

Machine Learning 201

Homework 4

Mike Bowles, PhD & Patricia Hoffman, PhD

1a) Use your favorite linear regularized regression technique on the red wine data set. What is the mean squared error?

1b) Next, use a second ordered polynomial basis expansion for this data set. What is the mean squared error? Was there an improvement?

1c) Finally, use spline functions with various degrees of freedom on this data set. Which set of spline functions produces the best mean squared error?

1d) Compare the mean squared error resulting from the different techniques used on this data set.

2) The Glass Data Set, which is used to classify different types of glass, can be found at the following web site:

<http://archive.ics.uci.edu/ml/machine-learning-databases/glass/> Use glmnet on this data set and measure the misclassification error. Next, use polynomial basis functions, and then use spline basis functions. Compare the resulting misclassification errors from these techniques.

3a) Inspect the basic relationships between factors in Joe's data set by creating a pair wise cross plot. Examine the covariance matrix for the various attributes in Joe's data set.

3b) Perform regularized regression on Joe's data set excluding the date and time variables.

3c) Propose a method to encode the date and time variable. Use regularized regression including these variables and compare the results with 3b.