Stochoastic Modeling and Simulation

Course Project

Due

The project asks you to develop two tools that you will use to analysis data that comes from a mixture of normal distributions. The first of these two tools will be a function that returns a two dimensional B-spline density estimator. You will use this nonparametric density estimator to visually inspect the data by looking at the graph of the nonparametric density estimate.

The second tool will be an implementation of the the EM algorithm for a mixture of multivariate normal distributions. The dimension of data and the number of classes should be arbitrary.

The third part of the project will be to determine the number of classes using the AIC (to be discussed shortly) and using your two tools to determine your best estimate of the number of classes and the maximum likelihood estimate.

The final portion of the project is to write a report that describes the tools that you have developed, how those tools where used and to explain your final analysis of the data. Short pieces of code can be used in the report, although the body of the code will be submitted via email. You should make good use of graphics in your report to help explain the data and how you determined the final parameters and number of classes.

Part I. Create and test a function whose input is a two dimensional sample, the minimum and maximum x and y values (say xmin, xmax,ymin and ymax), along with the bin width in each of the x and y directions. The function output is the values of the two dimensional B-spline density estimator based on that data evaluated at a set of points. You should vectorize the function so that the input can be a set of grid points. These values will be used to graph the two-dimensional density function.

Your report should provide examples, including graphs, that demonstrate that the code is working.

Part II. Create and test a function whose input is the initial estimate of the parameters of a mixture of multivariate normal density functions along with at least one parameter that is used as a stopping rule for the em algorithm. The function should be an implementation of the em algorithm and should work in any number of dimensions and any number of classes. The output should be maximum likelihood estimate of the model parameters.

Your report should mathematically describe the em algorithm that you have implemented (but not the underlying theoretical development) and provided examples that demonstrate that the algorithm is working effectively. A wise selection of a few examples is important to convince the reader that the code is working and is reasonably robust.

Part III. Now we are ready to try to find a model that gives us the best fit for a data set that I will provide. Note that when models are the same size, the likelihood function alone can be used to pick the best model. When the number of classes is variable and hence the number of parameters if variable, one can use the AIC to compare functions with different numbers of parameters. Develop tools that allow you to compare multiple models with a variety in the number of classes and different estimates of the maximum likelihood function. It is also appropriate for you to look at level contours of your B-spline density function.

Note that even for two models of the same size, the em algorithm may give quite different estimates of the model parameters. So you should use various starting values to help ensure that you have a good overall estimate.

Your report should describe the mathematics that was implemented and the strategy that you used to select good starting values to initiate the em algorithm along with other details as appropriate.

Par IV. The final portion of the assignment is to finalize the report. It should be well written, complete, but not extremely long or verbose. A good way to close the report would be to summarize your experiences with the project. Interesting problems that you had to solve are often useful. Warnings or an explanation of potential pitfalls to others who are asked to solve a similar problem are certainly appropriate. A reader who was not imminently familiar with this sort of problem should come away with a much better understanding of how to solve the problem.