STA4241 Homework 4, Fall 2021

Please turn in your own work, though you may discuss these problems with your classmates, professor, and TA. The assignment is due on Wednesday, November 10th at 11:59pm.

(1) Read in Problem1.csv off of the course website. We are interested in using the lasso to estimate the parameters of the model, $E(Y|X) = X\beta$. For this problem I want you to program the lasso solution manually. I want you to utilize the following algorithm for estimating lasso parameters:

Algorithm 1: Coordinate descent algorithm for the lasso

1. Initialize $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p)$.
2. Compute $r = \mathbf{Y} - \mathbf{X}\tilde{\beta}$
3. Repeat until convergence:
   (a) For $j = 1, \ldots, p$,
      i. Compute $r_j = r + \mathbf{X}_j\tilde{\beta}_j$
      ii. Compute $\beta^+ = \text{soft}\left(\frac{r_j^T\mathbf{X}_j}{\mathbf{X}_j^T\mathbf{X}_j}, \lambda\right)$
      iii. Set $[\tilde{\beta}]_j = \beta^+$
      iv. Compute $r = r_j - \mathbf{X}_j\tilde{\beta}_j$

The soft thresholding operator is defined as $\text{soft}(x, \lambda) = \max(|x| - \lambda, 0)\text{sign}(x)$.

(i) Find the lasso solution for $\lambda = 0.5$ by manually programming lasso. You may assume that there is no intercept in the model and therefore you only have parameters $\beta_1, \ldots, \beta_p$.

(ii) Find the lasso solution for $\lambda = 0.5$ using glmnet. Set intercept=FALSE to keep it from finding an intercept. Compare your solution to the one from glmnet.

(iii) Compute the lasso solution using your own code for a sequence of λ values between 0 and 2 and create a plot of the coefficients as a function of λ. This plot should look similar to the plots on slide 6 of week 8 lecture notes.

(2) Read in Problem2train.dat and Problem2testing.dat off of the course website. All models should be trained and fit exclusively on the training data set. The testing data set should only be used to evaluate the predictive performance of the final models found from the training data.

(i) Make a heatmap of the empirical correlation matrix of the covariates in this data set. This is a matrix where the (i, j) element is $\text{corr}(X_i, X_j)$. Given this plot, do you think that using PCA on this data set will be helpful?

(ii) Perform PCA on the training data and make a plot that shows the magnitude of the variance explained by each principle component. The plot should look similar to the one on slide 50 of week 8 lecture notes. Does this tell you anything about whether PCA will be useful for prediction in this data set?

(iii) Find the predictive performance of all approaches we have learned in this class to deal with high-dimensional data. These include:

1. Lasso regression
2. Ridge regression
3. Principle components regression with the number of components chosen by cross validation
4. Principle components regression with the number of components being set to the minimum number of components that maintains 95% of the variability in the original covariates. The proportion of variability explained by a set of PCs is the sum of their respective eigenvalues divided by the sum of all the PCs eigenvalues.
5. Partial least squares with the number of components chosen by cross validation
6. Extra credit (You will get 1 bonus point if you include this approach)! Fit the elastic net model with both tuning parameters chosen via cross validation.

(iv) Which approaches do best? Why do you think these approaches did best for this given data set?