- 제출일 : 9월 9일 일요일 오후 11시 59분까지
- R markdown으로 작성한 보고서 형식의 파일 (.pdf 또는 .docs)과 작성한 R markdown 파일 (.Rmd) 두 가지를 모두

  bdi.bd.test@gmail.com 로 제출합니다.
- 파일명, 메일제목은 이름으로 해주세요.

문제 1

It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose that $n = 2, p = 2, x_{11} = x_{12}, x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

1. Write out the ridge regression optimization problem in this setting.
2. Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.
3. Write out the lasso optimization problem in this setting.
4. Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique—in other words, there are many possible solutions to the optimization problem in 3. Describe these solutions.

문제 2

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

1. Which answer is correct, and why?

   (a) For a fixed value of IQ and GPA, males earn more on average than females.

   (b) For a fixed value of IQ and GPA, females earn more on average than males.

   (c) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

   (d) For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

2. Predict the salary of a female with IQ of 110 and a GPA of 4.0.

3. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

문제 3

강의 홈에 제공된 ‘‘`data3.xlsx`’’에 대하여 다른 변수들을 이용하여 “ideo_self”를 예측하는 모형을 구축 하고자 한다. 본인이 찾은 최적의 예측모형을 서술하고 10-fold cross-validation 을 이용한 시험오차(testing error)에 대한 혼동행렬(confusion matrix)을 계산하여라.

문제 4

강의 홈에 제공된 ‘‘data4.csv’’에 대하여

1. V2-V51의 자료를 군집의 갯수를 2개로 고정시킨 “적절한” 군집분석을 수행하여라. 수행한 군집분석을 설명하고 결과를 군집분석의 결과를 적는다.

2. V2-V51을 이용하여 V1의 class를 예측하는 binary classification문제를 고려하여 최적의 classifier를 찾고 10-fold closs-validation을 이용한 시험오차(testing error)에 대한 혼동행렬(confusion matrix)을 계산하여라.

문제 5

강의 홈에 제공된 ‘‘data5.xlsx’’을 이용하여 의원 들의 군집을 파악하여 보아라. 군집을 찾아가는 과정에 대한 설명을 제시함으로서 분석의 타당성을 이야기 하여야 한다. 자료의 대각원소는 해당 의원이 단독으로 (대표)발의한 법안의 수이다.