STSCI 3740/5740 Machine Learning and Data Mining **Fall 2024**

Nayel Bettache **Homework 4, due 15 December, 11:59pm**

Problem 1 (8 points)

In this problem you will compare the performance of lasso and ridge regression in different linear models. Use the following code to generate predictors X_train and errors eps_train in the training set and X_te and eps_te in the test set

```
p = 50
N = 100
set.seed(1)
X_train = array(rnorm(p*N),c(N,p))
eps_train = rnorm(N)
Nte = 10^3
X_te = array(rnorm(p*Nte),c(Nte,p))
eps_te = rnorm(Nte)
grid = 10^seq(10,-2,length = 100)
```

1. Create test and training data based on the model

$$Y_i = \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i$$

with $\beta_1 = ... = \beta_5 = 2$, and $\beta_j = 0, j > 5$. Fit a ridge regression and a lasso model with λ selected by cross validation on the grid defined above. Which method leads to a smaller test error?

2. Repeat the steps in the first part with $\beta_j = 0.5$ for $j = 1, ..., 50$. Which method performs better now?

3. Use a **for** loop to repeat part 1 with seeds set.seed(2),...,set.seed(50). Save the test error for both, lasso and ridge for all seeds. Together with the results from part 1, this should give you 50 test errors for each procedure. Make boxplots of the test errors and comment on the results. (Hint: you need to re-generate data under each seed.)

4. Use a **for** loop to repeat part 2 with seeds set.seed(2),...,set.seed(50). Save the test error for both, lasso and ridge for all seeds. Together with the results from part 2, this should give you 50 test errors for each procedure. Make boxplots of the test errors and comment on the results. (Hint: you need to re-generate data under each seed.)

Problem 2 (10 points)

Solve Problem 8 on page 363 in the textbook "Introduction to Statistical Learning" (second edition).