SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.570 Machine Learning, Fall 2017

Assignment 3

Last update: Thursday 26th October, 2017 08:46

Grading Policy and Due Date

- You are required to submit: 1) a report that summarizes your experimental results and findings, based on each of the following question asked; 2) your implementation (source code) of the algorithms.
- You are free to choose any programming language you prefer.
- Submit your assignment report and code to eDimension.
- This assignment is an individual assignment. Discussions amongst yourselves are allowed and encouraged, but you should write your own code and report.
- Submit your assignment to the eDimension by 11:59 PM on Thursday 16 November 2017. Late submissions will be heavily penalized (20% deduction per day).

1 Task: Mixture of Gaussians & EM

Download the file `em-data.zip` from the course web page and unzip it.

1. (10 pts) Assume the data comes from a mixture of $k = 2$ spherical Gaussian distributions. Now you would like to estimate the parameters for the two Gaussian distributions. Discuss what are the model parameters to be estimated, and what are the hidden variables for the mixture model.
2. (10 pts) Now, use the hard EM algorithm discussed in class to estimate the model parameters. Discuss how you initialize the model and when to terminate the parameter estimation process. Plot the value of the (log) joint likelihood associated with the data as a function of the number of EM iterations. Discuss the behavior of the curve. Think of a way to visualize the final result (*e.g.*, you may use graph to show (partial) membership of each individual data point).

   (Hint: to overcome potential numerical overflow issues, you might find the logsumexp trick useful. See this page: `https://en.wikipedia.org/wiki/LogSumExp`)
3. (10 pts) Now, instead of using the hard EM algorithm, let us use the soft EM algorithm, and repeat the above question.
4. (10 pt) Now, consider the soft-EM algorithm, assume we don't know the correct number of components k in the mixture model. Develop an idea to automatically choose the optimal k. Clearly present your idea, and clearly illustrate any outputs/observations that are used to make your decision on the selection of k.