T-61.3050 PROBLEMS 7/2007

In T1 on 2 November 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.

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

There is no problem session on 26 October and no lecture on 30 October due to the examination period. The problem sessions and lectures will continue on 2 November and 6 November, respectively.

This problem sheet has two pages.

"Fossil week": You should read the description of the FOSSILS data set and download it from the course web site at
`http://www.cis.hut.fi/Opinnot/T-61.3050/2007/problems#7`

Hint for problems 1–3: You may find the R-code with the FOSSILS data set at the course web site helpful.

1. Implement the Naive Bayes classifier for binary data, using Bernoulli distribution. (Hint: You should probably add a prior observation count to the estimated probabilities to avoid zero probabilities which cause trouble, if you take logs of them. In practice, this would mean that instead of equation (5.31) of Alpaydin (2004) [the below equation without α; the same equation appears in lecture slides] you should estimate the Bernoulli parameters as

$$\hat{p}_{ij} = \frac{\alpha + \sum_t x_j^t r_i^t}{2\alpha + \sum_t r_i^t},$$

where α is a positive constant ("prior observation count", you can for example use $\alpha = 1$).)

(a) Make a classifier that predicts whether Hipparion exists at a given fossil site, given information of the existence of all other taxa at various fossil sites. (Notice that in principle, you could use similar classifier to predict the taxa at a given fossil site, given taxa of all other fossil sites.)

(b) Look at the parameters of the classifier. Can you say which other taxa are important in predicting the existence of Hipparion and why?

(c) As explained in the description of the data at the course web site, the data has many "false zeroes" (taxon is not observed on a fossil site, even if the fossil site was dated within the taxon's lifetime). How could you use these classifiers to identify false zeroes?

2. Use the subset selection to find a good set of parameters for the classifier of Problem 1. Which taxa are most important in predicting the existence of Hipparion and why?

3. Analyze the fossil sites as well as the taxa using PCA.

   (a) Plot the fossil sites and taxa to a plane, using the first two principal components as coordinate axis. Can you interpret the plots?

   (b) Print out the largest eigenvectors. What do they tell you about the data?

   (c) The principal components are the eigenvectors of either the covariance or correlation matrix of the data. What is the practical difference between these two approaches?

   (d) Study the proportion of variance explained as a function of the number of eigenvectors. What would be a good choice for the number of principal components that could be used to present the data?

4. In the Problem Set 1/2007, problem 1b, spectral decomposition was presented. What is the relation between the PCA and spectral decomposition?