UNITED STATES MILITARY ACADEMY

TEE

CS485: SPECIAL TOPICS IN COMPUTER SCIENCE

SECTION G2

MAJ DANIEL RUIZ

By

CADET JACK SUMMERS ’22, CO A4

WEST POINT, NEW YORK

6 FEBRUARY 2022

_J.S__ MY DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

___ _ I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE:____

1. The first step in my path to my final model architecture involved analyzing the problem, and the data. I first investigated the data. I saw 11 attributes, with 10 possible answers (rating scale 1 - 10 inclusive). In turn, I knew that:

a. My first layer required 11 neurons because there are 11 values per sample.

b. I needed to use softmax activation due to the fact the problem is a multiclassification problem. My last layer required 10 neurons due to the 10 options for classification.

c. I needed to normalize the data due to the variety of ranges within each attribute.

d. I needed to vectorize the labels with a one hot encoding scheme.

e. I needed to use categorical crossentropy as the loss function due to the fact the problem requires multi class classification.

After building the basic model with minimal hidden neurons and a modest batch size I trained the model and analyzed the validation accuracy and model. I then added more hidden neurons and increased the number of epochs until the accuracy increased to a desired level. Then I began to implement callbacks and overfitting mitigation techniques. I stopped adjusting hyper parameters after I could not manually beat my best model's max accuracy for several days (cut sling load).

2. I knew I was overfitting when analyzing the graph produced by pyplot. When the training accuracy data points had significantly higher values than the validation data I knew I was overfitting. To mitigate the overfitting, I decreased the number of hidden layers and neurons. Additionally, I utilized I2 regularization and dropout to mitigate the overfitting. Finally, I used two callbacks: early stopping and reduce loss on plateau.

3. With the current data, I do not believe that this is an applied neural network problem. The small amount of data provided only allows for a maximum of 65% accuracy. However, if there was more data, I do believe that a multi class classification model is the most appropriate due to the fact each sample of wine has 10 classification options (1 – 10 inclusive).

4. One factor that may negatively affect the training process is a lack of variability of each classification. The vast majority of the data is labeled between 4 – 8. In turn, the model may not be able to predict lower (1-3) or higher quality wine (9-10).

Bibliography

[1] “Intro to data structures — pandas 1.4.0 documentation,” *pandas.pydata.org*. https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html#:~:text=DataFrame%20is%20a%202%2Ddimensional (accessed Feb. 07, 2022).

[2] “pandas.DataFrame.to_numpy — pandas 1.2.3 documentation,” *pandas.pydata.org*. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_numpy.html.

[3] “python - What does axis in pandas mean?,” *Stack Overflow*. https://stackoverflow.com/questions/22149584/what-does-axis-in-pandas-mean#:~:text=So%20a%20mean%20on%20axis.

[4] Real Python, “Reading and Writing CSV Files in Python,” *Realpython.com*, Jul. 16, 2018. https://realpython.com/python-csv/.