CIS 635 Data Mining

Homework 2

Objectives

- reinforce concepts related to preprocessing
- explore the R programming language

Instructions

You are given a data file: hw02data1.txt. This is a table of ratings for teaching assistants (TA) at some university (this is modified data that originally came from https://archive.ics.uci.edu/ml/datasets.php) There are 7 attributes:

a) id: unique number to identify each rating
b) type: type 1=TA native speaker, 2=non-native speaker (not important)
c) instructor: id number that identifies the instructor
d) classNbr: id number that identifies the course
e) semester: 1=summer, 2=fall, 3=winter
f) classSize: number of students in the class
g) rating: 1=Low, 2=Medium, 3=High

What type of data are these (nominal, ordinal or numeric)?

Part A – concepts

1. read the data from hw02data1.txt into R
2. delete the rows with missing data
3. identify any noise (3 occurrences) – change these to the mean (mode in this case)
4. identify any outliers (1 occurrence)
5. thinking about similarity
   a) is a rating of 1 more similar to 2 or to 3 or are they both the same?
   b) is a semester of 1 more similar to 2 or to 3 or are they both the same?

Part B document tables

Open the file hw02data2.txt in a text editor. It contains 5 documents. Each document is an email. Use these emails to fill in the docTerm matrix on the handin sheet. Fill each cell with a 0 or 1, 0 if the term(word) is not used in the document and 1 if it is.

Part C transactions

Open the file hw02data3.txt in a text editor. It contains 6 transactions. Each transaction is a list of items that someone purchased from a store. A pattern is when you spot an item that appears when a different item appears, like this: whenever you see A, you often see B. Give an example of one pattern you see in this list of transactions.

Part D – sampling

For this part you will use the data from hw02data1.txt in R.

1. calculate the mean value of the classSize.
2. write the code to sample 10 records without replacement using the R command sample.
3. write a function that takes a parameter for the number of samples and the mean (you can use the value from #1). In the function use a for loop that iterates 10 times. Each time it will sample the number of records using the parameter. Each time you sample, calculate the mean value of classSize. Calculate the absolute value of the difference between this mean from the parameter mean. After the loop, return the average of the differences calculated in the loop.
4. Call this function with 25, 50 and 100 samples and report the results