figure 1(b).

2.2 Tuned Neural Network Model

2nd neural network model in which hyperparameter tuning was done consists of 2 hidden layers and a dropout rate of 0.2. The input layer consists of 400 neurons. 1st hidden layer has 280 neurons and 2nd hidden layer has 180 neurons.



(a) Mean Squared Error Vs No. of epochs

(b) Actual Vs Predicted values

Figure 2: Neural Network with tuning

In figure 2(a), the mean squared error for training data decreases with increasing epochs but not much decrease in error is seen in mean squared error for test data, that's because, if compared with figure 1(a) where test error had gone upto value of 1 but in figure 2(a) error always remains in between 0.075 and 0.055, that means using lesser number of epochs would also have given us a good model. More number of points are closer to the regression line in figure 2(b) compared to figure 1(a) showing a better fit than the 1st model. The model gave an R^2 **value of 0.845** and **test mean squared error value of 0.0597**, overall showing somewhat better result than the neural network model which was not tuned.

2.3 Random Forest Model

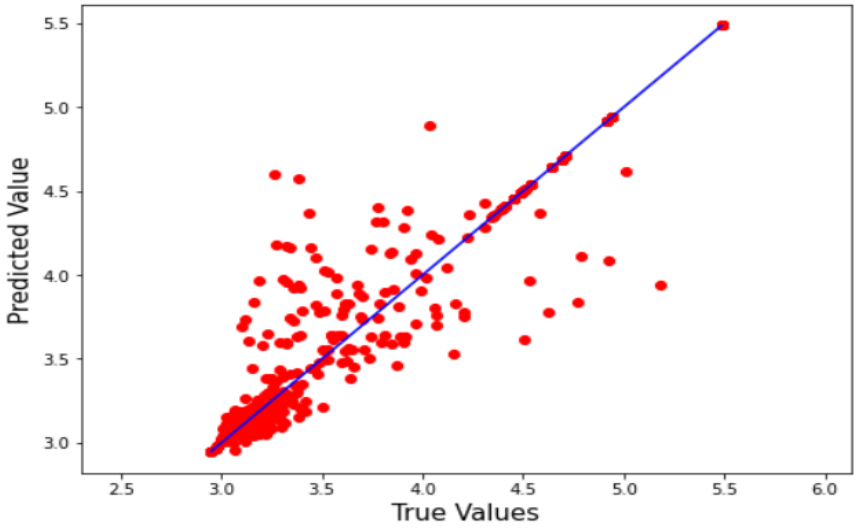


Figure 3: Actual Vs Predicted values

Random forest with 400 trees performed better than neural network model with R^2 **value of 0.8925** and **test mean squared error value of 0.0414**. Surprisingly, for my case random forest model worked the best. It showed a better fit than other model and more points are seen closer to the regression line.