STA 545 Statistical Data Mining I, Fall 2020

Homework 10, due: Wednesday 12/9/2020 (1PM)

Please submit the pdf file generated by R markdown in UBlearns. Please use tables, figures, or a few sentences to answer data analysis questions.

1. (40 points) Please download the spam data set in the ElemStatLearn R package. This dataset is for learning to classify e-mail as spam or real mail. There are 58 columns: 57 of them are features (see help(spam)), and the last one is a categorical variable ("factor"), called spam, with two values, email and spam. There are 4601 rows, representing 4601 different e-mails.

   (a) (5 points) Divide the data set randomly into a training set of 2301 rows and a testing set of 2300 rows. What fraction of each half is spam?

   (b) (10 points) Use bagging to fit an ensemble of 100 trees to the training data. Report the error rate of the ensemble on the testing data. Include a plot of the importance of the variables, according to the ensemble.

   (c) (10 points) Fit a series of random-forest classifiers to the training data, to explore the sensitivity to the parameter m. Plot both the OOB error as well as the test error against a suitably chosen range of values for m.

   (d) (10 points) Use the AdaBoost method with 100 boosting iterations. Report the error rate of the classifier on the testing data.

   (e) (5 points) Fit logistic regression to this dataset. Evaluate the model on the test set and compare to the Bagging, Random Forest and AdaBoost results.

2. (30 points) Fit a single hidden layer neural network to the spam data that is shown in the package "ElemStatLearn". Use the five-fold cross-validation method to determine the number of neurons to use in the layer.

3. (30 points) In this problem, we consider a simulation study.

   (a) (10 points) Generate a training dataset (100 observations) and a test dataset (1000 observations) from the model

   $$Y = \sigma(a_1^T X) + (a_2^T X)^2 + 0.3 \cdot Z,$$

   where $\sigma(v) = 1/(1 + \exp(-v))$ is the sigmoid function, Z is standard normal, $X^T = (X_1, X_2)$, each X_j being independent standard normal, and $a_1 = (3, 3)^T$, $a_2 = (3, -3)^T$

   (b) (10 points) Apply the projection pursuit regression model to this simulated data. Plot the training and test error curves as a function of the number of ridge functions.

   (c) (10 points) Apply the single hidden layer neural network to this simulated data. Plot the training and test error curves as a function of the number of neurons in the layer.