T-61.3050 PROBLEMS 2/2007

In T1 on 28 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 exceptionally no problem session on 21 September. These problems will be discussed in the 28 September session where you can also submit your solutions. We may have a separate problem sheet for the next week, also to be submitted on 28 September.

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. (Alpaydin (2004) Ch 2, Exercise 3) Why is it better to use the average of S and G as the final hypothesis?

2. (Alpaydin (2004) Ch 2, Exercise 10) Show that the VC dimension of the triangle hypothesis class is 7 in two dimensions. (Hint: For best separation, it is best to place the seven points equidistant on a circle.)

3. Familiarize yourself with a data analysis program of your choosing. If you do not already have a favourite or if you want to learn something new, you can try R or S. R can be downloaded freely from
   `http://www.r-project.org/`
   Download the data set for the 2/2007 problem session from
   `http://www.cis.hut.fi/Opinnot/T-61.3050/2007/problems#2`
   You can also find some example R (or S) code from the same location to get you started. The data is anonymized data set from a real natural source. The data set contains a header line and 4586 rows and 2 columns, "X" and "Y" (in addition to the rowname column). There are 20 rows in which "Y" is known and 4566 in which it is unknown

(NA). Your task is to make a regressor to predict "Y" where it is unknown (NA), given X, with the squared prediction error as small as possible. We will release some randomly chosen values of "Y" ("test set") on the above mentioned web page before 26 September 10 o'clock with which you can asses the performance of your predictor. Report what you have done and compute the average squared error on the test set for the 28 September problem session. If you want, you can try to choose the model complexity also using cross-validation, see section 14.2 of Alpaydin (2004).