CS909: 2017-2018

Exercise one: Exploring datasets, Data pre-processing, Decision Trees, Naïve Bayes

Submission: 12 pm Thursday February 15th 2018

Notes:

a) **This exercise will contribute towards 25% of your overall mark.**
b) **The use of R libraries is not allowed as part of this exercise, unless explicitly permitted.**

Preparation: Getting to know R

Install R from http://www.r-project.org/

Read the first ten chapters of Introduction to R and familiarise yourself with the contents of the following chapters:

Ch 1. how to use source(), rm(), q(), help(); R data format, where to find packages for R
Ch 2. how to create and populate vectors, arrays and do simple manipulations
Ch 3. numeric(), as.integer(), as.character()
Ch 4. factors
Ch 5. functions and manipulations of matrices
Ch 6. how to use lists and data frames
Ch 7. how to read data from files, especially using read.table().
Ch 8. how to use plot() with functions for descriptive statistics such as hist(), lines(), boxplot()
Ch 9. loops and conditions in R
Ch 10. how to implement a function in R

Please refer to online resources for relevant R documentation.

Part A: Exploring the iris dataset (40 marks)

1. Select and apply a suitable R command to discover the *number* and *names* of attributes in the iris dataset and the *number* of instances. (5)

2. What are the minimum, maximum, mean, median, and the first (25%) and third (75%) quartiles of the iris dataset attributes? (5)

3. Create a new object irisSubset containing rows 40 to 85 and save it in a file called irisSubset in the Rdata format. (5)

4. Remove object irisSubset from the R workspace. Load it back in from irisSubset.Rdata. (5)

5. Arrange the instances of irisSubset in descending order of attribute "Sepal.Length". (5)

6. Create a new subset irisSubsetSepal from iris with Sepal.Length < 5.4. (5)

7. Write a function that takes as its arguments an iris Species type and an attribute name and returns the minimum and maximum values of the attribute for that Species type. (10)

Part B: Data pre-processing (30)

1. Import the dataset irisMissing.csv into a data frame named irisMissing in your R workspace and use an R command to discover the row numbers of the instances that have missing values. (5)

2. Identify an R command that will drop missing values. Apply it to the irisMissing dataset to create a new data frame "irisDrop". **Briefly** describe three other strategies for handling missing values. Write your own R functions to implement each of these three strategies. (this should not take very long) (5)

3. Write an R function foo() that takes a data frame and a missing value function as arguments and returns a new data frame with the missing values replaced with values as determined by the missing value function. (10)

4. Use the hist() and boxplot() commands to compare results of applying each missing value strategy. Based on this, comment on their relative metrics. Save figures you generate as PDFs and include them in your Tabula submission. Make sure they are clearly labeled. (10)

Part C: Decision Trees (30)

1. Write a function disc() to discretise a dataset using equal width binning. It should take a data frame "dataset" and the number of bins as arguments and return "dataset" with non-ordinal attributes categorized. Load the "Loan" dataset into R and use your function to discretise it. (10)

2. Manually generate the decision tree for the Titanic passenger survival dataset below. Use Information Gain as your split measure. Sketch out the resulting decision tree and write out the equivalent rule set. (20)

Titanic passenger survival

Class	Sex	Age	Survived	Freq
1st	Male	Child	No	0

2nd	Male	Child	No	0
3rd	Male	Child	No	35
1st	Female	Child	No	0
2nd	Female	Child	No	0
3rd	Female	Child	No	17
1st	Male	Adult	No	118
2nd	Male	Adult	No	154
3rd	Male	Adult	No	387
1st	Female	Adult	No	4
2nd	Female	Adult	No	13
3rd	Female	Adult	No	89
1st	Male	Child	Yes	5
2nd	Male	Child	Yes	11
3rd	Male	Child	Yes	13
1st	Female	Child	Yes	1
2nd	Female	Child	Yes	13
3rd	Female	Child	Yes	14
1st	Male	Adult	Yes	57
2nd	Male	Adult	Yes	14
3rd	Male	Adult	Yes	75
1st	Female	Adult	Yes	140
2nd	Female	Adult	Yes	80
3rd	Female	Adult	Yes	76

<u>Part D: Naïve Bayes (40)</u>

1. View the HairEyeColor dataset. Assuming that the predicted class here is "Sex", express the prior probabilities and likelihood function of a Naive Bayes classifier model that can classify this data? (no coding required to answer this) (5)

2. Write a function NB that, given a data frame with discrete values and a class, returns the above parameters (priors and likelihood function). Verify your answers for the HairEyeColor dataset by using the Naive Bayes classifier in R (we recommend package e1071). (20)

3. What would happen if one of the features were zero, given your Naive Bayes function? e.g. if there were no red haired men in the dataset? How could you remedy this? (no coding required to answer this) (5)

4. Recall that in Part B the iris dataset had missing values. This time, rather than filling in the missing values using a mean or median, use Naive Bayes to help you with the task of finding missing values (we recommend using the R package klaR and the function NaiveBayes.) What assumptions have you made here and how do they influence the final classification into species? (10)