Exercises 1 **T-122.102, Spring 2003**

1. Give examples of 5 data sets where you would typically have binary values, either as a natural or as a transformed representation.

2. A binomial random variable Y, $Y \sim \text{Bin}(n, p)$, describes the probability of y successes out of n trials, where the probability of each success is p. Derive the binomial distribution from the Bernoulli distribution of an individual success,

$$P(R = 1) = r^p(1 - r)^{1-p}.$$

Calculate the expectation $E(Y)$ and variance $Var(Y)$.

Exercises 2 **T-122.102, Spring 2003**

1. Implement the Apriori algorithm (in e.g. Matlab) and try it with some small data matrix. The implementation does not need to be fast, but it should be correct. Especially the candidate generation phase can be tricky; you need not use the optimization mentioned in the presentation.

2. Show that the maximum entropy distribution has the product form given in the presentation:

$$p(x) = \mu_0 \prod_j \mu_j^{1[X_j \subseteq x]}.$$

(Hint: Lagrange multipliers.)

1. Describe an algorithm for the following task. You don't have to implement the algorithm.

   We have market-basket data from a supermarket, and we would like to find out which items are most similar to each other. The first idea for a similarity score might be simply co-occurrence or correlation, but it is not enough; for example, Pepsi Cola and Coca Cola are seldom bought at the same time, but they are intuitively quite similar, because they are bought in similar contexts of other items. Thus, items are similar if they occur in similar baskets; on the other hand, baskets are similar if they contain similar items.

1. Try out either HITS or PageRank on some actual data. Report your findings briefly, and also include a printout of your implementation (Matlab code, or whatever you use).

   If you have some suitable data, please use it. Otherwise, get a data file from the course web page. It is not required to implement an optimal algorithm. If the data is too large, make it small enough that you can run your experiment, but try to select a part of the data that gives nontrivial results. (You can try, e.g., selecting the nodes that have most links to other nodes.)

Exercises 5 **T-122.102, Spring 2003**

1. Consider the following model. There are two biased coins, called A and B. You observe $2n$ throws of the coins: m heads and $2n - m$ tails. You know that exactly half of the throws were performed with coin A, but you do not know which exact throws these are. You wish to determine each coin's probability of coming up heads.

   Explain in detail what it means that the model is unidentifiable.