Problem Set 3

Cian Stryker

10/16/2020

Concept Questions

Question 1

You should standardize your variables for Ridge, Lasso, KNN, PCR, and PLS. You do not need to standardize your variables when doing a logistic or linear regression.

Question 2

We cannot conclude that we should choose Model B. This is because model B's AUC (Area Under The Curve) is higher and, generally speaking, while a higher AUC indicates a model is better at predicting, it also has a higher chance for false positives. Model A has a lower AUC than Model B, which means it is further down on the ROC curve. This indicates that model A will likely produce fewer false positives than model B. We should therefore choose Model A because while it is worse at prediction overall, our primary goal is only to minimize the number of false positives and model A is better suited to this goal.

Question 3

If we run a Lasso with a lambda of 0 the model is equivalent to an OLS regression. As lambda approaches infinity, however, the shrinkage penalty grows and all coefficients will be reduced to zero, removing them from the model itself.

Question 4

When the number of principal components in a model is high and close to the number of features in the data set, Ridge and Lasso regressions are likely to outperform PCR.

Question 5

Yes I would conclude that X2 will receive more weight in the PLS regression because it is more highly correlated to our outcome variable (Y). PLS takes the correlation of predictors to the outcome variable into account when allocating weight, therefore, x2's higher correlation to Y would indicate it will receive more weight.

Question 6

No I would not necessarily conclude that X2 will receive more weight in the PCR regression. PCR does not take the relationship of predictors to the outcome variable (Y) into account when allocating weight, therefore, even though X2 has a higher correlation to Y, PCR will not take that into account when allocating weight.

Question 7

A polynomial non-spline regression imposes a global structure to data that doesn't necessarily take into account the differences between certain segments of the data range. A step function might be more appropriate for this situation because it divides the range of data into different bins or k distinct regions, each with a different constant. In the situation given, we could create a bin that shows years 10-11 and 11-12 separately which may show if that 12th year really does have a significant impact.

Question 8

When we determine how many knots or degrees of freedom to use for a regression spline we should use cross validation to try different degrees of freedom or knots (k values) and then choose the number with the lowest RSS.

Question 9

Natural regression splines add additional constraints at the tails of a model, which decrease the variance at the tails. In polynomial regressions the variance at the tails is often very high because data is limited. Natural regression splines address this issue and produces more stable estimates at the boundaries.

Question 10

For smoothing splines, if the smoothing parameter is 0 the function will be very rough, but if the parameter is equal to infinite it will become perfectly smooth.

Data Questions

1.)

a.) There are 40 predictors in the data set.

b.) 9230 observations were missing values.

c.) 34 variables are categorical variables, while six are numeric.

d.) 462 variables have variance less than 0.05.

2.)

a.) I would not say that the model with 10 principal components shows a big improvement in CV RMSE over the model with 0 components because the difference between the two is only 2.27

b.) The model with 20 principal components has a slightly lower CV RMSE than the model with 10, but only 0.24 lower, which is a very small improvement.

c.) 0 components has a CV RMSE of 36.91, 10 components has a CV RMSE of 34.64, and 20 components has a CV RMSE of 34.30.

3.)

a.) 39 component principled components corresponds to the lowest CV RMSE.

b.) The CV RMSE of 39 component parts is 24.944

c.) The test RMSE for the test data and 38 components is 25.890.

4.)

a.) 7 component principled components corresponds to the lowest CV RMSE.

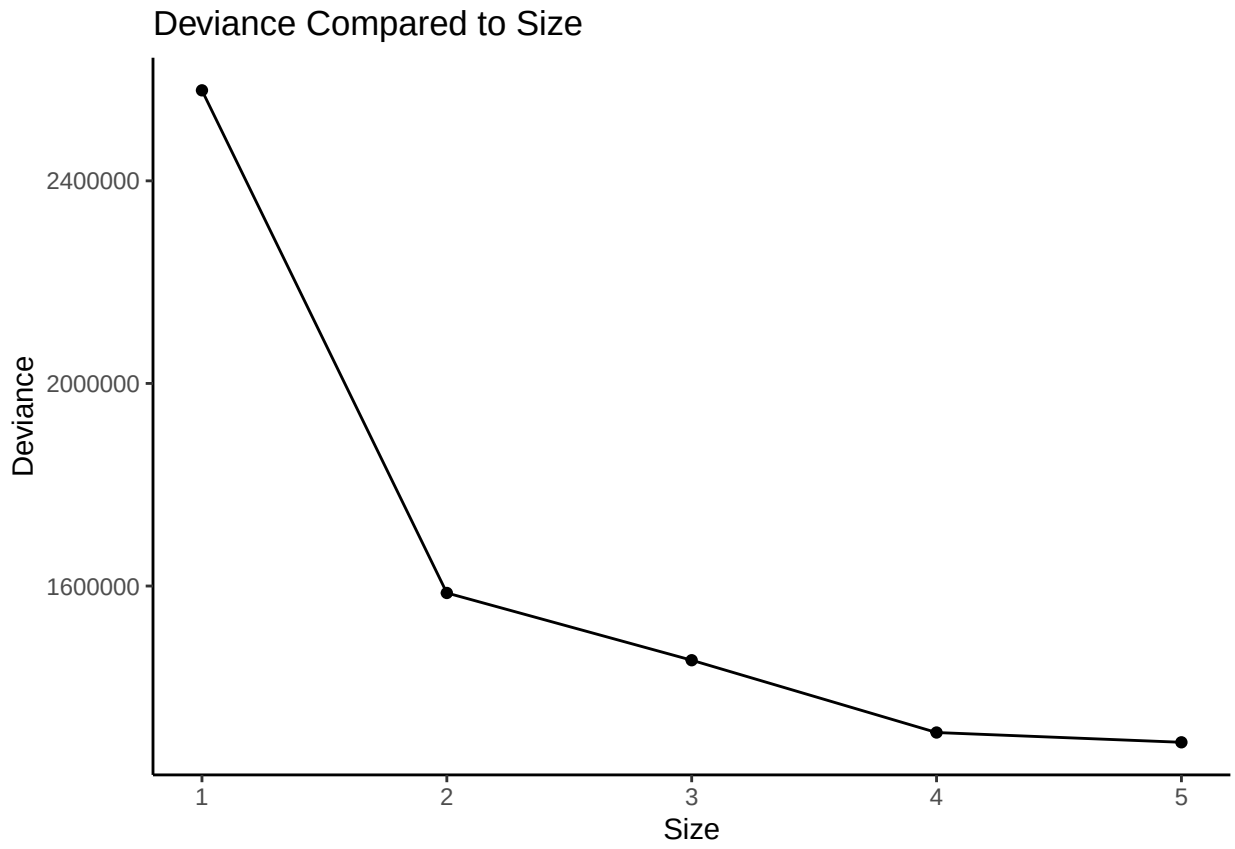
b.) The CV RMSE of 7 component parts is 24.958

c.) The test RMSE for the test data and 7 components is 25.853.

5.)

a.) The optimal size is 5.

b.)



c.) The test RMSE is 26.309

6.)

If I had to predict third grade reading scores using one of the methods tried in this problem set, I would choose PLS over PCR or decision trees because its test RMSE was the lowest of the three models.