ENGS/QBS 108 Fall 2017 Assignment 3

Due October 3, 2017

Instructors: George Cybenko and Saeed Hassanpour

Problem: K-Means Clustering [65 points]. In this problem, you will explore and attempt to solve a realistic machine learning problem. 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 predict whether a driver will fill an insurance claim given their feature vector.

1. [10 points] Data analysis is an important first step. Proceed to explore this problem's dataset by addressing the following:
   (a) What are the dimensions of the training and testing datasets?
   (b) Plot a histogram for
2. [20 points] Apply SVMs to the problem:
   (a) Separate the training data into a k-fold cross validation testing framework for k=10.
   (b) Train SVM models using your validation framework. Try a few different kernels. What kernel yields the best performance?
   (c) Apply your best model to the testing set. Report your classification accuracy.
3. [25 points] Apply a (simple) deep neural net to the problem:
   (a) Construct a deep neural network (layers? convolution?)
   (b) Train your model. Report your performance on the training data.
   (c) Apply your model to the testing set. Report your classification accuracy.