Data Mining Assignment 1

Your name and UIN here

February 3, 2017

I declare that all material in this assessment task is my own work except where there is clear acknowledgement or reference to the work of others and I have complied and agreed to the UIS Academic Integrity Policy at the University website URL: http://www.uis.edu/academicintegrity

Student Name: Your name goes here UID: Your UID here Date: February 3, 2017

Preparation Instructions

- You may use this TeX file as your base document remove the Preparation Instructions and the questions text and insert your answers in it, OR use any other editor to produce a document matching the published Typsetting Standards. Please remember:

    1. Leaving the Preparation Instructions and the questions text will contribute to a poor Turnitin similarity score and a breach of the report length limitation (I suggest removing them)

    2. If your submission similarity score exceeds 15% I do not grade the attempt and award you a summary 0 grade for submitting material that is not uniquely yours

    3. Reports must be authored using correct academic structure, that is Abstract, Introduction, Methodology, Results discussion, Conclusions (if any), references

4. Citations must be in the Harvard style (author and date), with correct bibliographical entries in the reference section

5. Expected report length 1000 to 1200 words

- This assessment task is due on Sunday night February 12, by 23:00 please note late submissions incur a 25% mark deduction penalty per 24 hours or part thereof.

---

1 Using the WEKA Workbench

Task 1: 4 Points

Ojective: Become familiar with the use of the WEKA workbench to invoke several different machine learning schemes.

- Use the following learning schemes to analyze the zoo data (in Zoo.arff):

Decision stump:	weka.classifiers.tree.DecisionStump
OneR:	weka.classifiers.rules.OneR
Decision table:	weka.classifiers.rules.DecisionTable -R (you will need to explore the parameter options to figure out what -R means and make a comment on this in your report)
C4.5:	weka.classifiers.trees.j48
PART:	weka.classifiers.rules.PART

The data you derive from the experiments described above are for your own use in order to prepare the solution to the question below (do not include "WEKA output" verbatim in your report):

- How do the classifiers determine whether an animal is a mammal, bird, reptile, fish, amphibian, insect, or invertebrate? Do the decisions made by the classifiers make sense to you?

- What can you say about the accuracy of these classifiers when classifying an animal that has not been used for training? Why does OneR perform so badly?

Task 2: 4 Points

Use the following learning schemes to analyze the bolts data (bolts.arff without the TIME attribute):

- *The bolts dataset describes the time needed by a machine to produce and count 20 bolts. (More details can be found in the file containing the dataset.) Analyze the data to answer the questions below.*

Decision stump:	weka.classifiers.tree.DecisionStump
OneR:	weka.classifiers.rules.OneR
Decision table:	weka.classifiers.rules.DecisionTable -R (you will need to explore the parameter options to figure out what -R means and make a comment on this in your report)
Linear regression:	weka.classifiers.functions.LinearRegression
M5P	weka.classifiers.M5P

The data you derive from the experiments described above are for your own use in order to prepare the solution to the question below (do not include verbatim in your report):

- What adjustments have the greatest effect on the time to count 20 bolts?
- According to each classifier, how would you adjust the machine to get the shortest time to count 20 bolts?

Write a report that records how your investigations proceeded and what results you found. Do not describe how to use the workbench or how the schemes in it work.

Experimental Research Presentation: 3 Points

1. Abstract:

   A summary of the contents of your report.

2. Introduction:

   State the purpose of the experiment

   Review the existing information or the theory

   A paragraph on theoretical model algorithms etc...

3. Methodology:

   Indicate experimental setup parameters, data collection techniques, number of iterations (per algorithm)

4. Results:

   Provide tables showing outcomes (do not copy WEKA output wholesale you will loose points for doing so) only cite data relevant to your findings

   Describe the uncertainties and offer reasons

   Provide graphs to support your arguments in "5"

5. Discussions/Conclusions:

   These should be supported by your results from above

Task 3: 4 Points

Objective: Explore your understanding of the theoretical aspects of Data Mining. Please answer the following questions, make sure you include a coherent rationale for each answer.

1. Your task is to offer good advice to marketing department. They evaluate customer satisfaction by keeping a record of the number of customer complaints for each product on the assumption that counts are ratio attributes, and thus product satisfaction must be a ratio attribute. The company auditors disagree and suggest that marketing has overlooked the obvious, rendering the satisfaction measure unreliable to misleading. By marketing department measure the company's best-selling line has the worst satisfaction as it has the greatest number of complaints. Identify the problem with the measure and propose a solution.

2. Describe how data mining can help a researcher design a digital document search tool by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied to a vast digital document collection.

Last updated: February 3, 2017

Typeset using TeXshop