Stat 897 Fall 2017 Data Analysis Assignment 7

Penn State

Due October 15, 2017

In this assignment we will use the Khan data found in the ISLR library. This dataset contains 2308 gene expressions on 83 individuals (already split into test and train) with a response variable of tumor measurement. We will use dimension reduction in prediction and compare results with shrinkage methods.

a) The data is alreay split into test and train in the `Khan` object. First, combine the xtrain and xtest datasets and run PCA on the combined predictor data set. Plot the components versus percentage variance explained.

b) Fit a PCR model on the training set (you will need library `pls`), with M chosen by 5-fold cross-validation. Report the test MSE obtained, along with the value of M selected by cross-validation. What is the percent variance explained on the combined predictor dataset for the number of components chosen here?

Please set the seed with "34"" here at the beginning of your code for part (b) and only here.

```
set.seed(34)
```

c) Fit a PLS model on the training set, with M chosen by 5-fold cross-validation. Report the test MSE obtained, along with the value of M selected by cross-validation. What is the percent variance explained on the combined predictor dataset for the number of components chosen here?

d) Perform Lasso and Ridge with lambda chosen by 5-fold CV. Compute the test MSE. How do your results compare to the PCR and PLS results? Which model would you prefer? (Think about both prediction quality and inference.)