Exercise Sheet 9

Exercise 32 Statistical Significance

Consider a transaction database with 1000 transactions over 100 items. Assume that each item occurs in a transaction with probability 0.05, so that each item is contained on average in about 50 transactions. Furthermore, assume that the transactions are independent and that the items occur independently in them, that is, the probability of any joint occurrence of items can be computed as the product of probabilities the individual occurrences of the items.

a) What is the probability that the four items with numbers 7, 16, 42 and 78 occur together in *exactly* four transactions?

b) What is the probability that the four items with numbers 7, 16, 42 and 78 occur together in *at least* four transactions?

c) If one wanted to compute the probability that *some* set of four items (*not* a specific set as in a) and b), but any set of four items) occurs in at least four transactions, how could one go about computing this? Where lie main problems? Why is a straightforward extension of the result of b) an overestimate?

Exercise 33 Surrogate Data and Pattern Signatures

In order to determine the statistical significance of found item sets, one may consider an approach based on surrogate (transaction) data sets and pattern signatures.

a) What is a surrogate data set? How could one generate a surrogate data set from a given transaction database? What do surrogate data sets represent?

b) What properties of the original data set should be preserved when generating surrogate data sets?

c) How many surrogate data sets should one generate? Is one enough? Are ten enough? Are 1000 enough? Is there a number that is generally right?

d) What is a pattern signature? Why is it better to consider pattern signatures rather than concrete patterns?

e) How are found item sets filtered with a surrogate data set approach?
Does such filtering yield only relevant patterns?

Exercise 34 Association Rules

a) What are association rules? With what measures are they evaluated?
What is the support of an association rule (two versions)?
What is the confidence of an association rule?

b) How are association rules induced? What steps are needed?
How is the induction related to frequent item set mining?

c) With what minimum support do we have to find frequent item sets for association rule induction? Is it the minimum support of association rules? Or does it depend on the choice of the rule support definition? If yes, how?

d) What relationships hold between the confidence values of different association rules formed from the same item set? How can we exploit these relationships in the generation of association rules?

Exercise 35 Association Rules

a) Are the two rule support definitions equivalent? That is, is there some transformation that one can apply to the used minimum support, so that the same association rules are found? Justify your answer!

b) Are association rules with more than one item in the consequent useful?
What additional information do they provide?

c) What are some measures other than minimum support and minimum confidence with which association rules may be evaluated? From what values are they usually computed?

d) What is the lift (value) of an association rule and how is it defined?
What is the information gain of an association rule?
What is the χ^2-measure of an association rule?