Problem Set 1: Ridge, Lasso, PCR and PCS

BUAD 5082 – Spring 2019

1. Objectives

The purpose of this problem set is to provide you with an opportunity to practice the kinds of skills that I expect you to be able to perform on an exam.

2. What You Will Need

- Access to a Windows computer with R

3. Solutions

Solutions to these problems will be posted several days after this Problem Set is posted..

4. Preliminaries:

Problem 1:

Predict the number of applications received using the other variables in the College data set from the ISLR package.

a) Set the seed to 5082 and split the data set into a training set and a test set (80/20 split).
b) Fit a linear model using least squares on the training set, and report the test error obtained.
c) Recreate the glm model with the full sample and save the coefficients in a (named) vector
d) Create a λ grid vector of 100 elements ranging from 10^{10} to 10^{-4}
e) Use cv.glmnet to fit a ridge regression model on the training set using this grid of *λ's*
f) Plot the MSE as a function of the log of lambda.
g) Display the best cross validated λ.
h) Report the test error obtained using the best λ.
i) Create a ridge model using the full sample and save the coefficients of the model with the best lambda.in a named vector
j) Display a dataframe with the glm and ridge coefficients. Name the two columns glm and ridge.
k) Repeat e) through i) with the lasso instead of ridge, then display a dataframe with all three sets of coefficients
l) In view of the coefficient report in the previous task, what can you conclude about the bias-variance tradeoff in this situation? Use one or more comment lines to respond to this question.
m) Set the seed to 5082 and fit a principal component regression model on the training set
n) Plot the cross-validated training MSEP as a function of the number of components.
o) Choose a value for *M* that will produce a significantly simpler model than the full model, even though it appears that the minimum cross-validated training error rate occurs at *M* = *p* (I chose 8). Compute and display the test MSE for this choice. Was this a good idea?

p) Repeat items m) to o) above with partial least squares regression instead of principal component regression, again choosing *M* with the purpose in mind of sacrificing training MSEP for simplicity (I chose 8).

q) Display the five test MSE's as a dataframe and comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

Problem 2:

Try to predict per capita crime rate in the Boston data set from the MASS package.

a) Construct lasso, ridge regression, principal component regression and partial least squares regression models. Present and discuss results for these approaches.

b) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross validation, or some other reasonable alternative, as opposed to using training error.

c) Does your chosen model involve all of the features in the original data set? Why or why not?