BIOST 2079 Homework 1

Distributed: 11/1/2023
Deadline: 11/10/2023

Theory

1. Derive the solutions of OLS and ridge regression in matrix form in the course slide: (5 points)
   (a) In OLS, we derive $\min_\beta (Y - X\beta)^T(Y - X\beta)$. Prove that $\hat{\beta}^{OLS} = (X^TX)^{-1}X^T$Y. (hint: For a given A, $\frac{\partial}{\partial\beta}(A\beta) = A$ and $\frac{\partial}{\partial\beta}(\beta^TA\beta) = 2A\beta$)
   (b) In ridge regression, we aim on $\min_\beta (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$. Prove that $\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^T$Y.
   (c) From the last question, prove that OLS is scale invariant but ridge regression is not. Scale invariant means that changing the scale of X does not change the prediction. For here, prove a simpler case where the entire design matrix X is multiplied by a constant (i.e. $X' = c \cdot X$).

2. If *X* is an orthogonal matrix (i.e. the predictors are uncorrelated: $Cov(X_s, X_t) = 0$ if $s \neq t$), prove that $\hat{\beta}_j^{ridge} = \hat{\beta}_j^{OLS}/(1+\lambda)$, $\hat{\beta}_j^{lasso} = sign(\hat{\beta}_j^{OLS})(|\hat{\beta}_j^{OLS}| - \lambda/2)_+$, where $(x)_+=0$ if $x<0$ and $(x)_+=x$ if $x \geq 0$. For simplicity, you may prove the one-dimensional case (p=1). (3 points)

Computing

* You are encouraged to use R Markdown (template provided) to generate pdf reports with embedded R codes and outputs.

3. This question is modified from Question 10 in Chapter 3.7 in ISLR. This question should be answered using the Carseats data set from the R package ISLR. (7 points)
   (a) Fit a multiple regression model to predict Sales using Price, Urban, and US.
   (b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!
   (c) Write out the model in equation form, being careful to handle the qualitative variables properly.
   (d) For which of the predictors can you reject the null hypothesis $H_0: \beta_j = 0$?
   (e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.
   (f) How well do the models in (a) and (e) fit the data?
   (g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).
   (h) Using the leave-one-out cross-validation and 5-fold cross-validation techniques to compare the performance of models in (a) and (e). What can you tell from (f) and (h)?

4. This question pertains to a prostate microarray dataset. You can access it by load("prostate.Rdata"). It has been preprocessed to have 210 gene and 235 samples. Lpsa value is the clinical outcome we want to predict. (5 points)
   (a) Randomly divide the data into one training dataset and one testing dataset (1:1).
   (b) Fit a linear model using OLS on the training dataset and calculate the test error in terms of RMSE. Report any problems you encountered.
   (c) Use ridge regression. Find the optimal lambda which will return the smallest cross validation error using the training data.
   (d) Build the ridge regression model using the training data and the lambda in (c) and then predict test error in terms of RMSE.
   (e) Repeat steps in (c) and (d) using lasso. Derive the RMSE in the testing dataset.