T-61.3050 PROBLEMS 1/2007

In T1 on 14 September 2007 at 10 o'clock.

You should solve the problems before the problem session and give the solved problems to the assistant. Please write clearly and leave a wide (left or right) margin. The solutions should be stapled together with a cover sheet containing your name, student number and the numbers of problems you have solved.

For the problems where a "correct" solution exists (math and algorithm questions) the assistant will present one possible solution during the session. In some cases the questions do not have a single correct answer, but the idea is that you think about the problem and are prepared to discuss it with the assistant and other students during the session.

There is no problem session on 21 September; the next problem session will take place on 28 September 2007. See
`http://www.cis.hut.fi/Opinnot/T-61.3050/2007/problems`
for up-to-date information of the problem session.

This problem sheet has two pages.

1. Prerequisite knowledge. The following problems test your prerequisite knowledge. Don't be scared of rather formal notation. If some of the problems is totally alien to you then you may need a refresh on the prerequisite courses.

   (a) (ALGEBRA, PROBABILITIES) Let X be a finite set of items, and $P : X \mapsto \mathbb{R}$ a probability measure on X, that is, P is any function that satisfies $P(x) \geq 0$ for all $x \in X$ and $\sum_{x \in X} P(x) = 1$. Let $f : X \mapsto \mathbb{R}$ be an arbitrary function on X. Define the *expectation of* f by $E[f(x)] = \sum_{x \in X} P(x) f(x)$. Using these definitions, show that:

      i. $E[]$ is a linear operator.

      ii. $E\left[(f(x) - E[f(x)])^2\right] = E[f(x)^2] - E[f(x)]^2$. (Hint: The proof is short if you use linearity.)

   (b) (MATRIX CALCULUS) Let $A \in \mathbb{R}^{n \times n}$ be a symmetrix matrix with eigenvalues λ_i, $i \in \{1, \ldots, n\}$, defined by $Av_i = \lambda_i v_i$, where $v_i \in \mathbb{R}^n$, $v_i^T v_i = 1$ and the eigenvectors have been ordered so that $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$.

i. Show that matrix B has the same eigenvectors and -values as matrix A (that is, A and B are identical matrices) when $B = \sum_{i=1}^{n} \lambda_i v_i v_i^T$ (spectral decomposition).

ii. Compute the Frobenius norm $||A - C||_F^2$, where $C = \sum_{i=1}^{k} \lambda_i v_i v_i^T$ and $k < n$. The Frobenius norm for a matrix $D \in \mathbb{R}^{n \times n}$ is defined as $||D||_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}^2$. (Hint: You should probably use $||D||_F^2 = \mathrm{Tr}\left(D^T D\right)$.)

(c) (ALGORITHMS) The Fibonacci numbers $F(i)$ are defined for $i \in \{1, 2, \ldots\}$ recursively as $F(i+2) = F(i+1) + F(i)$, with $F(1) = F(2) = 1$. Using pseudocode, write down an algorithm that outputs the Fibonacci numbers from 1 to n. Analyze the time complexity of your algorithm using the O-notation. What can you say about efficiency of your algorithm?

2. Read the letter by the SIGKDD Executive Committee on privacy and data analysis, available at
`http://www.acm.org/sigs/sigkdd/civil-liberties.pdf`
Write a short (at most $\frac{1}{2}$ page) position comment on how the privacy issues should be taken into account in practical data analysis.

3. Alpaydin, Chapter 1, Exercise 6 (Section 1.5 on page 15): How can we predict the next command to be typed by the user? Or the next page to be downloaded over the Web? In what ways does this problem differ from the credit-risk example given in the first lecture? In your answer (at most 1 page) you should try to think of one or two ways to formalize the task as a machine learning problem. (I.e., what is the input of your model, what is it's output?)