CS 422-04: Data Mining (CRN: 19881)

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Fall 2017: Homework 2 (10 points)

Due date: Wednesday, Oct 04, 2017 11:59:59 PM Chicago Time

1 Exercises (3 points divided evenly among the questions)

1.1 Chapter 4

Exercise 2, 3, 5.

2 Practicum problems

2.1 Problem 1 (6 points)

The ILPD.csv file contains a dataset of Indian liver patients (see https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29). The dataset contains 583 instances and 10 attributes; the 583 instances are composed of 416 liver patient records (class label = 1) and 167 non-liver patient records (class label = 2). Note that the classes are imbalanced here, there are more liver patients in the database than non-liver patients.

The attribute information is as follows:

age	Age of the patient
sex	Gender of the patient
tb	Total Bilirubin
db	Direct Bilirubin
aap	Alkphos Alkaline Phosphotase
sgpaa	Sgpt Alamine Aminotransferase
sgoaa	Sgot Aspartate Aminotransferase
tp	Total Proteins
alb	Albumin
ag	Ratio of Albumin to Globulin
label	Class Label (1 = Liver Patient, 2 = Non-liver Patient)

IMPORTANT: Set the seed to 100 (i.e., set.seed(100)) first. Then, divide the dataset into two parts: train and test with a random 60/40 split (i.e., 60% of the data is used for training and 40% for testing). Perform the following analysis on this dataset.

(a) For the training dataset, produce a correlation scatterplot of the variables. Use the psych package to do so. For example, here are the R command to get the correlation plot of the iris data using this package:

```
> install.packages("psych")  # Do only once
> library(psych)
> data(iris)
> pairs.panels(iris[1:4])
```

In the resulting graph, the diagonal is the attribute value, the upper triangle contains the numeric correlation coefficient and the lower triangle contains the correlation graphs.

Study the correlation graphs of the ILPD dataset and determine: (i) Which pair of attributes have the strongest correlation? (ii) Which pair has the weakest correlation? (iii) Which pair is the most negatively correlated? (iv) Which variables appear to follow a Gaussian distribution?

(b) Do you think that normalizing or scaling the attributes here will help with the classification task? Please justify your answer. If you think it will help, which attributes would you normalize or scale?

(c) Run rpart decision tree algorithm on the training data to create a model using all predictor variables. Then, using that model evaluate the out-of-sample accuracy for the model using the predict() function. What is the accuracy of the model on out-of-sample data? (The accuracy is evaluated by the confusionMatrix() function as shown in class.) What is the TPR? What is the TNR? What is the PPV?

(d) Investigate the model created in (c) using the plotcp() and printcp() functions to determine where to prune the tree. Prune the tree as needed and re-evaluate the out-of-sample accuracy for the model using the newly pruned tree. Is the accuracy of the pruned tree on out-of-sample data better, worse or remains unchanged compared to the one you obtained in (c)? Why do you think the newly pruned tree provides a better, worse, or unchanged accuracy?

(e) Create a new model on the training data, but this time, instead of using all predictor variables, do some exploratory analysis to determine if you can reduce the number of predictor variables (perhaps those pairs that have high correlation with each other can be reduced to one predictor; or if a number of predictors have similar mean and standard deviation, you can keep only one of them; or think of other creative ways to reduce predictor variables). Create a new model on the in-sample (training) data using the reduced set of predictors and evaluate the new model on the out-of-sample data. What is the accuracy of the model on out-of-sample data? (The accuracy is evaluated by the confusionMatrix() function as shown in class.) What is the TPR? What is the TNR? What is the PPV?

(f) For each of the model created in (c) and (e), provide the following:

(i) a ROC curve using the ROCR package.
(ii) AUC.
(iii) Which model performs better and why?

2.2 Problem 2 (1 points)

The original ILPD dataset contained some missing values, which I have filled in the dataset you are being provided for this problem. Download the original dataset from https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29 and determine which attribute had the missing values, and how many instances were missing. Then, compare the corresponding instances in the dataset you are provided for Problem 1. Use R commands to figure out which instances for which attributes were missing values, and how these missing values were imputed in the dataset you analyzed for Problem 2.1. Prepare a brief report (or include a knitted document) that includes the R commands you used to figure out the missing values and how they were imputed.

Hints: See Lander book, page 50. Also see http://www.statmethods.net/input/missingdata.html