581550 Data mining — tietämyksen muodostaminen
Autumn 2002
Hannu Toivonen

Exercises 2 (due Sep 23–27)

1.-2. Recall the definitions of conditional probability and statistical independence. The *conditional probability* of X given that Y is true is $P(X|Y) = \frac{P(X \text{ and } Y)}{P(Y)}$. X and Y are *(statistically) independent* if and only if $P(X \text{ and } Y) = P(X) \cdot P(Y)$.

Consider now the following frequent sets in some 0/1 relation r:
$fr(\{A\}, r) = 0.505$
$fr(\{B\}, r) = 0.410$
$fr(\{C\}, r) = 0.301$
$fr(\{A, B\}, r) = 0.370$
$fr(\{A, C\}, r) = 0.156$
$fr(\{B, C\}, r) = 0.097$
$fr(\{A, B, C\}, r) = 0.096$

Let's denote the probability $P(X \subseteq t | t \in r)$ of item set X occurring on a random row in r simply by $P(X)$. Obviously the (relative) frequency of an item set is this probability: $fr(X, r) = P(X)$.

- What is confidence $conf(\{A\} \Rightarrow \{B\})$, i.e., the conditional probability $P(B|A)$? What is $conf(\{B\} \Rightarrow \{A\})$?
- What is the joint probability $P(A \text{ and } B)$ of A and B? What is $P(A) \cdot P(B)$?
- What do the above results and the plain frequencies $fr(\{A\})$ and $fr(\{B\})$ tell us about the relationship of A and B? What do similar results about A and C tell?
- What is the J-measure for rules $\{A\} \Rightarrow \{B\}$, $\{B\} \Rightarrow \{A\}$, and $\{A\} \Rightarrow \{C\}$? What does it tell about the relationship of A with B and C?
- Explain, e.g, with a drawing, what is the idea in testing the (nominal) statistical significance of a rule. Use rules $\{B\} \Rightarrow \{A\}$ and $\{A\} \Rightarrow \{C\}$ as examples. (It is not necessary to actually compute the significance levels, as it goes beyond the scope of this course. But if you are interested in doing this, assume the number of rows is 200. It can be a good idea to use normal approximation to the binomial distribution.)

3. Consider supermarket basket analysis as an application for association rules. Combinations of cheap items only are probably not very useful for the shop keeper, so we might want to set a threshold for the total value of an item set, in addition to a frequency threshold. Can you do this easily with Apriori? If yes, how? If not, why not?

4. Rules might be pruned from a discovered set because they do not provide any additional information. Suppose the user is shown the rule

$$\{A, B, C\} \Rightarrow \{E\} \text{ (confidence}_1\text{, frequency}_1\text{)}$$

Under what circumstances would it be advisable to omit showing the rule

$$\{A, B, C, D\} \Rightarrow \{E\} \text{ (confidence}_2\text{, frequency}_2\text{)}$$

5. Association rules can also be applied to text. Use the program mentioned in exercise 1 (last week) to study associations between words in sentences. Take a document (a web page, a report you have written, program source code, ...) and preprocess it to a suitable input format: one line of the input file should correspond to one sentence, with the words of the sentence separated by white space. For plain text this preprocessing can be done with reasonable results with this unix shell command:

   ```
   tr '\012.' ' \012' < document > newfile
   ```

   Alternatively you can analyze letters in words: each input line consists of white space separated letters of one word in the document, and results then tell how characters tend to co-occur in words.

   Run some simple experiments using your data (on text in some language). Do you find anything interesting?

6. What kind of problems does the episode formulation suffer from? (Hint: consider e.g. windows of width 3 on sequences containing fragments such as

   – ...A...A... vs. ...AA...
   – ...$AAABBB$... vs. ...AB...
   – ...AB...AB... vs. ...A...B.....A...B...

   where each position (a letter or a dot) denotes a time point. How do frequencies and confidences of episodes behave?