CS583A: Course Project

Santander Customer Transaction Prediction

Ravi Patel

May 19, 2019

Summary

I participated in an active (with late submission) competition of predicting if a customer would perform an action based on an anonymized dataset. The model I choose is a K-Fold neural network, which takes in 1 X 200 matrix as an input and outputs a class label. I am implementing the neural network using Keras and run the code on Google Colab. Performance is evaluated on the classification accuracy. By the time I was wrapping up my solution the competition ended and was only accepting late submissions.

Problem Description

The competition is at https://www.kaggle.com/c/santander-customer-transaction-prediction/overview. The is a binary classification problem to classify if a customer will perform an action in the future. The data is given as a .csv file that contains 1X202 vector of numbers. The number of training samples is n = 200,000 and number of classes (target) is 2. The training set is not balance: $n_{target=0}$= 179902 and $n_{target=1}$= 20098. The challenges that can arise: since there are large number of training samples, there might lead to some underfitting; since the data is anonymous is it unclear to understand the importance; since the training set is not balance, how accurate will the model predict.

Solution

The model I choose a neural network that uses StratifiedKFold method. The neural network structure is an input followed by a combination of three Dense, BatchNormalization and Dropout layers with a Relu activation layer and lastly a Dense layer with a Sigmoid activation layer. As a loss function I used was binary cross entropy and the metrics is accuracy. For the optimizer, I used Adamax with a learning rate of 1E-4, epoch of 30/40, batch size of 32/64. Advance tricks used for this model is to have batch normalizations and dropouts and data normalization.

Compared Methods

Several different methods were used to get a higher accuracy score. To get the proper idea how well each method does we need a baseline.

For the baseline, I used the model and trained it using the full training and target dataset then used the model to predict target values of the training data. The result of the baseline: 0.7275.

The next method will split the training data into training and validation data then have the model train with the training portion of the dataset and check against the validation portion of the dataset. The result of that step is: (enter result). Then use the whole dataset to train the model and the result is: 0.7609.

The next method will be to use K-fold (k=4 and shuffle=true, for time constraints). Using the K-Fold method, from SKLearn library, will split the training data to training data and validation data. The number of epochs=40 and batch size=64. Once completed then train the model on the full dataset. This results in: 0.666.

The final method will use Stratified K-fold (k=4 and shuffle=true, for time constraints). Using the Stratified K-Fold method, from SKLearn library, will split the training data to training data and validation data. The number of epochs=30 and batch size=64. Once completed then train the model on the full dataset. This results in: 0.7351

Outcome

After submitting the code, I got a score of 0.84651. The python notebook and code write up can be found on the github link https://github.com/rpatel1291/Santander-Customer-Transaction-Prediction.