Jackson Swalley

Final Project Overview

Applied Deep Learning

This script utilizes the tensorflow, scikeras and sklearn modules in Python to tune the hyperparameters of a neural network model intended for classification of patients as diabetic or not diabetic. The data set from Kaggle includes the following data to predict if a patient is diabetic or not:

1. Number of times pregnant
2. Plasma glucose concentration
3. Diastolic blood pressure
4. Tricep skin thickness
5. 2 hour serum insulin levels
6. BMI
7. Diabetes pedigree function
8. Age
9. Outcome (1 for diabetic, 0 for not)

Many binary classification models currently use a fixed set of hyperparameters for their model. The code attached to this “Read me” trains and tests several models with different parameters using GridSearchCV() and GridSearchCV.fit(). These functions find the model within the set of hyper parameters that produces the highest accuracy. The set of model hyper parameters are as follows:

Number of layers: [2,3]

Nodes in first layer: [64,32,16]

Epochs: [30, 60]

Activation Function: [sigmoid, relu, tanh]

Loss Function: [Binary Cross Entropy, Hinge]

For binary classification, accuracy is defined as the ratio of true positives plus true negatives divided by total number of data points. This metric for each model will be associated with it’s respective accuracy. The script included displays the highest accuracy calculated with its associated model parameters.

A screen shot of a computer code

Description automatically generated

*Figure 1: Highest accuracy and associated model parameters*

*Figure 1* shows the output for the set of model parameters previously listed. It is important to note that this method of varying hyperparameters has the potential to be computationally expensive, as many different models are being trained and tested in the same execution of the code. The data set used has 8 categories and 768 subjects and the run time was roughly 25 minuets while using a single GPU. As complexity increases, computational power required will increase exponentially.

Another important implication of this model is that it may be subject to overfitting. If a model happens to effectively memorize a dataset, it will naturally have the highest accuracy score. Because the objective of this script is to find the highest scoring parameters, there may be a tendency to retrieve parameters that inherently overfit the data.

**Previous work and Citations**

1. Kaggle use of binary classification with fixed parameters on diabetes data set: <https://www.kaggle.com/code/karthik7395/binary-classification-using-neural-networks>
2. Data set: <https://www.kaggle.com/datasets/mathchi/diabetes-data-set>
3. Manual optimization of neural networks: <https://machinelearningmastery.com/manually-optimize-neural-networks/>
4. Overfitting and prevention of overfitting: <https://medium.com/analytics-vidhya/the-perfect-fit-for-a-dnn-596954c9ea39>