RSM8512 Assignment - Model Selection

Yanbing Chen 1009958752 2023/10/19

Question 1 [20 marks]

We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p+1$ models, containing $0, 1, 2, \ldots, p$ predictors. Explain your answers.

(a) Which of the three models with k predictors has the smallest training RSS?

(b) Which of the three models with k predictors has the smallest test RSS?

(c) True or False:

i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.

(ii) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$ -variable model identified by backward stepwise selection.

(iii) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$ -variable model identified by forward stepwise selection.

(iv) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$ -variable model identified by forward stepwise selection.

(v)The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k+1)$ variable model identified by best subset selection.

Question 3 [10 marks]

Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\Sigma_{i=1}^{n}(y_i - \beta_0 - \Sigma_{j=1}^{p}\beta_j x_{ij})^2$$

subject to

$$\Sigma_{j=1}^{p} |\beta_j| <= s$$

For a particular value of s. For parts (a) through (e), indicate which of i.through v. is correct. Justify your answer.

(a) As we increase s from 0, the training RSS will:

i. Increase initially, and then eventually start decreasing in an inverted U shape. ii Decrease initially, and then eventually start increasing in a U shape.
ii. Steadily increase.
iii. Steadily decrease.
iv. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

Question 8 [30 marks]

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

(a) Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as a noise vector ε of length $n = 100$.

(b) Generate a response vector Y of length $n = 100$ according to the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$ where $\beta_0, \beta_1, \beta_2, \beta_3$ are constants of your choice.

(c) Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, ..., X^{10}$. What is the beset model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the `data.frame()` function to create a single data set contraining both X and Y.

(d) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

(e) Now fit a lasso model to the simulated data, again using $X, X^2, ..., X^{10}$ as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of $lambda$. Report the resulting coefficient estimates, and discuss the results obtained.

(f) Now generate a response vector Y according to the model

$$Y = \beta_0 + \beta_7 X^7 + \varepsilon$$

and perform best subset selection and the lasso. Discuss the results obtained.

Question 10 [20 marks]

We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

(a) Generate a data set with $p = 20$ features, $n = 1000$ observations, and an associated quantitative response vector generated according to the model

$$Y = X\beta + \varepsilon$$

where β has some elements that are exactly equal to zero.

(b) Split your data set into a training set containing 100 observations and a test set containing 900 observations.

(c) Perform best subset selection on the training set, and plot the traning set MSE associated with the best model of each size.

(d) Plot the test set MSE associated with the best model of each size.

(e) For which model size does the test set MSE take on its minimum value? Comment on your results. If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.

(f) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficeint values.

(g) Create a plot displaying $\sqrt{\Sigma_{j=1}^{p}(\beta_j - \hat{\beta}_j^r)^2}$ for a range of values of r, where $\hat{\beta}_j^r$ is the jth coefficient estimate for the best model containing r coefficients. Comment on what you observe. How does this compare to the test MSE plot from (d)?