CMPT459 Fall 2017
Data Mining
Martin Ester
TA: Zhilin Zhang

Programming Assignment 2

Total marks: 100
Due date: October 18, 2017

Data
Our dataset is a collection of data about the Titanic passengers, and the goal of the assignment is to predict the survival of a passenger based on some features such as the class of service, the sex, the age etc. The dataset can be downloaded as

titanic3.csv

from http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets under Data for Titanic passengers.

Here is a description of the attributes:

```
pclass          Passenger Class          (1 = 1st; 2 = 2nd; 3 = 3rd)
survived        Survival                 (0 = No; 1 = Yes)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                 (C = Cherbourg; Q = Queenstown; S = Southampton)
boat            Lifeboat (ID of the lifeboat the passenger took)
body            Body Identification Number (ID of the corpse)
home.dest       Home/Destination
```

Tasks
In this assignment, you will gain practical experience with data preprocessing and classification methods. Solve the tasks using R and answer the questions.

1. Read in the dataset and split the dataset randomly into 80% training data and 20% test data using the function sample(). To make sure that everybody uses the same training/test split, set the seed of sample to 1 using command set.seed(1).

2. Report the number of missing values per attribute in the training and test dataset.

3. You can use only past data to predict the future. Assume that you want to predict the survival of a passenger at the time of the accident, i.e. when the Titanic hit the iceberg. With this assumption in mind, which attributes do you use as features?

4. How do you deal with missing values in the different attributes? Report your plan. Preprocess the dataset according to your plan.

5. Learn a logistic regression model from the training data. What are the most significant three attributes of your model?

6. Apply the logistic regression model to predict the class labels of the test data. Plot the confusion matrix. What is the accuracy of the model?

7. Plot the ROC curve of your logistic regression model for varying probability thresholds. What is the AUC of your logistic regression model?

8. Learn SVM models from the training data, using linear and radial kernels. Using the function tune(), obtain the best parameters for linear and for radial kernels. What are the best parameters fort he linear and for the radial kernel? Discuss the results.

9. Apply the best SVM model to predict the class labels of the test data. Plot the confusion matrix. What is the accuracy of the model?

10. Plot the ROC curve of your best SVM model for varying probability thresholds. What is the AUC of the model?

Submission

Submit your R code in pa2.r and a report report2.pdf answering the questions in CourSys.

Note: we do not accept handwritten submissions, but only typed reports!