T-61.3050 PROBLEMS 8/2007

In T1 on 9 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.

This problem sheet has two pages.

1. Linear Discriminant Analysis

   (a) Consider two classes, C_1 and C_2, generated by two bivariate Gaussians ($C_i \sim N(\mu_i, \Sigma_i)$) with $\mu_1 = [5, 4]$, $\mu_2 = [-3, -2]$ and $\Sigma_1 = \Sigma_2 = I$, where I is the identity matrix. What is the direction of the linear discriminant? (A geometric argument is sufficient.)

   (b) Implement LDA using R or Matlab (Or any other programming language that has routines for computing matrix inversion and the spectral decomposition). Generate points from the Gaussians of part a) and compare your result with the one you obtained above.

   (c) Download the data set from
   `http://www.cis.hut.fi/Opinnot/T-61.3050/2007/lda_test_1.txt`
   and plot it to two dimensions using LDA. Do the same using PCA and compare the results.

2. k-means and Lloyd's algorithm.

   (a) Show that the error

   $$\mathcal{E}(\{\mathbf{m}_i\}_{i=1}^{k} | \mathcal{X}) = \sum_{t=1}^{N} \min_i ||\mathbf{x}^t - \mathbf{m}_i||^2$$

   can not increase in either step of Lloyd's algorithm.

(b) Implement Lloyd's algorithm for solving k-means. Select the initial centroids by picking k data points uniformly at random.

(c) Download the data set from `http://www.cis.hut.fi/Opinnot/T-61.3050/2007/kmeans_test_1.txt` and cluster the points using your algorithm with $k = 2$. Run your algorithm a number of times using different choices for the initial centroids. What do you observe? Why does this happen? Suggest an alternative way of initializing the centroids to overcome the problem.

(d) We have already talked about model selection in the previous problem sessions. In case of clustering the question is how to set k, i.e., what is the correct number of clusters? Download the data set from `http://www.cis.hut.fi/Opinnot/T-61.3050/2007/kmeans_test_2.txt` and run your algorithm on it. How many clusters are there?