DATA MINING W4240

HOMEWORK 4 QUESTIONS

March 25, 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/hw4/
3. Open MATLAB and navigate to the folder containing the downloaded material

In this homework you need to implement variational Bayesian inference for the gaussian mixture model. You will also need to implement the k-means algorithm. The clustering obtained by k-means will then be used as an initialization for the variational algorithm.

1. (25 points) Implement a k-means clustering algorithm. This should be done by filling out the two functions **update_cluster_centers.m** and **k_means.m**. The final return will be a list of cluster assignments, one for each data point. After writing these functions you should be able to cluster d-dimensional data using the algorithm. Since you will be using the output as an initialization you should program the algorithm in such a way that it returns k clusters, i.e., that it does not collapse to fewer clusters than the input number k. If you find that a cluster has no data points assigned to it during the algorithm, one potential solution is to assign that cluster mean to the value of a randomly chosen data point. K-means is an iterative algorithm, but in this code you can hardcode the number of iterations to 10.

2. (75 points) Implement the variational Bayes inference algorithm for a gaussian mixture model and evaluate its performance on Fisher's Iris data. To do this, you will need to implement the functions **get_r.m**, **get_other_parameters.m**, **expected_rand_index.m**, **simulate_z.m**, and **variational_lower_bound.m**. More explicit direction regarding the functionality of these programs can be found in the header of each file. The naming of variables in these programs is designed to mirror the example given in your text book. If you are confused about what certain variables are please refer there first before asking the TA. You will also need to fill out values for K and the parameters of the prior distributions in the

file **main.m**. As the iteration progresses you should note that the variational lower bound continues to increase and the algorithm only stops when there is apparent convergence.
After convergence is reached, the code in the main file is designed to answer the question, how many of the K mixture components actually show up in the data. Given that the data is a classification of Iris plants based on measurements of the plant, this question translates into "How many species of Iris are in the population we are examining?". To get a valid answer to this question you need to set K high enough that it can collapse to a value supported by the data. The answer to this question is estimated through sampling. That is, given the estimated distribution over the cluster assignments Z, we sample Z and then consider how many cluster centers the sample is allocated. This allows us to get an empirical estimate of the distribution of the number of cluster centers in the estimated posterior. To get a point estimate we consider the expected number of cluster centers which we estimate with the mean of the sample.
The Rand index is a metric for comparing clusterings of data. A low value of the Rand index indicates that two clusterings are close. Consider a clustering $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$ where a_i is the set of elements in the data X in cluster i, and define the function SameCluster$(i, j, \mathcal{A})$ $= 1$ if $x_i, x_j \in a_h$ and 0 otherwise. Then, given two clustering $\mathcal{A}$ and $\mathcal{B}$ we define the Rand index as

$$\frac{1}{\binom{N}{2}} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left[(\text{SameCluster}(i, j, \mathcal{A}) + \text{SameCluster}(i, j, \mathcal{B})) \mod 2\right].$$

Since the result of the variational Bayes procedure is a distribution, we do not have a clustering directly, but instead have a distribution over clusterings. So, instead we need to use the expected Rand index to compare a clustering to the true clustering. Since the distribution from the variational approximation has independent cluster assignments this means we can inspect each pair of observations separately to find the probability they are in the same cluster for each of the K clusters. Then, compare the probability that they are in the same cluster to the true labels and calculate the expectation of the contribution the pair makes to the Rand index. This is done separately for each pair of data points and the result is divided by $\binom{N}{2}$ to get the expected Rand index.
You will notice in the main file that the output from the k-means clustering algorithm is used to inform the initialization of the variational Bayes algorithm. Also, included in the HW material is a function which plots d-dimensional data by plotting the first two PCA components. This is just a trick to help in seeing the clusters in d-dimensional data when $d > 2$.

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) **update_cluster_centers.m**
   (b) **k_means.m**
   (c) **get_r.m**
   (d) **get_other_parameters.m**
   (e) **expected_rand_index.m**
   (f) **simulate_z.m**
   (g) **variational_lower_bound.m**
   (h) **main.m**

   It is imperative that the names be exactly as described here. There should be no folders attached, only raw .m files. You may attach other MATLAB code files if they act as utility functions for the other programs.
3. The subject will be exactly your Columbia UNI followed by a colon followed by hw4. For example, if the TA were submitting this homework the subject would read **nsb2130:hw4**
4. If you submit hw more than once, later files will overwrite earlier files.