T-61.5060 Algorithmic methods in data mining

Exercises 1, September 13, 2007

These exercises cover some of the material from previous courses that will be needed in this course.

1. Associative arrays are arrays that can be indexed by, e.g., arbitrary strings. In some programming languages associative arrays are a primitive data type: for example, in Awk we can write
   `A["asdfdfadsf"]=A["asdfdfadsf"]+1;`
   this increments the value associated with the string "asdfdfadsf" by 1. Associative arrays are typically implemented by using hashing. Explain the details how you would implement an associative array.

2. (Some very simple material on probability.) Suppose we are tossing a fair coin 5 times. (a) What is the probability that we get exactly 3 heads? (b) What is the probability that we get at least 3 heads? (c) What is the probability that we get exactly 3 heads given that we get at least 3 heads?

3. Suppose there are three types of coins: type 1 (probability p of a head is 0.5), type 2 ($p = 0.25$) and type 3 ($p = 0.75$). We first pick a coin from a collection of coins, where each type is equally frequent. We then throw the selected coin three times, and get three heads. What is the probability that the coin is of type 1, 2, or 3?

4. What does it mean when we write that the running time of an algorithm in $O(n^2)$?

5. One of the most important facts about random variables is that expectation is a linear operator. Let X be a random variable. Using linearity of expectation derive the equation $Var(X) = E(X^2) - E(X)^2$.

6. Generate (by using, e.g., Matlab) a 0-1 matrix M with 1000 rows (observations) and 100 columns (variables, attributes) such that the entries $M(i, j)$ are 0 or 1 with probability 0.5, independent from each other. Given two rows a and b of the matrix defined their distance $D(a, b)$ as the number of positions j in which $M(a, j)$ and $M(b, j)$ are different

(Hamming distance). Plot the distribution of the distances $D(a, b)$ for a particular row a, and also when a and b vary over all the rows of the table. Comment on the results.

7. As the previous exercise, but let the probability that $M(i, j) = 1$ be proportional to $1/j$. That is, the distributions of the variables is skewed.