POL 350C, Homework 7

May 19, 2016

Assigned: 5/19/2016
Due: 5/26/2016

In this problem set we're going to use machine learning to predict student's drinking habits. The data come from a public health study of Portugese students. You can read more about the variables (and their interpretation) here:

http://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION#
.
We're going to model the sum of weekday and weekend drinking activity.

The data are stored in `StudentDrinking.RData` on canvas. The dependent variable is, `alcohol`, which is a measure of alcohol consumption. The bigger `alcohol` is, the more students drink. The covariates are stored in `X`.

1 Comparing Coefficients from OLS, LASSO, Ridge, and Elastic Net

We first want to explore the behavior of OLS, LASSO, and Ridge applied to the data.

i) Fit a linear regression of alcohol on the covariates in the included data

ii) Using `cv.glmnet` fit a LASSO regression of alcohol on the covariates

iii) Using `cv.glmnet` fit a Ridge regression of alcohol on the covariates

iv) Using `cv.glmnet` fit an elastic-net regression of alcohol on the covariates, with $\alpha = 0.5$. Explain what $\alpha = 0.5$ implies about the model you're fitting.

v) Using your models from (i-iv) let's examine the behavior of the coefficient on `male` as λ increases

    a) Suppose `glmnet.obj` contains the results from applying `cv.glmnet`. To obtain the coefficient values for the sequence of λ values tested in `cv.glmnet`, we use the coefficient function `coef(glmnet.obj, s = glmnet.obj$lambda)`.

Use this function to obtain a matrix of coefficients for the models used in (ii-iv).

b) Using the matrix for each method, plot the coefficient on `male` against the value of λ from the models in ii-iv. Include the coefficient from OLS as a flat line. What do you notice as λ increases?

2 Cross-Validation, Super Learning and Ensembles

We're going to assess the performance of five models, an unweighted average, and a super-learning average of the methods.

i) First, set the first 20 rows to the side for use as the validation set.

ii) We'll first estimate the (unconstrained) super learner weights.

   a) On the training data (all but the first 20 rows) perform ten fold cross validation, including (1) linear regression, (2) LASSO, (3) Ridge, (4) Elastic-Net, and (5) Random Forest. Obtain 5 predictions for each observation in the training set, one from each observation

   b) Regress the dependent variable on the out of sample prediction, (without including an intercept).

   c) Store those weights as $\boldsymbol{w}$

iii) Now, fit all 5 models from (ii)-(a) to the entire training data set and predict the drinking level from the vallidation set (the data put off to the side).

iv) Obtain two ensemble predictions.

   a) Take the unweighted average of the predictions from the methods

   b) Take the weighted average, using the weights $\boldsymbol{w}$.

v) You should have 7 predictions. Store those in a matrix and report the correlation between the predictions

vi) Using the average absolute difference as a loss function assess the performance of each method. Which method performs best? Which performs the worst?

   The average absolute difference for method k is defined as

$$L(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}_k) = \sum_{i=1}^{N_{\text{validation}}} \frac{|Y_i - \widehat{Y}_{ik}|}{N_{\text{validation}}}$$

   where $N_{\text{validation}}$ refers to the number of observations in the validation set $\widehat{\boldsymbol{Y}}_k$ refers to the predictions from the k^{th} method,