BMTRY 790 Spring 2023: Assignment 2

Exercises

1. Recall the loss function for ordinary least squares regression, $L(\beta) = (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta)$.

   i. Show that the solution that minimizes $L(\beta)$ is $\hat{\beta}^{ols} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$. Note that for matrix derivatives, if a and b are $K \times 1$ vectors, then $\frac{\partial a^T b}{\partial b} = \frac{\partial b^T a}{\partial b} = a$. Also if b is a $K \times 1$ vector and $\mathbf{A}$ is a $K \times K$ symmetric matrix then $\frac{\partial b^T \mathbf{A} b}{\partial b} = 2\mathbf{A}b = 2b^T\mathbf{A}$.

   ii. Similarly, given the loss function for ridge regression, $L(\beta) = (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) + \lambda\beta^T\beta$, show that the solution that minimizes the squared error loss function is $\hat{\beta}^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T y$.

   iii Show that the ridge regression estimates can also be obtained by ordinary least squares regression on an augmented dataset in which the centered $\mathbf{X}$ matrix includes p additional rows $\sqrt{\lambda}\mathbf{I}$, and augmented y with p zeros. How does this relate to *ElasticNet*?

2. A study was conducted to examine the association between prostate specific antigen (PSA) with different clinical measures in patients with prostate cancer. Clinical markers in the data include log(cancer volume), log(prostate weight), patient age, log(amount benign prostatic hyperplasia), seminal vesicle invasion, log(capsular penetration), Gleason score, percent of Gleason 4 or 5, and log(PSA). Download the data from the class website and split the data into $\frac{2}{3}$ training and $\frac{1}{3}$ test sets. Fit OLS, ridge, lasso, elasticnet, pcr, and plsr models to the training data. Compare discuss the model coefficients from each model and prediction performance on the test data.

3. Below is a simple R function to generate data in R.

```
gendat<-function(n, decay)
{
  x1<-runif(n, 0, 1)
  x2<-x1+rnorm(n, 0, decay)
  x3<-x2+rnorm(n, 0, decay)
  x4<-x3+rnorm(n, 0, decay)
  x5<-runif(n, 0, 1)
  x6<-runif(n, 0, 1)
  x7<-x6+rnorm(n, 0, decay)
  x8<-x7+rnorm(n, 0, decay)
  x9<-x8+rnorm(n, 0, decay)
  x10<-runif(n, 0, 1)
  y<-x1+x2+x3+x4+x5+0.1*x6+0.1*x7+0.1*x8+0.1*x9+0.1*x10+rnorm(n, 0, 5)

  dat<-cbind(y,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10)
  colnames(dat)<-c("y", paste("x", 1:10, sep=""))
  return(dat)
}
```

i. Describe the data generated by the above function, include a discussion of the impact of changing the magnitude of the decay parameter.

ii. Statisticians often run simulations to compare performance characteristics of different methods. Write a function to conduct a simulation to compare parameter estimation and prediction performance of OLS with (i) one of ridge, lasso, or elastic net, and (ii) one of PRC or PLSR on training and test sets in data with different levels of correlation between the x_j's. At a minimum, simulation input parameters should include sample size for training and test data, decay, and number of repetitions. Output should include the mean squared error for the training and test sets and the coefficient values from each iteration. Note, training and test data should be generated within the simulation function using the data generation function above. Additionally, you need to consider how you will select your final models for the different methods.

iii. Using your simulation function, conduct a simulation to compare prediction performance between the three methods included in your simulation function for training data with $n = 200$, test data with $n = 100$, and $decay = 1, 0.2,$ and $, 0.01$ for 200 iterations.

a. Plot the mean training and test error for each method. Also plot the mean estimated coefficient values under each scenarion and te mean squared error of the coefficient values.

b. Discuss your results in terms of the impact of correlation on prediction performance of each of the methods base don prediction error in test and training data and on error in the coefficient estimates.