CMPT459 Fall 2018
Data Mining
Martin Ester
TA: Sahand Khakabimamaghani

Programming Assignment 2

Total marks: 100 (50 for the code, 50 for the report)
Due date: October 24, 2018

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 from

https://coursys.sfu.ca/2018fa-cmpt-459-d1/pages/titanic.

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. Using package tree, learn a decision tree from the training data. Plot the resulting tree. What is the size of the tree? What is the accuracy of the tree on the test dataset?

6. Analyze the importance of attributes in your decision tree. Report the top three most important attributes in decreasing order of importance and explain your choice. What knowledge about the survival of passengers can you learn from the decision tree?

7. Prune your decision tree learnt in task 5. To do so, use cost complexity pruning and perform cross-validation in order to determine the optimal level of tree complexity. What is the size of the pruned decision tree?

Submission

Use R Markdown for doing this assignment. R Markdown allows you to have both report text and R code in one place while providing simple text formatting facilities. Please visit https://rmarkdown.rstudio.com/ for more information and tutorials.

To help you start, there is a template below that you can copy and paste and fill in the areas indicated as <bold text>. Submit your file pa2_<your student number>.Rmd in CourSys.

Template for the R Markdown code

~~~
---
title: Programming Assignment 2
author: <your name - your student number>
output:
  html_document:
    mathjax: default
---
```{r include=FALSE}
<libraries are loaded here>
```


## Task 1

```{r}
<your code for task 1 comes here>
```
<report text for task 1 comes here>

## Task 2

```{r}
<your code for task 2 comes here>
```
<report text for task 2 comes here>

<  and so on …>
~~~