581550 Data mining — tietämyksen muodostaminen
Autumn 2002
Hannu Toivonen

Exercises 1 (due Sep 16–20)

1. Data set

   http://www.cs.helsinki.fi/hannu.toivonen/teaching/timuS02/datasets/ilmo

   contains data about the courses to which some students have registered. What types of knowledge could one find about this data set?

2. Have a look at the data set

   http://www.cs.helsinki.fi/hannu.toivonen/teaching/timuS02/datasets/ice-cream.txt

   Find software tools that make it possible for you to draw the data set and analyze it somehow. (Suggestions: SAS, SPSS, Matlab, Awk & Gnuplot, etc.) What can you find about the relationships of temperature and ice cream consumption?

   (If you have not used any statistical software package before, all these analyzes can be done using for example awk or perl for data manipulation and gnuplot for drawing. For example, in gnuplot the command

   ```
   plot 'datafile', 3*x-1
   ```

   will plot the points in file 'datafile' (each point on a row, two first fields are used as x and y coordinates) and the line $3x + 1$. The awk program (from the unix shell command prompt)

   ```
   awk '{print $1, $2-(3*$1-1)}' < datafile > residuals
   ```

   computes the residuals of fitting the line to the data. This file can again be plotted. The residuals can also be plotted directly in gnuplot by

   ```
   plot "< awk '{print $1, $2-(3*$1+1)}' < plotexample"
   ```

   For more information about gnuplot, see for instance

   http://www.duke.edu/~hpgavin/gnuplot.html

3. Study the proof of the rule generation algorithm (Theorem 2.10, page 14) and prepare to explain why it works correctly.

4. Find frequent sets in the following 0/1 relation over $\{A, \ldots, H\}$ using Apriori candidate generation with frequency threshold $min\_fr = 0.3$.

Row ID	Row
t_1	$\{A, C, E, G\}$
t_2	$\{A, B, E, F, H\}$
t_3	$\{B, E, H\}$
t_4	$\{B, H\}$
t_5	$\{C, E, F, G, H\}$

- What are the collections of potential candidates, candidates $\mathcal{C}_l$, and frequent sets $\mathcal{F}_l(r)$ for size $l = 1, 2, ...$?
- How many sets does Apriori need to test against the database? How large a fraction is this from the whole search space of all item sets?
- What association rules hold in this relation when the confidence threshold is $min\_conf = 0.7$?
- Use the program

  /home/fs/htoivone/timuS02/programs/covsets1.v09

  (Sun binaries, runs e.g. on hydra) on the above input. Suitable options are

  covsets1.v09 -s0.3 -c0.7 -e1 filename

  (Some help is available by `covsets1.v09 -h`. Each input line represents a database row, with white-space separated strings as the items on the row. The order of items on a row does not seem to be significant. Terminology in the program:
  support = frequency, tuple = row, covering set = frequent set.)

5. Assume the largest frequent set is of size k.

   - How many database passes does Apriori need?
   - How many frequent sets are there?
   - Test the above program on an input line containing just one line of the form

     A B C D E F G H

     What do you notice?
   - Assume k is "large", say $k > 15$. Discuss the effect of database size (very large vs. very small) on the relative costs of database pass and Apriori candidate generation.

6. Use the above mentioned program on some subset of the course enrollment data. Did you find any interesting rules?