Northeastern University
Khoury College of
Computer Sciences

Practicum 2

The practicum week is intended to deepen that material and foster discussion where you can learn from your peers and from problems encountered in solving the various questions. Therefore, interact on the discussion board, explore, investigate, share, and respond to each other. You can earn bonus points should you need to make up for shortcomings in Practicum 1 -- alternatively, you can choose which problems to solve.

This is an individual practicum which means that you may not work in groups.

Problem 1 (70 Points)

1. (0 pts) Download the data set Census Income Data for Adults along with its explanation. Note that the data file does not contain header names; you may wish to add those. The description of each column can be found in the data set explanation.
2. (0 pts) Explore the data set as you see fit and that allows you to get a sense of the data and get comfortable with it.
3. (5 pts) Split the data set 75/25 so you retain 25% for testing using random sampling.
4. (20 pts) Using the Naive Bayes Classification algorithm from the **KlaR**, **naivebayes**, and **e1071** packages, build an ensemble classifier that predicts whether an individual earns more than or less than US$50,000. Only use the features *age*, *education*, *workclass*, *sex*, *race*, and *native-country*. Ignore any other features in your model. You need to transform continuous variables into categorical variables by binning (use equal size bins from in to max). Note that some packages might not work with your current version of R and may need to be downgraded.
5. (10 pts) Create a full logistic regression model of the same features as in (4) (*i.e.*, do not eliminate any features regardless of *p*-value). Be sure to either use dummy coding for categorical features or convert them to factor variables and ensure that the *glm* function does the dummy coding.
6. (5 pts) Add the logistic regression model to the ensemble built in (4).
7. (15 pts) Using the ensemble model from (6), predict whether a 35-year-old white female adult who is a local government worker with a doctorate who immigrated from Portugal earns more or less than US$50,000.

8. (15 pts) Calculate accuracy and prepare confusion matrices for all three Bayes implementations (**KlaR**, **naivebayes**, **e1071**) and the logistic regression model. Compare the implementations and comment on differences. Be sure to use the same training data set for all three. The results should be the same but they may differ if the different implementations deal differently with LaPalace Estimators.

Problem 2 (50 Points)

1. (0 pts) Load and then explore the data set on car sales referenced by the article Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579.
2. (5 pts) Are there outliers in the data set? How do you identify outliers and how do you deal with them? Remove them but create a second data set with outliers removed. Keep the original data set.
3. (5 pts) What are the distributions of each of the features in the data set with outliers removed? Are they reasonably normal so you can apply a statistical learner such as regression? Can you normalize features through a log, inverse, or square-root transform? Transform as needed.
4. (5 pts) What are the correlations to the response variable (car sales price) and are there collinearities? Build a full correlation matrix.
5. (0 pts) Split the data set 75/25 so you retain 25% for testing using random sampling.
6. (10 pts) Build a full multiple regression model for predicting car sales prices in this data set using the complete training data set (no outliers removed), *i.e.*, a regression model that contains all features regardless of their p-values.
7. (10 pts) Build an ideal multiple regression model using backward elimination based on p-value for predicting car sales prices in this data set using the complete training data set with outliers removed (Question 2) and features transformed (Question 3). Provide a detailed analysis of the model using the training data set with outliers removed and features transformed, including Adjusted R-Squared, RMSE, and p-values of all coefficients.
8. (5 pts) On average, by how much do we expect a leather interior to change the resale value of a car based on the models built in (6) and in (7)? Note that 1 indicates the presence of leather in the car.
9. (5 pts) Using the regression models of (6) and (7) what are the predicted resale prices of a 2005 4-door Saab with 61,435 miles with a leather interior, a 4-cylinder 2.3 liter engine, cruise control, and a premium sound system? Why are the predictions different?
10. (5 pts) For the regression model of (7), calculate the 95% prediction interval for the car in (9).

Submission Details

- Total Points: 120/100 (100 + 20 bonus points). This means you can earn extra points to boost your final course grade or you may choose which questions to work on.
- Your submission must contain two files: the *.Rmd* notebook **and** a knitted PDF or HTML (from the notebook). Name your *.Rmd* R Notebook, DA5030.P2.*LastName*.{Rmd,[pdf|html]}, where *LastName* is

your last name.

- The *.Rmd* file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it.) Code that doesn't execute, stops, throws errors will receive -- naturally -- receive no points. If the graders have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.
- Not submitting a knitted PDF or HTML will result in reduction of 30 points.
- Not submitting the *.Rmd* file (or both) will result in a score of 0.
- We must be able to run your code.
- Submit by the due date posted on Blackboard. No late submissions are acceptable past the dates posted on Blackboard.
- The submission link is in the Assignments folder on Blackboard. No email submissions are accepted.
- **No late submissions will be accepted. Submit early and often in case of computer, network, or Blackboard glitches.**

Useful Resources

- Quick Guide to Creating Scatterplots in R with ggplot
- Cross Validation for Naive Bayes Classifiers

Search