

Problem Set 3

Cian Stryker

10/16/2020

Concept Questions

Question 1

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

Question 2

We can conclude that we should choose Model B because the AUC (Area Under The Curve) is higher and, generally speaking, when the AUC is higher the model is better at predicting 0s as 0s and 1s and 1s. So since model B has a higher AUC than Model A, we should choose Model B because it will have fewer false positives.

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 coefficients will be reduced to zero, removing them from the model itself.

Question 4

When we have data where the outcome variable is strongly correlated with our predictors, Ridge and Lasso will likely outperform PCR. PCR does not take the outcome variable into account when deciding principal components so it may generate components that explain the most variance overall, but with little relationship to our outcome variable. Similarly, it may have dropped predictors that are highly correlated to the outcome variable in its first step.

Question 5

Yes I would conclude that X_2 will receive more weight in the PLS regression because it is more highly correlated to our outcome variable Y . PLS takes correlation of variables to the outcome variable into account when allocating weight.

Question 6

No I would not necessarily conclude that X_2 will receive more weight in the PCR regression. PCR does not take the relationship of variables to the outcome variable Y into account when allocating weight, therefore, just because X_2 has a higher correlation to Y , PCR will not take that into account.

Question 7

A polynomial regression imposes a global structure to data that doesn't necessarily take into account the differences between certain parts of the data range. A step function might be more appropriate for this situation because it divides the range of data into different bins. 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

The number of knots used in a regression spline is often determined by using cross validation to try different K values and then choose the number with the best scores.

Question 9

Natural regression splines add additional constraints that at the tails of a model the model must be linear, which decreases 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.

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 very 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.) 498 variables have variance less than 0.05.