T-61.3050 PROBLEMS 11/2007

In T1 on 7 December 2007 at 10 o'clock.

This is the last problem session!

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.

See `http://www.cis.hut.fi/Opinnot/T-61.3050/2007/problems` for up-to-date information of the problem sessions.

This problem sheet has two pages.

1. In comparing classification or regression models, it is often useful to have a *dummy model* as a baseline. By definition, the prediction of a dummy model is constant with respect to the covariates. Any reasonable classification or regression model (for a given data) should be able to beat the dummy model (sometimes the dummy model is surprisingly good, however, for example, in classification task when the class distribution is strongly skewed). What is a good dummy model for a (i) classification and a (ii) regression task? What would be a good dummy model for an unsupervised learning task, such as clustering?

2. Repeat the analysis of problem set 2/2007, problem 3, using K-fold cross-validation. Do you get a better result in the test data set?

3. Consider a simple binary classifier using decision boundary at $w \in \mathbb{R}$, where the data is given by $\mathcal{X} = \{(r^t, x^t)\}_{t=1}^N$, where $r^t \in \{0, 1\}$ and $x^t \in \mathbb{R}$.

   (a) Create a toy data, find an optimal classifier and write down the confusion matrix.

   (b) Sketch a Receiver Operating Characteristic (ROC) curve for the classifier.

(c) What is the relation of the ROC curve to the losses and risks (Alpaydin, Ch 3.3)?

(d) How could you draw a ROC curve for a more complex classifier, say, Naive Bayes classifier that outputs a class probability $P(r \mid \mathbf{x})$?

(e) What would a ROC curve for the dummy model look like?

(f) Some people use the area under the ROC curve as a measure of classifier performance. Why does this make sense? (And invent at least one scenario where the area under the ROC curve is not a good measure of classifier performance.)

4. Consider the data and classifier of Problem 3. Show, for example by using McNemar's test, that your classifier beats the dummy classifier in the validation set.

Due to lecturer's flu there was no lecture on these topics. The K-fold cross validation, ROC curve and hypothesis testing is explained in Alpaydin (2004), Chapter 2004, and in other literature on statistics and machine learning. Also, there are several sources freely available at the web (use your favourite search engine).