Tutorial 7

Quentin Bouet

2024-09-04

Model Selection and Regularisation

In the last week lecture, we focused on regression analysis methods with embedded features of model selection. Specifically you've seen some methods in dealing with multiple and often competing predictor variables. To avoid overfitting your model, resampling methods are a must to employ. This week we focus on regression method to avoid NP problems. In the lecture, we looked • Subset selection (best subset selection & stepwise selection) • Shrinkage methods (ridge regression & the lasso) • Dimension reduction methods (PCR & PLS) In this tutorial, we will see how we can apply these methods in R.

Discuss the following topics

- What are advantages and disadvantages of linear model?
- What is N-P problem? Why is it a problem?
- Propose three solutions to overcome N-P problem
- What are the difference between forward and backward subset selection?
- Explain why backward subset selection can still subject to NP problem
- Explain how shrinkage methods work. How is it different from OLS?
- Discuss the differences between ridge and lasso regression?
- What is the role of complexity parameter in ridge or lasso regression?
- Discuss how complexity parameter regulates the model structure?
- How do you choose the value of complexity parameter?
- Explain how Principal component regression works

Independendent Learning

Exercises

- Question 2, Chapter 6.6
- Question 3, Chapter 6.6
- Question 4, Chapter 6.6
- In this exercise, we will predict the number of applications received using the other variables in the College data set.

1. Split the data set into a training set and a test set.
2. Fit a linear model using least squares on the training set, check assumptions and report the test error obtained.
3. Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.
4. Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates.
5. Fit a PCR model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selcted
6. Fit a PLS model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selcted
7. Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

- We will now try to predict per capita crime rate in the Boston data set.

1. Try out some of the regression methods explored in this week, such as best subset selection, the lasso, ridge regression, principal component regression and partial least square regression. Present and discuss results for the approaches that you consider.
2. Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross validation, or some other reasonable alternative, as opposed to using training error.
3. Does your chosen model involve all of the features in the data set? Why or why not?