DATA MINING W4240

HOMEWORK 3 QUESTIONS

March 10, 2011

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. The model being fit here is as follows

$$\begin{aligned} Y_i &\sim \text{Normal}(X_i^t \vec{w}, 1/\beta) \;\; \forall\, i \in \{1, \ldots, n\} \\ w_j &\sim \text{Normal}(0, 1/\alpha) \;\; \forall\, j \in \{1, \ldots, \text{length}(X_i)\} \end{aligned}$$

The only parameters of interest are α and β and we wish to fit their values using maximum likelihood. This requires the EM algorithm because you will integrate over the values of the weights w. You will need to consult the book and the programs provided to understand the function signatures. Data is provided in the main file which you should use while developing your code. However, make sure the functions will run on any dimensional design matrix. I recommend you test your functions by 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. The data you are using is the Fisher iris data. Each row of data consists of four measurements made regarding an iris flower. You will need to cluster the measurements

using EM on a gaussian mixture model. You should make sure that the program runs no matter what choice of k you make. Also, 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 testing them on synthetic data.

3. (25 points) For w6240 ONLY You need to write and submit a main_6240.m file which implements the a stochastic gradient ascent version of the EM algorithm for the gaussian mixture model. If you need to create additional functions to make this work, please submit those as well. By online stochastic gradient ascent I mean the following; consider that at each step of the EM algorithm the parameters of interest are updated to maximize the expectation (taken with respect to the posterior distribution conditioned on all the data) of the log likelihood (of all the data). To create an online stochastic gradient ascent approach, first perform the e-step on only a subset of the data. Then, find the values of the parameters which maximize the expected log likelihood of that same subset of data. Step the parameter values in the direction of the calculated maxizing parameters (on the subset) of length η. Please implement this by looping through the data and using small, sequential chunks of the data. Iterate until convergence, but note that the log likelihood of the entire data set will not increase at every step.

Submitting your HW

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.spring2011.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.