Northeastern University
College of Professional Studies

Regularization

Overview and Rationale

In order to consolidate your theoretical knowledge into technique and skills with practical and applicational value, you will use the glmnet() package in R to implement Ridge and LASSO functions to build linear and logistic models through Ridge and LASSO regression over values of the regularization parameter lambda.

Course Outcomes

This assignment is directly linked to the following key learning outcomes from the course syllabus:

- Conduct regularization method for models to describe relationships among variables and make useful predictions.

Assignment Summary

Use the College dataset from the ISLR library to build regularization models by using Ridge and Lasso (least absolute shrinkage and selection operator). Predict Grad.Rate for all models.

1. Split the data into a train and test set – refer to the Feature_Selection_R.pdf document for information on how to split a dataset

Ridge Regression

2. Use the *cv.glmnet* function to estimate the lambda.min and lambda.1se values. Compare and discuss the values.
3. Plot the results from the *cv.glmnet* function provide an interpretation. What does this plot tell us?
4. Fit a Ridge regression model against the training set and report on the coefficients. Is there anything interesting?
5. Determine the performance of the fit model against the training set by calculating the root mean square error (RMSE). sqrt(mean((actual - predicted)^2))
6. Determine the performance of the fit model against the test set by calculating the root mean square error (RMSE). Is your model overfit?

LASSO

7. Use the cv.glmnet function to estimate the lambda.min and lambda.1se values. Compare and discuss the values.
8. Plot the results from the *cv.glmnet* function provide an interpretation. What does this plot tell us?
9. Fit a LASSO regression model against the training set and report on the coefficients. Do any coefficients reduce to zero? If so, which ones?
10. Determine the performance of the fit model against the training set by calculating the root mean square error (RMSE). sqrt(mean((actual - predicted)^2))
11. Determine the performance of the fit model against the test set by calculating the root mean square error (RMSE). Is your model overfit?

Comparison

12. Which model performed better and why? Is that what you expected?

13. Refer to the Intermediate_Analytics_Feature_Selection_R.pdf document for how to perform stepwise selection and then fit a model. Did this model perform better or as well as Ridge regression or LASSO? Which method do you prefer and why?

Report

Refer to the attached rubric for more details on the report. The report should contain a well written cover/title page, introduction, body, conclusion, and references. It must follow APA format and have at least 1000 words (excluding title page and references page. All R code used for your report should be included in an appendix at the end of the report.

Graphs, figures, charts, and tables are very useful visual effects to communicate your results and impress your readers. However, such items should not be included in the report unless they are well described and interpreted. Please use subtitles to make your assignment more reader friendly as well.

Be sure to include your code in the Appendix.

Format & Guidelines

The report should follow the following format:

(i) Title page

(ii) Introduction

(iii) Analysis

(iv) Conclusion/Interpretations

(v) References

(vi) Appendix

Assignment Rubric

Category	Above Standards	Meets Standards	Approaching Standards	Below Standards
Introduction **15%**	Clearly and briefly introduces the goals of the project, the question that needs to be answered and the methods used in the analysis. The goals, questions and methods outlined are consistent with one another.	Introduction provides a brief and intelligible overview of the goals and methods of the assignment.	Introduction provides an overview of the goals and methods of the assignment, but is ambiguous or not concise.	Does not introduce project goals, project questions or methods.
Analysis **25%**	Incorporates R code and the outputs. Provides detailed analysis of the output focusing on significance results. Uses visualizations to make major points.	Provides all R code and the outputs. Includes interpretation of the output, graphs, figures, charts, and tables and the significance of the results in the analysis.	Provides R codes and outputs, but the R code does not match the outputs or is missing some code or outputs. Includes limited interpretations, charts, and tables and the significance of the results in the analysis.	Does not provide R code or its outputs or minimal R code is provided. Includes few interpretations, charts, or tables. Does not identify the significance of the results in the analysis.
Data Visualizations **25%**	Data visualizations are appropriate for the level and type of analysis. . Uses graphs, figures, charts, and tables to increase visual effects of the main points being made based on the results.	Data visualizations are appropriate for the level and type of analysis. Graphs, figures and tables communicate insights and significance to the reader.	Data visualization are useful for the level and type of analysis, but graphs, figures and tables do not clearly communicate the significance of the results to the reader.	Data visualization are used minimally or not at all. If graphs, figures and tables are used, it is unclear what they are intended to communicate or why.
Interpretation & Conclusions **25%**	Wraps up the findings in a conclusion that provides an answer to the question(s) posed in the introduction. Makes specific recommendations based on the data presented.	The conclusion summarizes and makes sense of the results, making good points that reflect clear understanding of the assignment material.	The conclusion summarizes and makes sense of the results, making good points that reflect a basic understanding of the assignment material.	The conclusion does not summarize or attempt to make sense of the results. Conclusions do not reflect an understanding or reflect a misunderstanding of the material.

Report: Writing Mechanics, Title Page, & References **10%**	There are no noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	There are no noticeable errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	There are very few errors in grammar, spelling, and punctuation; and completely correct usage of title page, citations, and references. The report contains approximately 1,000 words.	There are more than five errors in grammar, spelling, and punctuation; or the usage of title page, citations, and references are incomplete; or the report contains far less than 1,000 words.