

# Assignment 2: Density Estimation

## Overview of Statistical Machine Learning

due 08.05.2006

In this assignment we'll try various density estimation techniques on the Fischer Iris data set (from the course website). This is unsupervised learning, so make sure you don't use the class labels (5th column) :-)

**Parametric:** Calculate the mean and covariance matrix for the 4-D Gaussian that best fits the data in the maximum likelihood sense. Calculate the log-likelihood of the data under this model.

**Nonparametric:** Now model the data density with Parzen windows, using isotropic (*i.e.*, spherical) Gaussian kernels of fixed width  $\sigma = 5$ , and report the resulting log-likelihood of the data.

**Semiparametric:** Now fit a mixture of 3 Gaussians to the data, using the EM algorithm. Plot the log-likelihood of the data against the EM iteration number, for 10 restarts from different random initial conditions (*i.e.*, means and covariances). Do you find the same solution each time?

For one of the EM solutions: assign each data point to the most likely Gaussian. How well does that correspond to the class labels (5th column) of the data? In general, do you think mixture density estimation can serve as a basis for classification? In what situation (common in practice) would you have to do this?

**Comparison:** Briefly compare the above 3 methodologies in terms of expressive power (*i.e.*, achieved log-likelihood), ease of implementation, and computational cost (especially how that scales to large data sets). When would you use which?

Please hand in your calculations, results, plots, *etc.* to Jin Yu.