Homework 3

MA506 Probability and Statistical Inference: Fall 2022

Due: November 7 (Monday), 11:59pm

100 points

Model Selection in Regression

For this assignment, we will be analyzing different regression models for a given dataset. The Regression models we will explore are as follows:

- Linear Regression
- Ridge Regression
- Lasso Regression

Here, we will be using the diabetes dataset (it is available from sklearn *load_diabetes*()). Each row/sample in this dataset contains 10 features containing different information about the patients. These 10 features are to be used to predict the target variable which represents a quantitative measure of disease progression for each patient. Hence it is a regression problem. The total number of samples are 442.

Question 1: Preprocessing (5 points)

1. Divide the dataset into 2 parts: D_{tr} (training set), and D_{te}(testing set) by randomly placing 80% of the data into D_{tr} and 20% in D_{te}. For this, use train_test_split() from sklearn with a random state of 0.

Question 2: Linear Regression (45 points)

1. (**2 points**) Divide D_{tr} further into 2 different sets: D_{train} (training set) and D_{val} (validation set). Place 80% data in D_{train} and remaining in D_{val}. Again use random state 0.
2. (**32 points**) Fit the following 4 models separately on D_{train}:
   (a) $pred = \beta_0 + \beta_1 \cdot bmi + \beta_2 \cdot bp$
   (b) $pred = \beta_0 + \beta_1 \cdot bmi + \beta_2 \cdot s5$
   (c) $pred = \beta_0 + \beta_1 \cdot bp + \beta_2 \cdot s5$

(d) $pred = \beta_0 + \beta_1 \cdot bmi + \beta_2 \cdot bp + \beta_3 \cdot s5$

and print the R^2 value of each of these model on both D_{train} and D_{val}.

3. (**6 points**) Choose the best model based on your analysis from previous part and fit that model on D_{tr} and display the R^2 on D_{tr} and D_{te}

4. (**5 points**) Explain the problems with the model selection approach you implemented in part 1, 2 and 3 of this question.

Question 3: Ridge Regression (30 points)

For the linear model:

$$pred = \beta_0 + \beta_1 \cdot bmi + \beta_2 \cdot bp + \beta_3 \cdot s5$$

with X and Y containing the appropriate data, we now wish to fit a ridge regression model which has the following optimal weights:

$$\hat{\beta}^{ridge} = (X^T X + n\lambda I)^{-1} X^T Y$$

Hence for fixed X and Y matrices, the problem of fitting a ridge regression model boils down to finding the right λ. In this regard:

1. (**20 points**) Use Generalized Cross Validation (CV) metric to compute and display the best λ on D_{tr}.

2. (**5 points**) Using the best obtained λ, Compute and display the corresponding $\hat{\beta}^{ridge}$ on D_{tr}.

3. (**5 points**) Using $\hat{\beta}^{ridge}$ from previous part, compute and display R^2 on D_{te}

Question 4: Lasso Regression (20 points)

Again for the linear model:

$$pred = \beta_0 + \beta_1 \cdot bmi + \beta_2 \cdot bp + \beta_3 \cdot s5$$

with X and Y containing the appropriate data, we now wish to fit a lasso regression model. Here again the objective of fitting a lasso regression model boils down to finding the right λ. In this regard:

1. (**10 points**) On D_{tr}, Use LassoCV from sklearn (again use a random state of 0) to do a K-fold cross validation based selection of best λ. Display the best λ you obtain. Use K = 5 here.

2. (**5 points**) Using the best obtained λ in part 1 above, compute the predictions on D_{te} and display the corresponding R^2 on D_{te}

3. (**5 points**) Comment on the relative performance of linear regression, ridge regression and lasso regression for this problem.