Caen University Master 2 SAAD/DECIM

Practical work

Gaussian mixture model (GMM), EM, and model selection

by

Faicel Chamroukhi

Professor

Generate data from a Gaussian mixture model:

- Generate a two-dimensional dataset from a K-component Gaussian mixture density with different locations (means) and different covariance matrices.
- To do this, for each data point,
  1. first, the class label (z_i) (the label of the Gaussian component to be generated from it) is selected according to a multinomial distribution with parameter vector the mixing proportions (pi_1,...pi_K) which you would also choose

     Store the class labels for each generated data point (to perform later comparison)
  2. Given the class label (z_i), generate a data vector (x_i) according to the corresponding Gaussian component N(mu_{z_i},Sigma_{z_i})

EM for a GMM

- Implement the EM algorithm to estimate a K-component Gaussian mixture density:
- Initialize the mixing propositions and the covariance matrices (e.g., equal mixing proportions and Identity covariance matrices)
- Initialize the means locations “randomly” (by your own choice of K vectors from R^d) or initialize them with standard K-means clustering (use your own K-means code or the one provided by Matlab, or your own function :))
- in the EM training loop, store the value of the observed-data log-likelihood at each iteration
- At convergence, plot the log-likelihood curve and plot the estimated density and the corresponding MAP partition (use the scatter plot, gscatter ..); you may need the function to draw ellipse densities (function)

Model Selection (next practical work)

- Now select the number of mixture components by computing the values of a chosen model selection criterion (BIC, AIC, AIC3, ICL,...) for K varying from 1 to 10. Each EM run would correspond to a value of the model selection criterion
- Compare your results with the ground truth (in terms of the chosen number of mixture components; and in terms of classification error rate for K=3)

Real data:

- load the iris dataset
- Do the same job (You can do the same job with other data sets)

Enjoy!