Homework 5

April 4, 2018

Problem 1. In this assignment, you will be using neural network to classify whether a transaction is fraudulent or not. We provided the required dataset and you will have to design a neural network architecture that accurately identifies the frauds.

The data contains 28 undefined features which define the type of transaction.

1. Find the distribution of each class and explain why the data is skewed.
2. There are many approaches that were developed for handling the skewed data. For this problem, we will use an approach called 'under sampling' where we randomly delete instances from over-represented class. You have to reduce the dataset such that both classes have equal distribution.
3. Split the reduced data into train, valid and test datasets. Now train a neural network classifier (your choice) and calculate precision & recall using confusion matrix.
4. Explain what each metric (precision and recall) represents for this dataset. Which do you think is more important?
5. Use k-fold approach and check whether you can improve the metric chosen in part 4.
6. Tune the classification threshold and find best model using ROC curves. Explain why the model you chose is the best.
7. Use your best network on complete dataset and report the performance.