DARSHAN PATEL

FORDHAM UNIVERSITY GABELLI SCHOOL OF BUSINESS

ISGB 799Z – DEEP MACHINE LEARNING – SPRING 2019

Assignment 1

ExeCUTIVE SUMMARIES FOR CHOLLET, CHAPTER 3, AWS

Determining Quality of Movie Based on Review

1. Research Question

In this study, a dataset is shared from the IMDB (Internet Movie Database) that holds information on 50,000 movie reviews. The task here is to determine whether a movie is positively acclaimed or negatively acclaimed by each review writer. This is done simply by looking at the text semantic of each review.

1. Method

Neural networks will be used to categorize reviews. Using a fixed subset of the reviews, a network will build a model that looks at each review, find key words and phrases that relate to the review being good or bad. The complete dataset has 50,000 reviews, of which 15,000 are used to create a model for determining quality of movie, 25,000 to test the model (at production) and 10,000 to report the behavior of the model. The model will run over the training data 20 times, After each time, it is tested on the validation set, adjusts for better performance, and its accuracy is reported. At the end, the testing set can be used for testing the model at the production level.

1. Results and Discussion

Figure

With each run of the dataset, the model gained greater insights into how text semantics of movie reviews. According to Figure 1, it can be seen that with each run of the network, the network learned how to distinguish reviews better and better, decreasing loss on the training data. Unfortunately this leads to an increase in the loss in the validation data. Nonetheless, the network learned how to more or less understand movie reviews based on text semantics. According to Figure 2, the model garnered textual insights in the training data to accurately determine whether a review was good or bad. But when tested on the validation set, the network failed to make huge improvements in the accuracy. In fact, as the network tried to learn more and more, it started to lose its ability to accurately predict for the future. The network has overanalyzed the reviews after 3 run throughs of the data.

Figure 2

1. Conclusion and Implications

In this study, a mechanism to find out if a review is good is bad based on text semantics was explored. It was found that by simply going through reviews a handful number of times is enough to properly categorize reviews. The model predicted some reviews accurately and some poorly. Using the test set for production, it was found that the model performed with an error of 42%. This is not ideal for practical use. Almost half of reviews were incorrectly identified as positive or negative. IMDB would be better off using a number system to properly determine if a movie is good or bad than simply basing it off the text semantics of a movie view.

Determining Topic of Newswires

1. Research Question

In this study, a dataset of short newswires, provided by Reuters, is provided. Each of these newswires come with the text itself, as well as a category of news topic it most corresponds to. The goal of this study is to figure out the topics itself based upon textual information in the newswire without being provided the topic at hand.

1. Method

Using a neural network, a network is trained through 7,982 newswire to determine words that correlate to the news topic provided. It does through several passes of the newswires to learn new information on news topics and then calculate the loss and accuracy on a separate dataset of 1,000 newswires. After training and validating the network, the model is ready for production and testing on completely new newswires.

1. Results and Discussion

It was found that the trained neural network did not benefit from multiple run throughs of the dataset. In Figure 3, it can be seen that loss in the training data decreases exponentially while the validation data loss decreases at first and then hovers around a certain point and then increase. This suggests that although the model tried to minimize extraneous information in the training data, it did not appear to do so when being tested. Furthermore, after a small number of runs through the newswires, the model did not seem to improve or deteriorate. As seen in Figure 4, after ten runs of the neural network, the model does not improve on the accuracy of finding the right news topic for the newswire in the validation set of articles. This shows that as more and more information is attempted to be gathered about the newswire, it does not help to generalize what the news topic is actually. The key idea here is that too much of an investigation into the newswire will not provide additional insight.

Figure 3

Figure 4

1. Conclusion and Implications

Finding a specific news topic for a newswire is a challenge since there are multiple different topics it can belong to. This study tried to do so using a neural network model and a split of data from Reuters. As seen in the results, it was found that 8 runs of the network was enough to find the right topic for newswire. When this model was then tested on a production level dataset, it was found that news topics were predicted with 78% accuracy. This is a satisfactory accuracy score. Being able to predict newswire topics accurately 8 out of 10 times is not bad.

Determining House Prices

1. Research Question

In this study, a dataset of median house prices in Boston from the 1970s is given, as well as some description of the house location, such as crime rate, average number of rooms, distance to employment centers, etc. The goal here is to use characteristics of the suburbs to determine the median house price in Boston.

1. Method

Given a set of 506 house prices, a split was made where 80% of the data was trained on and 20% of the data was left for testing on the production level. A simple neural network configuration was utilized. In addition, since the number of records is small, the training data was split into four pieces where 3 were used for training the mechanism and 1 for testing it. This was done 4 times where the test data was different in each and the error score would be the average of the four runs. In each of these runs, multiple run throughs were made through the data so that the network could enhance its prediction power and the final error score was marked down.

1. Results and Discussion

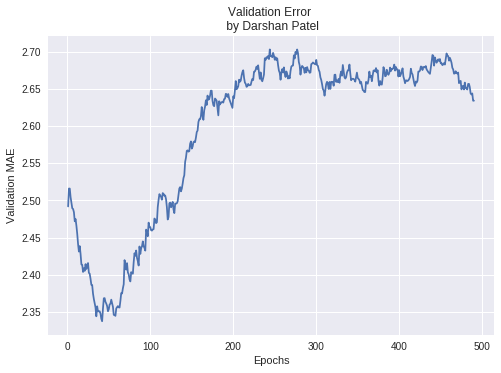
Multiple run throughs of the data is not required for more accurate predictions in house price. According to Figure 5, it can be seen that after 80 pass throughs of the data, the minimum averaged error in house prices in Boston was found. After that, error gradually went up. Going through more than 240 runs of the dataset made higher mispredictions. Using the parameter 80 run throughs, the model was made on the entire training data and then tested using the testing set withhold in the beginning. It was found that the median house prices in Boston was still off by $2,675, on the average. Considering that the range of the median house prices was $10,000 to $50,000, this could mean a prediction could be off by a maximum of 26% at the lower end of prices. This is a huge chunk of error (and money).

Figure 5

1. Conclusion and Implications

It was found that the model will predict the median house prices in Boston with a total average error of $2,675. In the scale of house prices at the time, it can either make a huge difference or a small difference. For better prediction skills, it may be helpful to make a model for lower median house prices and one for higher median house prices. Nonetheless, if these numbers are used for geographical data, it will be sufficient. If these numbers are to be used for pricing houses in Boston in the near future, it can benefit from additional modeling.