ENGS/QBS 108 Fall 2017 Assignment 3

Due October 17, 2017

Instructors: George Cybenko and Saeed Hassanpour

Prepared by: Benjamin Priest

Problem: Porto Seguro Dataset [65 points]. In this problem, you will explore and attempt to solve a realistic machine learning problem hosted on the Kaggle competition platform. In order to download the data and access other competition resources, you will need to create a kaggle account. The Porto Seguro dataset includes a number of anonymized features of vehicle drivers collected over the course of a year, as well as a target flag indicating whether the driver filled an insurance claim during the year. Your task is to train models to predict driver safety using this data.

1. [10 points] Data analysis is an important first step. The first two fields in train.csv, "id" and "target," are an index and the label, respectively. The other 57 fields are features with partially anonymized names. The elements of the "target" column are in $\{0, 1\}$, where 1 indicates the driver filed an insurance claim during the year. Proceed to explore this problem's dataset by addressing the following:

   (a) What are the dimensions of the training and testing datasets?

   (b) Features ending in "_cat" are categorical, meaning that their entries are categories rather than numerals. How many categories does each such feature include?

   (c) Features ending in "_bin" are binary, taking entries in $\{0, 1\}$. Are any of these features "one-hot" encodings of a single categorical feature? If so, which ones, and how many categories does it include?

   (d) Explore the discussion page of the competition. What else can you learn about the features or dataset from what others have done? Include code if relevant.

2. [25 points] In this part, you will attempt proscribed a proscribed solution to the competition. You will build a (simple) deep neural net whose desired output is a probability that a driver will file a claim given the input feature vector. You will report results in terms of a ROC curve, where you vary the decision threshold between 0 and 1.

   (a) Hold out 20% of the training dataset as the validation set for this problem.

   (b) Construct a deep neural network with three dense layers of 100, 50, and 20 neurons, respectively, as well as a single sigmoidally activated output layer.

   (c) Train your model on the remaining training data using cross-entropy loss. Report your ROC curve on the training data.

   (d) Apply your model to the validation set and report the corresponding ROC curve.

   (e) Retrain your model on the *whole* training set and apply your model to the testing set. The output format is specified in sample_submission.csv. Submit your output to Kaggle. What performance does it return?

3. [25 points] In this part, you will attempt your own solution to the competition. Repeat problem 2 with a model of your choice. You are free to browse existing kernels for ideas, but your work should ultimately be your own. Credit any resources that you use referenced extensively in your solution.

4. [5 points] How do your solutions in problems 2 and 3 compare? Explain your observations.