Northeastern University
Khoury College of
Computer Sciences

Practice 6

For all of the problems below, keep in mind that feature selection is one of the most difficult issues in model building, particularly regression. We have introduced several automatic feature selection approaches: forward and backward fitting using either *p*-value, Adjusted *R*-Squared, or *AIC*. In addition we also have Principal Component Analysis (*PCA*). However, in practice you may also need to choose from derived or combined features, *e.g.*, ratios or sums. This makes feature selection a combinatorially problem. In practice, you need to use domain expertise to choose features that you suspect or know contribute to the response variable. So, in this problem, use your domain knowledge and intuition.

R implements backward and forward fitting using *AIC* with the *step()* function. You may use it for the problems if you wish. Note that *AIC* will likely produce a model that includes coefficients with a *p*-value > 0.05. That is because *AIC*-based selection is based on adding or eliminating features that reduce the information in the model -- it is not based on statistical significance. Also note that elimination is a greedy algorithm and will not produce an optimal model. Like finding an optimal decision tree, finding an optimal set of features is an *NP*-Complete problem and thus is computationally intractable requiring suboptimal solutions that can be identified in polynomial time.

In addition, just because a model A has a higher Adjusted R-Squared or a lower *AIC* compared to model B doesn't mean that it is better. Model A may have a lower mean squared error (smaller mean residuals) but that difference in mean error could be due to sampling. Thus, data scientists confirm that model A is really better than model B by running a one-way *ANOVA* or a *t*-test that compares mean residuals. In R you can simply use *a <- anova(modelA, modelB)* followed by *summary(a)* to determine if the difference in the model performance is statistically significant.

Compare your answer with those of your peers using the discussion forum.

Problem 1 (60 Points)

Download the data set on student achievement in secondary education math education of two Portuguese schools (use the data set *Students Math*). Using any packages you wish, complete the following tasks:

1. (10 pts) Create scatter plots and pairwise correlations between *age*, *absences*, *G1*, and *G2* and final grade (G3) using the *pairs.panels()* function in R.
2. (10 pts) Build a multiple regression model predicting final math grade (G3) using as many features as you like but you must use at least four. Include at least one categorical variables and be sure to properly

variables and then state the final model as an equation. State the backward elimination measure you applied (*p*-value, AIC, Adjusted R2). This tutorial shows how to use various feature elimination techniques.

4. (10 pts) Calculate the 95% confidence interval for a prediction -- you may choose any data you wish for some new student.
5. (10 pts) What is the *RMSE* for this model -- use the entire data set for both training and validation. You may find the *residuals()* function useful. Alternatively, you can inspect the model object, *e.g.*, if your model is in the variable *m*, then the residuals (errors) are in *m$residuals* and your predicted values (fitted values) are in *m$fitted.values*.

Problem 2 (40 Points)

For this problem, the following short tutorial might be helpful in interpreting the logistic regression output.

1. (5 pts) Using the same data set as in Problem (1), add another column, PF -- pass-fail. Mark any student whose final grade is less than 10 as F, otherwise as P and then build a dummy code variable for that new column. Use the new dummy variable column as the response variable.
2. (10 pts) Build a binomial logistic regression model classifying a student as passing or failing. Eliminate any non-significant variable using an elimination approach of your choice. Use as many features as you like but you must use at least four -- choose the ones you believe are most useful.
3. (5 pts) State the regression equation.
4. (20 pts) What is the accuracy of your model? Use the entire data set for both training and validation.

Problem 3 (10 Points)

1. (8 pts) Implement the example from the textbook on pages 205 to 217 for the data set on white wines.
2. (2 pts) Calculate the RMSE for the model.

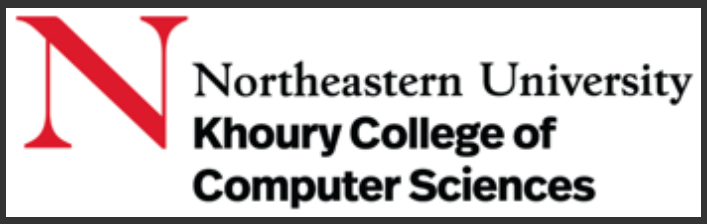
Submission Details

- Practice Problems are for learning and practice and therefore are not graded and no submission is required. You are encourage to discuss and review them with your peers. Additionally, they are reviewed during weekly recitations. If you desire, you may ask for individual feedback from the instructional staff during office hours. Completing practice problems will prepare you for the graded practicums and their completion is critical to doing well on the practicums and the final project..

Useful Resources

- R Markdown Notebooks

Northeastern University
Khoury College of
Computer Sciences



Learning

Blackboard
Lynda.com
Data Camp

Support

Contact Instructor
Virtual Office
Book Appointment

Search