DATA MINING W4240

HOMEWORK 3 QUESTIONS

October 14, 2010

Professor: Frank Wood

Preliminary Instructions

1. Download the skeleton code for the assignment at
   http://www.stat.columbia.edu/~fwood/w4240/Homework/index.html

2. Unzip the downloaded material in an appropriate folder, something like w4240/hw3/

3. Open MATLAB and navigate to the folder containing the downloaded material

In this home work you will need to implement the expectation and maximization steps of the EM algorithm for a Bayesian linear regression and a classical gaussian mixture model.

1. (50 points) Implement the the functions **e_step_linear_regression** and **m_step_linear_regression** to implement the EM algorithm for Bayesian linear regression. You will need to consult the book and the programs provided to understand the function signatures. Make sure the functions will run on any dimensional design matrix with a real valued output. I recommend you test your functions by writing a wrapper for the functions and testing them on synthetic data.

2. (50 points) Implement the functions **e_step_gaussian_mixture**, **m_step_gaussian_mixture**, and **log_likelihood_gaussian_mixture** to implement the EM algorithm for a gaussian mixture model. Make sure the functions will run on any dimensional real vector valued data with any chosen number of components. I recommend you test your functions by writing a wrapper for the functions and testing them on synthetic data.

You must complete this HW assignment on your own, you are not permitted to work with any one else on the completion of this task. Your grade will reflect your ability to implement a working version of the procedure. Submitted code must run on my machine in less than 3 minutes. Grading will be automated and the submitted files will be run, therefore to submit the HW you will need to follow the following directions exactly.

1. Send an email to w4240.fall2010.stat.columbia.edu@gmail.com

2. Attach your updated MATLAB files

   (a) e_step_linear_regression.m

   (b) m_step_linear_regression.m

   (c) e_step_gaussian_mixture.m

   (d) m_step_gaussian_mixture.m

   (e) log_likelihood_gaussian_mixture.m

   It is imperative that the names be exactly as described here. There should be no folders attached, only raw .m files. You may not attach other MATLAB code files.

3. The subject will be exactly your Columbia UNI followed by a colon followed by hw3. For example, if the TA were submitting this homework the subject would read **nsb2130:hw3**

4. If you submit hw more than once, later files will overwrite earlier files.