Homework 2

Mining Association Rules from Gene Expression Data

Code Submission Due: October 15 2016 1:59PM
Hard Copy Report: before class starts on October 15

The gene expression data can be found on Piazza. The first column denotes the sample id, and the last column is disease name. For the rest columns, each column represents one gene. For example, the row " Sample4 Down Down Down Up ... AML " can be interpreted as: “Sample4 G1_Down G2_Down G3_Down G4_Up ... AML”.

Requirements

1. Implement the Apriori algorithm to find all frequent itemsets. Report the number of frequent itemsets for support of 30%, 40%, 50%, 60%, and 70%, respectively.

   You cannot directly call a function or package that implements Apriori. You need to implement the algorithm by yourself. If you are not sure about whether it is OK to use a certain function, please post your question on Piazza.

   In your report, please list the results for different support values like this:
   Support is set to be 50%
   number of length-1 frequent itemset: ??
   number of length-2 frequent itemset: ??
   number of length-3 frequent itemset: ??
   Total: ??

2. Generate association rules based on the templates. Test templates:
   Template 1:
   {RULE|BODY|HEAD} HAS ({ANY|NUMBER|NONE}) OF (ITEM1, ITEM2, ..., ITEMn)
   Template 2: SizeOf({BODY|HEAD|RULE}) ≥ NUMBER.
   Template 3: Any combined templates using AND or OR. For example:
   HEAD HAS (1) OF (Disease) AND BODY HAS (NONE) OF (Disease)

   In your report, list all the generated rules for a given support and confidence value.
   We will release the specific sample queries two days before the code submission deadline on Piazza.

3. Prepare your submission. Your final submission should be a zip file named as Homework2.zip. In the zip file, you should include:

- A folder “Code”, which contains all the codes used in this assignment. Inside the folder, please have a file “README” which describes how to run your code

- Report: A pdf file named as Homework2.pdf. The report should at least consist of the following parts: 1) describe your implementation details about Apriori Algorithm; 2) results for above requirement 1; 3) results for above requirement 2.

4. Log in any CSE department server and submit your zip file as follows:
submit_cse601 Homework2.zip

Please refer to Course Syllabus for late submission policy. We will take the submission time recorded by the server as the time of your submission.

This assignment must be done independently. Running your submitted code should be able to reproduce the results in the report. Note that copying code/report from another group or source is not allowed and may result in an F in the grades of all the team members. Academic integrity policy can be found at http://www.cse.buffalo.edu/shared/policies/academic.php