DA5030 | Intro to Machine Learning & Data Mining

Northeastern University
Khoury College of
Computer Sciences

Practicum 3

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. Build separate R Notebooks for the two problems.

Problem 1

1. (0 pts) Download the data set Bank Marketing Data Set. 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. Use the *bank-additional-full.csv* data set. Select an appropriate subset for testing. Use *bank-additional.csv* if your computer cannot process the full data set.
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. Is there distributional skew in any of the features? Is there a need to apply a transform?
3. (15 pts) Build a classification model using a support vector machine that predicts if a bank customer will open a term deposit account.
4. (15 pts) Build another classification model using a neural network that also predicts if a bank customer will open a term deposit account.
5. (20 pts) Compare the accuracy of the two models based on AUC.
6. (10 pts) Calculate precision and recall for both models. See this article to understand how to calculate these metrics.

Problem 2

1. (0 pts) Download the data set Plant Disease Data Set. 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. This assignment must be completed within a separate R Markdown Notebook. Use *read.transaction()* from the **arules** package to read the data.
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. Is there distributional skew in any of the features? Is there a need to apply a transform?

3. (40 pts) Use association rules to segment the data similar to what was done in Hämäläinen, W., & Nykänen, M. (2008, December). Efficient discovery of statistically significant association rules. In Data Mining, 2008. *ICDM'08. Eighth IEEE International Conference* on (pp. 203-212). IEEE.
4. (+30 pts) Are there clusters in the data? Can plants be segmented into groups? Build a *k*-means clustering model to investigate.
5. (+10 pts) Visualize the clusters.

Submission Details

- Submit the both the *.Rmd* R Notebook files and the knitted HTML output files containing your embedded code and explanations -- separate notebooks and HTML files for each problem. As always, zip all files together before submitting; you cannot upload HTML files to Blackboard. You will lose 20 points if your submission is not an R Notebook.

Useful Resources

- TBD

Northeastern University
Khoury College of
Computer Sciences

Learning

Blackboard
Lynda.com
Data Camp

Support

Contact Instructor
Virtual Office
Book Appointment

Search

Created by Martin Schedlbauer, PhD