Homework 3: Linear Model Selection and Regularization

Overview

Due Sunday by 11:59 pm.

Fork the `problem-set-3` repository

Conceptual exercises

Training/test error for subset selection

1. (5 points) Generate a data set with $p = 20$ features, $n = 1000$ observations, and an associated quantitative response vector generated according to the model

$$Y = X\beta + \epsilon$$

   where β has some elements that are exactly equal to zero.

2. (10 points) Split your data set into a training set containing 100 observations and a test set containing 900 observations.

3. (10 points) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size. For which model size does the training set MSE take on its minimum value?

4. (5 points) Plot the test set MSE associated with the best model of each size.

5. (5 points) For which model size does the test set MSE take on its minimum value? Comment on your results.

   > If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you generate the data previously until you create a data generating process in which the test set MSE is minimized for an intermediate model size.

6. (10 points) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient sizes.

7. (10 points) Create a plot displaying

$$\sqrt{\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^r)^2}$$

   for a range of values of r, where $\hat{\beta}_j^r$ is the jth coefficient estimate for the best model containing r coefficients. Comment on what you observe. How does this compare to the test MSE plot?

Application exercises

The General Social Survey is a biannual survey of the American public.[1]

> The GSS gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. Hundreds of trends have been tracked since 1972. In addition, since the GSS adopted questions from earlier surveys, trends can be followed

[1] Conducted by NORC at the University of Chicago.

for up to 70 years. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.

In this problem set, you are going to predict individual feelings towards egalitarianism. Specifically, `egalit_scale` is an additive index constructed from a series of questions designed to measure how egalitarian individuals are – that is, the extent to which they think economic opportunities should be distributed more equally in society. The feature ranges from 1 (low egalitarianism) to 35 (high egalitarianism).

`gss_*.csv` contain a selection of variables from the 2012 GSS. Documentation for the other features (if not clearly coded) can be viewed here. Some data pre-processing has been done in advance for you to ease your model fitting:

- Missing values have been imputed
- Nominal variables with more than two classes have been converted to dummy variables
- Remaining categorical variables have been converted to integer values

Your task is to construct a series of statistical/machine learning models to accurately predict an individual's egalitarianism using model selection and regularization methods. *Use all the available predictors for each model unless otherwise specified.*

1. (10 points) Fit a **least squares linear** model on the training set, and report the test MSE.
2. (10 points) Fit a **ridge** regression model on the training set, with λ chosen by 10-fold cross-validation. Report the test MSE.
3. (10 points) Fit a **lasso** regression on the training set, with λ chosen by 10-fold cross-validation. Report the test MSE, along with the number of non-zero coefficient estimates.
4. (10 points) Fit an **elastic net** regression model on the training set, with α and λ chosen by 10-fold cross-validation. That is, estimate models with $\alpha = 0, 0.1, 0.2, \ldots, 1$ using the same values for λ across each model. Select the combination of α and λ with the lowest cross-validation MSE. *For that combination*, report the test MSE along with the number of non-zero coefficient estimates.
5. (5 points) Comment on the results obtained. How accurately can we predict an individual's egalitarianism? Is there much difference among the test errors resulting from these approaches?