Stat 365/665 Homework #2: Variable Selection and Shrinkage Methods in Regression
Due on Monday, Feb 14, 2011, at the beginning of class (on paper).

Format of the report: Explain clearly your methods and your results. Point out what you find interesting and try to make sense of it. Use tables and plots if they better represent your results and explain them. Please put your code and less important output in the appendix.

1. The lasso estimate is defined by

   $\hat{\beta}^{lasso} = \arg\min_\beta \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$

   or equivalently

   $\hat{\beta}^{lasso} = \arg\min_\beta \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$ subject to $\sum_{j=1}^{p} |\beta_j| \leq t$

   Prove the following result: In case of orthogonal inputs, the lasso estimates translate least squares estimates by a constant factor, truncating at zero, that is, $\hat{\beta}^{lasso} = sign\left(\hat{\beta}_j^{LS}\right)\left(\left|\hat{\beta}_j^{LS}\right| - \gamma\right)_+$,$\gamma$ is a one-to-one transformation of the bound t appearing in the equivalent definition of lasso problem.

2. Explore "Communities and Crime" data, which is available from UCI respository. The data and its description is available at

   http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.

   There are 1994 instances and 128 variables. The task is to predict crime rate from socio-economic variables and law enforcement variables using linear models for regression. I did some pre-processing and preliminary analysis of the data and noticed a couple of issues with the data.

   (a) Missing values: we can impute the missing values if there are a relatively small proportion of them. Commonly used methods include imputing the missing values with column means and means of K-nearest neighbors. However I would not mind removing the variable if there are a large proportion of missing values for that variable.

   (b) Categorical input variables: A variable with K categories is coded as K-1 dummy variables. If you are not very familiar with this treatment, please think carefully about how to interpret the coefficients of the dummy variables. If K is very large, including the categorical variable is effectively increase the number of input variables greatly. So you need to consider whether it is a reasonable thing to do.

   To make everyone's results comparable, I will randomly pick 400 instances as our test data. The indices of these instances are available on classesv2. So you will use the rest 1594 instances as your training set to fit a model and choose tuning parameters. Then you use the fitted model to predict on the test data. Variable selection and shrinkage methods should be considered, including stepwise regression, ridge regression and lasso. For ridge and lasso, use 10 fold cross-validation to choose tuning parameters. Interpret your results and comment on any interesting findings.