Regularization: Ridge regression, the LASSO, and Elastic Net

Fall 2022, MATH8050: Homework 10
Your Name, Section XXX

Due November 30, 12:00 PM

General instructions for homeworks: Please follow the uploading file instructions according to the syllabus. Each answer must be supported by written statements as well as any code used. Your code must be completely reproducible and must compile. For writing mathematical expressions in R Markdown, refer to the homework template posted on Canvas, a 30-minute tutorial, or LaTeX/Mathematics.

Advice: Start early on the homeworks and it is advised that you not wait until the last day. While the professor and the TA's check emails, they will be answered in the order they are received and last minute help will not be given.

No late homeworks will be accepted.

R Working Environment

Please load all the packages used in the following R chunk before the function `sessionInfo()`

```
# load packages


sessionInfo()
```

Total points on assignment: 10 (reproducibility) + 90 (Q1)

Reproducibility component: 10 points.

1. (90pts total, equally weighted) Comparison of ridge regression, lasso regression, and elastic net regression.

We'll construct a matrix with n = 1000 observations (you can think about these as 1000 participants) and p = 2000 predictors.

```
set.seed(123)

n <- 1000
p <- 2000

pred <- matrix(rnorm(n*p), nrow = n, ncol = p)
```

We created an outcome variable "dv" which is calculated, first, by the sum of predictors 1 through 5. Then by adding the sum of predictors 6 through 10 (but only influenced by these rows 80% as much as by the first 5). Then 11 through 15, time (60% as much) and so forth. Additionally, some random noise is added to each outcome variable.

```
dv <- (rowSums(pred[,1:5]) + .8*rowSums(pred[,6:10]) +
         .6 * rowSums(pred[,11:15]) + .4*rowSums(pred[,16:20]) +
         .2 * rowSums(pred[,21:25]) + rnorm(n))
```

We then center the predictors

```
pred <- scale(pred)
```

Let's now split our data (both the predictors and dv) into train and test sets. We'll go with a 80-20 split.

```
train_rows <- sample(1:n, .8*n, replace = F)

pred.train <- pred[train_rows,]
dv.train <- dv[train_rows]

pred.test <- pred[-train_rows,]
dv.test <- dv[-train_rows]
```

a. Perform the ridge regression using the package `glmnet`, in which you need to set 200 different lambda values (the default for the function `glmnet` is 100). You need to generate the plot with x-axis being the log of lambda and y-axis being the training MSE. Please also report the MSE on the testing dataset with the model generated with a lambda equal to `lambda.lse` (use `?glmnet` for its explanation.)

b. Repeat the same steps above with Lasso Regression: fit the lasso regression with `nlambda=200`, plot the training MSE against log-lambda, generate prediction on testing set using the model with `lambda.lse` and compute MSE on testing set.

c. Perform Elastic Net Regression: fit the elastic net regression with `alpha` set to be `seq(0, 1, by=1/20)`, plot the testing MSE against each alpha value