CS 6604 Assignment #1
Assigned: Sep 1, 2007
Date Due: Sep 12, 2007, in class, before class starts

---

1. (5 points) In class, we have formally characterized the lattice induced by subsets and defined (two) closure operators over them. Similarly, formally define the lattice induced by all boolean expressions and identify closure operators over this lattice (*Hint:* think of subsets as conjunctions and extend the formalism to general boolean expressions). Cookup a small dataset and show the lattice of closed sets induced by your operators. For some closed sets, show their minimal generators.

2. (5 points) The database literature references a concept called 'functional dependencies' (FDs) which sound superficially similar to association rules. But there are important differences. Whereas association rules are posed over data instances (items and itemsets), FDs are posed over the database schema. Whereas association rules are mined with user-specified confidence levels, FDs are typically intended to have 100% confidence. Pickup a database book, learn about the concept of FDs, and suggest how we can mine FDs in a levelwise manner. For full credit, explain how the space of patterns is organized levelwise, such as along a lattice, what the 'edges' of the lattice mean, and what properties can be exploited to prune the search for FDs. Define terms similar to support and confidence as criteria for mining functional dependencies.

3. (10 points) The Brin, Motwani, and Silverstein paper investigates the use of the χ^2 measure for determining correlated sets of items. Reconsider this problem but with the Pearson's correlation as the measure of interest. Explain whether this measure obeys upward-closure or downward-closure and how you can design an algorithm to find correlated sets.

4. (10 points) You are given a market basket database of n transactions, each of which has exactly m items, chosen from a total of i items. If the support threshold is s, give the tightest bounds possible on the minimum and maximum numbers of frequent itemsets.

5. (20 points) The authors of the paper 'ABS: Adaptive Borders Search of Frequent Itemsets' claim that there is a bug in the *Pincer Search* paper. Read both papers and either confirm that the bug exists or why their argument is incorrect.

6. (50 points) Many algorithm implementations for frequent itemset mining are available in the FIMI repository (http://fimi.cs.helsinki.fi/). In this exercise you will pick an algorithm from this repository and apply it on the Netflix recommendation dataset (http://www.netflixprize.com) to find frequently rated (not purchased) itemsets. Observe that you must first render the dataset in the binary item-transaction format necessary for association mining. Write a detailed report on how you chose the algorithm, how the parameters were selected, how the dataset was structured, timing graphs after mining, and key insights gained from the data.

Turnin a typed (not handwritten) paper copy giving answers to the questions, plots, including a brief description of how you solved each question. Write enough to convince us that you completed the assignment independently. The paper copy should mention a web URL where we can go and check out your codes and download them, if necessary. Assignments without the web URL will not be graded.