Poornima Andukuri

Introduction:

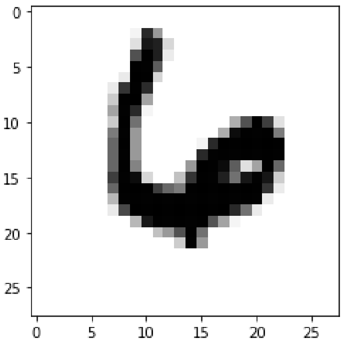
Assignment 6 uses the application of neural networks to assess the classification performance accuracy and processing times to predict the handwritten digits in the MINST dataset. Use Tensorflow library to build a model that along with accurately predicting the digit that is written, will run efficiently on a larger data set.

The steps involved in the following analysis are

1. Use Tensorflow to build a 'Deep Neural Network' (DNN) model
2. Use train data set to fit and evaluate the various designs of the model using hidden layers
3. Use test data set to predict the hand-written digits
4. Record the processing time and model accuracy on train and test datasets
5. Compare the model timing, model (F1, Precision and Recall) scores.

Exploratory Data Analysis:

There were two data sets provided. The first data set is train, that is used to create the model and the second data set is to test the model and predict the digit. The train data has 42000 rows of data and 785 variables. The test data set has 28000 rows and 784 variables. The 784 variables represent an image of 28x28 with each pixel value associated, indicating the darkness of that pixel. The pixel value is an integer between 0 and 255. The image on the right is a binary plot of showing a row of data that has a response value of 6.



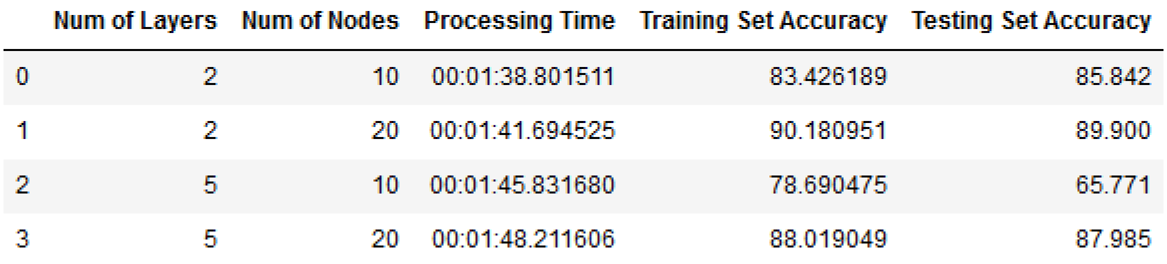
Data Preparation and Overview of Programming:

We use TensorFlow library to fit four different neural networks. The first step to get the data ready to create the model, is to drop the response variable from the train data set and convert both train and test data set into an array. All models are created using the DNNClassifier method in TenorFlow.

The first model is created to use 2 hidden layers with 10 nodes per layer. The second model is created to use 2 hidden layers with 20 nodes per layer. The third model is created to use 5 hidden layers with 10 nodes per layer and the fourth model is created to use 5 hidden layers with 20 nodes per layer.

The number of classes is set to 10, the number of training examples utilized in on iteration (batch_size) is set to 50 and the maximum number of times the program can iterate over the entire dataset (num_epochs) is set to 50 for all models. TensorFlow Estimator from V1 is used to train the model.

The table below lists all the models that were created in the process of the analysis. Performance is based on the accuracy scores of training and test data sets and processing times.



	Num of Layers	Num of Nodes	Processing Time	Training Set Accuracy	Testing Set Accuracy
0	2	10	00:01:38.801511	83.426189	85.842
1	2	20	00:01:41.694525	90.180951	89.900
2	5	10	00:01:45.831680	78.690475	65.771
3	5	20	00:01:48.211606	88.019049	87.985

Insights & Conclusions:

A model with 2 hidden layers with 20 nodes in each layer is recommended. The model takes about 101 seconds to build and evaluate the model. The accuracy on the training data set is 90% and close to 90% on test data set.

Appendix:

The ipynb notebook and an html version of the notebook along with the output and Kaggle submission scores are included in the submission.