Homework 6 – Regularization, PCA, Partial Least Squares

Due Nov. 12, 2017, 5pm

Learning objectives: we are finishing up model tuning and selection in this assignment. In a previous assignment, you worked on subset selection techniques and cross validation to build an improved model. This week, you are going to add another skill to your "skill belt" by comparing the performance of last week's work with this weeks.

Recall the setup for questions 5 from homework 5:

"Now, we consider a validation set. In this case, please use the technique used in the text to create the test and training sets in the text. Set your seed to 1 and then create a vector indicating the rows that should be in the test and training set. If you are using Python, do whatever is analogous to this so that grading is easier and more consistent…. Use 3-fold cross validation."

Instead of using a subset selection technique we will now use a regularization technique. We would like to compare the performance of all of the techniques – whatever best model you got from homework 5, ridge and lasso. The general idea here is to reinforce that all of these techniques (subset selection and regularization) help tune a model and is orthogonal to using cross validation and what metric you choose.

1. Use ridge regression to find a solution to predicting mpg as a function of all features except for name. Remove "name" from the data set and treat origin as a categorical variable.
    a. Plot the MSE as lambda ranges from 0 to 10^10. If there is no "minimum" (you see a value of MSE that has values to the left and right that are higher) within this range, please try more values of lambda until you generate a minimum.
    b. What is the lambda with the minimum MSE (optimal) and what is the MSE?
    c. What is the MSE when lamba is 0?
    d. What is the MSE when lambda is 100?
    e. What are the coefficients when lamba is 0, 100, and optimal?
2. Use lasso regression to find a solution to predicting mpg as a function of all features except for name. Remove "name" from the data set and treat origin as a categorical variable.
    a. Plot the MSE as lambda ranges from 0 to 10^10. Include at least 10 values of lambda. If there is no "minimum" within this range, please try more values of lambda until you generate a minimum.
    b. What is the lambda with the minimum MSE (optimal) and what is the MSE?
    c. What is the MSE when lamba is 0?
    d. What is the MSE when lambda is 100?
    e. What are the coefficients when lamba is 0, 100, and optimal?
    f. Which generated a lower MSE, ridge or lasso?
    g. Refer to homework 5. Do the MSEs in either ridge or lasso improve over those in homework 5?

3. We will now repeat the above exercise using principle component regression. Use the pcr method described in the book, however, use the split for the test sets above to conduct cross validation.
    a. Plot the MSE as a function of number of principle components.
    b. Plot the variance explained as a function of number of principle components.
    c. What is the number of principle components in the best (lowest MSE) model?
    d. What is its MSE?
    e. Is there another number of principle components you might consider? Why?
4. You knew I would ask this. Repeat the above using PLS.
    a. Plot the MSE as a function of number of random segments.
    b. Plot the variance explained as a function of number of random segments.
    c. What is the number of random segments in the best (lowest MSE) model?
    d. What is its MSE?
    e. Is there another number of random segments you might consider? Why?

PCA – The text shows how to do PCR – principle components regression. This is essentially creating a predictive model using PCA. But let's do what Andrew Ng demonstrates in his online class and try to do the same by computing our own principle components. Recall from class, you should standardize variables, generate a covariance matrix, run SVD, get the eigen vectors, then transform the original data…

5. One use of dimensionality reduction is to try to "understand" the data. People find the "top" principle components of a data set and then cluster these components and try to "manually" determine groupings. In this exercise, please compute the first two principle components of the data set and then plot them. Do you see any clusters?
    a. Use the algorithm for PCA described in class (also see Andrew Ng's videos on PCA). (Note: please "scale" the data.) Plot the amount of variance explained as a function of number of principle components.
    b. How much variance is explained by the first two principle components?
    c. Plot the first two principle components on a 2D graph.
    d. Do you see any clusters? Eyeball possible clusters in the plot and compare the points in them – you can circle the points you think belong to the same cluster or plot the points using different colors / markets - to explain what the clusters mean (e.g., cluster 1 is high mpg cars with high horsepower). If there is no "obvious" clustering, just pick a point where you can split the data in two and describe what's in either split.
        i. Comment: Using PCA to reduce dimensions in order to plot the points is one application of PCA, besides regression. A related technique has been used to group movies together in the Netflix data set.
        ii. Comment 2: Huh? What do you mean describe what's in a cluster? The ability to tell a story from the data is a typical data science exercise. Whether it's correct or interesting is a different story... Just give it a try.

Closing comments: I hope that you see from these exercises that there is a wealth of ways to perform the bias-variance tradeoff, where it's generally not clear which one is best. You just have to learn by doing and find your style. What is common among all these techniques is they take as inputs the same data and use the same overall framework for evaluation. Finally, please go over the other questions at the end of the chapter and ask question if you have any!