

Lab #7: Penalized Regression

Biostat 216: Machine Learning in R for the Biomedical Sciences

Subset Selection

1. Install the `leaps` package
2. Load the `leaps` package
3. Read in the breast cancer imaging data “`ispy1doctored.csv`” into a data frame called `dat`
4. Perform best subset selection using the default parameters with `MRI_LD_Tfinal` as outcome (the longest diameter of the breast cancer tumor at the final visit pre-surgery) and all other variables as predictors. Look at a summary of the fitted models. What is the max number of predictors in any model imposed by the default settings?
5. How many predictors do you have available in the dataset?
6. Perform best subsets selection allowing models that could include all predictors
7. Output a list of names of the objects in the summary of best subset selection fit, and examine the output of summary.
8. Generate plots for each of the metrics R^2 , Adjusted R^2 , C_p , and BIC , vs. number of variables in the model. Add a point to the plot indicating where optimal point is. Tip: writing a function may save you some typing/clutter in your code.
9. Now perform forward and backward stepwise selection allowing for potentially the full set of predictors in the models
10. Compare the coefficients of the different models based on restricting the number of variables from 1 through 8. Tip: writing a loop will save you a lot of typing. What is the largest size of model where all 3 methods agree? What is the smallest size model where all 3 methods disagree?
11. Try the plot function associated with `regsubsets` on each of the 3 selection procedures (best subset, forward, and backward). Use the `scale` argument to plot R^2 . Feel free to also look at the corresponding plots for the Adjusted R^2 , C_p , and BIC statistics too. Again you may want to play with writing a function to avoid clutter.

Ridge regression

12. Install the package `glmnet`
13. Load the `glmnet` package
14. Convert the data into components `x` and `y` that are suitable for use in the `glmnet` package (we will continue to consider `MRI_LD_Tfinal` as the outcome)
15. Generate a set of values to try for `lambda` from $\lambda = 10^{10}$ to $\lambda = 10^{-2} = 0.01$
16. Fit ridge regression to the dataset with `MRI_LD_Tfinal` as outcome and all other variables as candidate predictors for all of the `lambda` values you have created
17. Look at the value of `lambda` associated with `ridgeMod[75]`
18. Display the fitted coefficients associated with this value of `lambda`
19. And then the l_2 norm associated with this value of `lambda`

20. Repeat steps 17 to 19 for `lambda` associated with index 55. Is the difference in the l_2 norm in the direction that you would expect? Why/why not?
21. Use the `predict` command to find the ridge regression coefficients for $\lambda = 3000$

Lasso

22. Fit lasso regression to the dataset with `MRI_LD_Tfinal` as outcome and all other variables as candidate predictors for the same set of lambda values you considered in fitting ridge regression
23. Repeat steps 17 to 19 (again with index 75) but this time for the lasso fit
24. We have not performed any evaluation of prediction error here or comparison of methods. Outline the steps to compare the predictive accuracy in terms of mean squared error (MSE) between ridge regression and lasso using training/validation/test partitions of the data.
25. The dataset is not particularly large here. How might cross-validation be used to make better use of the full data?