CSI 5387: Data Mining and Concept Learning

Winter 2020

Assignment 1

Submission Deadline: 12th February 2020 on Brightspace.

Overview:

This assignment covers some basic data mining tasks: exploratory data analysis, data pre-processing, and a simple classification model construction. The assignment is in two parts; the first part focuses on exploratory data analysis techniques and data preprocessing while the second part involves the construction of a Decision Tree.

Part A (60 points)

1) Compute the dissimilarity or similarity measures for the following vectors.
   a. X = {1, 1, 1, 1}, Y = {2, 2, 2, 2} Manhattan, Euclidean, Supreme
   b. X = {0101010001}, Y = {0100011000} cosine, Jaccard coefficient
2) Some women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, were tested for diabetes according to World Health Organization criteria. Given a sample of the data for their *age* and *BMI* in Table 1 below, answer the following:
   a. Calculate the mean, median, variance and standard deviation of *age* and *BMI.*
   b. Find the Q1 and Q3, IQR and Range.
   c. Draw the boxplots for *age* and *BMI*
   d. Draw a scatter plot and a q-q plot based on these two variables.
3) Using Table 1, answer the following:
   a. Normalize the two attributes based on *z-score* normalization
   b. Calculate the *correlation coefficient* (Pearson's product moment coefficient).
   c. Are these two attributes positively or negatively correlated?
   d. Compute their covariance.

Table 1: Age and BMI of Pima Indian Women

Age	24	55	35	26	23	52	25	24	63	31
BMI	30.2	25.1	35.8	47.9	26.4	35.6	34.3	25.9	32.4	43.3
Age	33	45	59	44	29	21	21	51	42	21
BMI	43.1	30.9	30.1	27.6	41.3	23.2	25.4	36.6	29.3	22.1

Part B (40 points)

For this part of the assignment, you will experiment with either Weka, R or Python. If using Weka, note that there are filters you can use to complete the data preprocessing component. Like the classifiers in Weka, you can select filters from the menu and tailor them to your requirements. The performance of the classifier should be evaluated by 10-fold cross validation on the provided dataset.

A benchmark dataset to predict the type of thyroid disease a patient has is being provided. Two versions of the dataset is attached in CSV and Arff format. Use the applicable dataset file type suitable for your choice of platform to complete the following:

a. Perform feature selection on the dataset and state briefly why feature selection is sometimes important.

b. Create a C4.5 or C5.0 model using the selected attributes from your dataset to predict the type of thyroid disease a patient has. Visualize the decision tree. Describe the first few splits in the decision tree. Can you extract some rules?
c. Try different ways to improve your decision tree algorithm. Some examples of strategies include using different splitting strategies, and pruning tree after splitting.
d. There are some missing values in the dataset. Several strategies can be used to handle them, e.g., remove cases with unknowns, fill in unknown values by exploring correlation, and by exploring similarity between cases. Apply one of these methods to address the missing values.

Deliverables:

- Solution to the questions from parts A & B.
- Screen shots of the developed decision tree.
- A set of rules extracted from the developed decision tree.
- Brief description of how you implemented the tree and why you chose your approaches.
- Accuracy result of your implementation (with the confusion matrix).
- Explain improvements that you made and why you think it works better (or worse).

Weka Resources:

- Weka can be downloaded from https://www.cs.waikato.ac.nz/ml/weka/.
- The online appendix for the 4th edition of the book Data Mining: Practical Machine Learning Tools and Techniques can be obtained from https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf to get more details on how to use Weka.
- Another great resource are videos by Rushdi Shams on YouTube https://www.youtube.com/watch?v=11naa4Brp9s&list=PLJbE6j2EG1pZnBhOg3_Rb63WLCprtyJag&index=35
- You can find more resources online…