Dr. Tony Diana
DATA 602 Introduction to Machine Learning
Homework I Lecture Week 10

Neural Networks

- Use the diabetes database.
- Plot the histograms of the variables. Visualize the difference in distribution between diabetics and non-diabetics with separate plots using Seaborn. Create a subplot of 3 X 3 and plot a density plot for each variable. Hide the 9^{th} subplot.
- Check the data (null data and number of rows with 0 for each variable). Replace the zero values with NaN. Re-check the data.
Replace the NaN values with the mean of the non-missing values.
- Standardize the data. Take a look at the mean, std. deviation, and the maximum of each of the transformed variables.
- Create a train and test set (80/20).
- Use the sequential class from Keras to build the MLP. Add two hidden layers (with the respective node values of 32 and 16, 'relu' activation) and one hidden layer ('sigmoid' activation for output layer). Use 'adam' optimizer and 'binary crossentropy' loss. What does cross-entropy mean and refer to?
- What is the outcome of 100 and 200 epochs? Any difference? Explain.
- Evaluate the training and testing sets' accuracy.
- Provide the confusion matrix using Seaborn. What are your conclusions?
- Provide the ROC graph. What is the area under the curve? What are your conclusions?
- Is the MLP better than a logistic regression model? Do you get a better accuracy with a Random Forest model? Why? Show the outcomes of the different models in a table format.