Statistical Data Mining I
Final Homework

Due: Sunday December 15th (11:59 pm)
30 points

Directions: Submit all source codes with write up.

1) (10 points; Exercise 11.7) Fit a neural network to the spam data of Section 9.1.2. The data is available through the package "ElemStatLearn". Use cross-validation or the hold out method to determine the number of neurons to use in the layer. Compare your results to those for the additive model given in the chapter. When making the comparison, consider both the classification performance and interpretability of the final model.

2) (10 points) Take any classification data set and divide it up into a learning set and a test set. Change the value of one observation on one input variable in the learning set so that the value is now a univariate outlier. Fit separate single-hidden-layer neural networks to the original learning-set data and to the learning-set data with the outlier. Use cross-validation or the hold out method to determine the number of neurons to use in the layer. Comment on the effect of the outlier on the fit and on its effect on classifying the test set. Shrink the value of that outlier toward its original value and evaluate when the effect of the outlier on the fit vanishes. How far away must the outlier move from its original value that significant changes to the network coefficient estimates occur?

3) (10 points; ISLR modified Ch9ex8) This problem involves the OJ data set in the ISLR package. We are interested in the prediction of "Purchase". Divide the data into test and training.

   (A) Fit a support vector classifier with varying cost parameters over the range [0.01, 10]. Plot the training and test error across this spectrum of cost parameters, and determine the optimal cost.

   (B) Repeat the exercise in (A) for a support vector machine with a radial kernel. (Use the default parameter for gamma). Repeat the exercise again for a support vector machine with a polynomial kernel of degree=2. Reflect on the performance of the SVM with different kernels, and the support vector classifier, i.e., SVM with a linear kernel.