API 222: Problem Set 3

Machine Learning and Big Data Analytics

Due October 22, before midnight EST

Conceptual Questions

1. For which of the following methods should you standardize your variables? PLS, PCR, Lasso, Ridge, Linear Regression, Logistic Regression, KNN? **(0.25 pts each)**

2. I want to minimize the number of false positives in my data. I run two models: Model A and Model B. AUC for Model A run on the is 0.7 and for Model B is 0.8. Therefore, can we conclude that we should choose Model B? Why or why not?**(0.5 pts)**

3. If we run a Lasso with a λ of 0, what is this model equivalent to? As λ approaches ∞, what will happen to the regression coefficients? **(0.5 pt)**

4. When will Ridge and Lasso outperform PCR? **(0.5 pts)**

5. We want to perform PLS regression. Let's denote two of our variables as X1 and X2. The correlation between X1 and the outcome variable is 0 and the correlation between X2 and Y is .9. Should you conclude X2 will receive more weight in the PLS regression. Why or why not? **(0.75 pts)**

6. We want to perform PCR regression. Let's denote two of our variables as X1 and X2. The correlation between X1 and the outcome variable is 0 and the correlation between X2 and Y is .9. Should you conclude X2 will receive more weight in the the first factor loading of PCR regression. Why or why not? **(0.75 pts)**

7. I want to account for the fact that different levels of education are associated with different types of accreditation. For example, the impact of the change from 10 to 11 years of education may be different than the change from 11 to 12 years of education, due to the fact that individuals with 12 years of education tend to have received a high school diploma. I'm considering accounting for this impact in my regression. Why might I choose a step function rather than the standard (non-spline) polynomial regressions? **(0.75 pts)**

8. How should we determine how many knots or degrees of freedom to use for a regression spline? **(0.5 pts)**

9. What is the main advantage of natural regression splines compared to standard polynomial regressions? **(0.5 pts)**

10. For smoothing splines, if the smoothing parameter λ=0, the function will be very smooth or rough? If λ=∞, the function will be very smooth or rough? **(0.5 pts)**

Data Questions

For this section, use the `STAR` data available from the `AER` package.

We are interested in predicting third grade reading scores (read3). Drop the following columns: birth, readk, read1, read2, mathk, math1, and math2. Also, make sure all numeric values are saved as type numeric rather than type integer (be careful in making this conversion to ensure values are preserved). Note that if you have trouble with this, Google is a good resource. Of course, the teaching staff is happy to help you during office hours if you have trouble.

1. Data Cleaning **(1 pt)**

   (a) How many predictors are in the data set (after you drop the variables according to the directions)?

   (b) How many observations have missing values? Drop those observations.

   (c) How many categorical variables are in the data set? If the variable is a binary, replace it with a numeric where 1 indicates "yes" and 0 indicates "no". Convert all categorical string variables to a complete set of indicator variables (e.g. if the variable takes on four unique values, make four indicator variables).

   (d) How many of the variables (including the newly generated variables) have variance <0.05? Drop all columns with variance <0.05. This step ensures your subsequent code will run without erroring.

   After completing the previous question, randomly split the data into a training set using `set.seed(13194)`(1894 observations) and a test set (remaining observations). All models should be trained only on the training data, so when it asks you to use cross validation (CV), you should run CV only using training data. Use 10 Fold CV

2. Run Principal Components Regression on the training data.

   (a) Would you say the model with 10 principal components shows a big improvement in CV RMSE over the model with 0 principal components?

   (b) What about the model with 20 principal components compared to the model with 10 principal components?

   (c) What are the CV RMSE associated with 0, 10, and 20 components.

   Note: RMSE is Root Mean Squared Error. RMSE is simply the square root of the MSE. **(1 pt)**

3. Evaluating Principal Components Regression **(1.5 pts)**

   (a) How many principal components yields the best CV RMSE?

   (b) What is the corresponding CV RMSE?

   (c) What is the test RMSE?

4. Run Partial Least Squares on the training data. **(1.5 pts)**

   (a) What number of components correspond to the lowest CV RMSE?

   (b) What is the CV RMSE (using only your training data)?

   (c) What is the test RMSE (using only the test data)?

5. Run a decision tree on your training data and use cross-validation to choose the optimal size. **(2 pts)**

   (a) What is the optimal size?

   (b) Include a plot of the deviance compared to the size. As a reminder, a smaller deviance indicates that the tree provides a good fit for the data.

   (c) What is the test RMSE?

6. If you had to predict Third grade reading scores using one of the methods tried in this problem set, which method would you use and why? **(1 pt)**