BT5152 AY18/19 Semester 1

Assignment 5

Due: 18 November 2018, 11:59 PM

Question 1 (4 Marks):

Let's use the A3 training dataset for this question. You can keep only two columns, X4 and X8, and the label is still y. Use the first 1500 rows for training and last 500 rows for testing. We will compare the performance of 3 different SVM kernels by "svm" function of "e1071" library. Based on the sample code provided, please Train, Tune, and Visualize SVM using the following 3 kernels: (1) Polynomial Kernel, (2) RBF Kernel, (3) Sigmoid Kernel. The initial data preparation part has been done for you in the code sample (A5-1_template.R)

Using the best parameters to make predictions on the testing set. Briefly report that how is the best prediction performance across 3 kernels. How does the best model compare to other algorithms that you have tried in A3?

Grading of this question is mostly based on the correctness of your code. However, if your result is far from the median performance of this class of 3 methods, penalty may be imposed.

Question 2 (5 Marks): Classification for Unbalanced Dataset

The dataset is the HomeCredit dataset from Kaggle.com. Please use the pre-processed dataset uploaded on IVLE for this assignment and a code sample (A5-2_template.R) is provided.
The files are named:

- **application_v3.csv**: This is your main data file for this question
- **column_descriptions.xlsx**: This is an Excel spreadsheet that describes the columns

(2-1) Please try using XGBoost as the baseline model to predict "TARGET", which is the customer will default or not. The performance metric is AUC.
(2-2) Please try under-sampling, over-sampling, under- and over-sampling together, and SMOTE with the XGBoost. Report which method gives you the best prediction performance. Similarly, if your best performance is too poor, penalty may be imposed on this question.

Question 3 (3 Marks): Please provide your feedback to BT5152 for future improvement.

Visit this link: https://goo.gl/forms/B71TmWdjfkSMMqx13

(3-1): Which week's lecture is most interesting and/or helpful to you? (Dr. Huang will consider expand it in the future.)
(3-2): Which week's lecture is LEAST interesting and/or helpful to you? (Dr. Huang will consider drop it in the future.)

(3-3): Any suggestions or feedbacks for future improvement?

Submissions and Grading

- You need to submit 4 files in a zip file (A0123456X.zip):
    - A5-1.R (or .Rmd) file
    - A5-2.R (or .Rmd) file
    - A5-1.PDF (or .html generated from .Rmd) file
    - A5-2.PDF (or .html generated from .Rmd) file

  You may use any packages, but make sure your R file is runnable and has all the dependency packages imported e.g. library(C50).
- The page limit of the pdf file is maximum 2 pages including everything. The formatting is A4, default margin, 12 font size, single-spacing. There is no need to try to fill 2 pages. Correct answers are much more important than the length of your answers for grading.
- You may revise and submit as many times before deadline. Make sure to remove any old version that you don't wish to be graded.
- If you have questions about the assignment, feel free to email TA and cc me. Later, if you have questions about grading of assignment, then you can email TA and cc me because TA (not me) will grade your assignment by following my grading rules listed below.

Grading Policy

- Zero tolerance for Plagiarism.
- Every day of late submission will result in 3 marks deducted, i.e. 4 days late = 0 mark.
- Submissions without a runnable R/Rmd file will receive a failing grade. Make sure all the dependency packages are imported e.g. library(C50)
- TA can judge the quality of your code and deduct up to 2 marks. For example, if you included quite a number of unnecessary codes (which shows you do not really know which line of R command is the real one that helps you conduct analysis). You can provide comments into your codes to show your understanding.
- TA can give you up to +2 bonus if you did an excellent job. The max of A1 is still 12 marks.