AM11: Machine Learning for Big Data

Assignment IV: Neural Nets & Deep Learning Workshop Report

Due to Friday, February 21, 2020 @9pm

Instructor: Christos Nicolaides

During our workshop we went through two scripts.

A. We build a Neural Network to predict customer churn in telecommunications company.
B. We use LSTM to predict GE stock price using historical data.

In this assignment you are asked to respond to the following questions regarding the two cases of the workshop.

Customer Churn Case:

1. Why did we remove 'customerID' Column during data preprocessing?

2. Data in the real world are rarely clean. On the contrary, they tend to be incomplete, noisy and inconsistent. Hence, it is an important task to handle missing values.

   a. In our dataset, we dropped some rows that include null values instead of handling. Explain why and how did this affect the dataset and the Neural Network training process as a whole?

   b. Explore different techniques of handling missing values in machine learning and give a quick description for each one.
   c. Implement one of these techniques, check the results. Did this have any effect on the overall accuracy of your ANN model? Why or why not? Discuss.

3. Why was a good idea to perform log-transformation on the ‘TotalCharges’ variable?

4. We centered and scaled the data. Explain why?

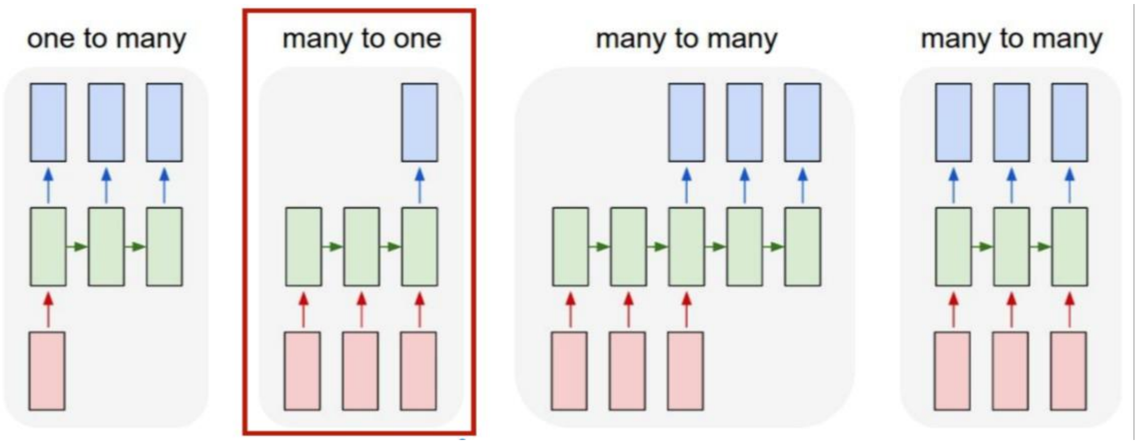
5. In the output layer of the ANN we build using keras, the activation function used is the *sigmoid*. The loss function is *binary-crossentropy*. Explain. What should have we used in case of multiclass classification problem

6. Perform Hyperparameter tuning to better the performance of your model.
   a. Present the architecture and parameters of your best possible model.
   b. Does your model suffer from underfitting or overfitting? Discuss using plots.
   c. Inspect the performance with all the metrics used in class (Accuracy, F1, AUC, etc.).

7. Check your “best” model performance with respect to any classifier you know (logistic, SVM etc.)

GE Stock Price Prediction Case:

8. In our time-series predictions, we used a many-to-one LSTM model.



In our model we used the closing values from 4 previous days to predict one value.

   a. Instead of using one feature *closing value* in each time step, use any combination of features (3 combinations at least) to predict the *closing value* of one upcoming day. Discuss results. Did the performance improve after adding more features?
   b. Repeat (a), but instead of using LSTM, use simple RNN with one feature (closing value) and any combinations of more than one features to predict the closing value of one upcoming day. Discuss results. Did the performance improve after adding more features?
   c. Perform Hyperparameter tuning to better the performance of your LSTM model. (Hyperparameter includes changing the number of previous days, number of layers, number of units in each layer, etc.). Choose only two of them to tune. Discuss your results.
   d. Perform Hyperparameter tuning to better the performance of your RNN model. (Hyperparameter includes changing the number of previous days, number of layers, number of units in each layer, etc.). Choose only two of them to tune. Discuss your results.
   e. Can we predict for more than one day? (e.g. feature values from 8 previous days to predict the *closing values* for 2 upcoming days).

Implement such a model, using RNN or LSTM, perform hyperparameter tuning, plot predictions, and discuss your results.

f. Discuss in half A4 page what are the limitations of Deep Learning on predicting stock market.

g. *Bonus*: Build a CNN model that uses the values of multiple features of *x* previous days to predict the *closing value* for *y* upcoming days. Perform hyperparameter tuning, plot predictions, and discuss your results.

Please return back two codes (one for the customer churn case and one for the GE stock price prediction case) in the form of ".html" + (".R" or "RMD"). Please use comments to guide your work. Use "*surname_name_am11_asmIV_CH.html*", "*surname_name_am11_asmIV_GE.html*" & "*surname_name_am11_asmIV_CH.R*", "*surname_name_am11_asmIV_GE.R*" as the names of the 4 files you will return. Please keep it short. Do not print more than 6 lines of data at any case (just use *head()*)

Please return a 3 to 5 A4-page report where you include your discussion to the several question. Please use font size greater or equal to 11. Use "*surname_name_am11_asmIV.pdf*" as the name of your report.

Good Luck!