EECS E6892: Bayesian Models for Machine Learning

Columbia University, Fall 2015

## Homework 4: Due Monday, December 14, 2015 by 11:59pm

**Please read these instructions to ensure you receive full credit on your homework.** Submit the written portion of your homework as a *single* PDF file through Courseworks (less than 5MB). In addition to your PDF write-up, submit all code written by you in their original extensions through Courseworks (e.g., .m, .r, .py, etc.). Any coding language is acceptable. Do not wrap your files in .rar, .zip, .tar and do not submit your write-up in .doc or other file type. Your grade will be based on the contents of *one* PDF file and the original source code. Additional files will be ignored. We will not run your code, so everything you are asked to show should be put in the PDF file. Show all work for full credit.

**Late submission policy:** Late homeworks will have 0.1% deducted from the final grade for each minute late. *Your homework submission time will be based on the time of your* ***last*** *submission to Courseworks. I will not revert to an earlier submission!* Therefore, do not re-submit after midnight on the due date unless you are confident the new submission is significantly better to overcompensate for the points lost. Submission time is non-negotiable and will be based on the time you submitted your last file to Courseworks. The number of points deducted will be rounded to the nearest integer.

### Problem Set-up

We are given observations $X = \{x_1, \dots, x_n\}$ where each $x_i \in \mathbb{R}^d$. We model this as being generated from a Gaussian mixture model of the form

$$x_i \,|\, c_i \sim Normal(\mu_{c_i}, \Lambda_{c_i}^{-1}), \qquad c_i \overset{iid}{\sim} Discrete(\pi)$$

In this homework, you will implement three algorithms for learning this mixture model, one based on maximum likelihood EM, one on variational inference and one on Gibbs sampling. Use the data provided for all experiments.

**Problem 1.** (30 points)

In this problem, you will implement the EM algorithm for learning maximum likelihood values of $\pi$ and each $(\mu_j, \Lambda_j)$ for $j = 1, \dots, K$. The algorithm is given in the notes, and also in Section 9.2 of Bishop's book.

a) Implement the EM-GMM algorithm and run it for 100 iterations on the data provided for $K = 2, 4, 8, 10$.

b) For each $K$, plot the log likelihood over the 100 iterations. What pattern do you observe and why might this not be the best way to do model selection?

c) For the final iteration of each model, plot the data and indicate the most probable cluster of each observation according to $q(c_i)$ by a cluster-specific symbol. What do you notice about these plots as a function of $K$?

**Problem 2.** (35 points)

In this problem, you will implement a variational inference algorithm for approximating the posterior distribution of the GMM variables. We therefore require prior distributions on these variables. For this problem, we use

$$\pi \sim Dirichlet(\alpha), \qquad \mu_j \sim Normal(0, cI), \qquad \Lambda_j \sim Wishart(a, B)$$

For this problem, set $\alpha = 1$, set $m$ to be the empirical mean of the data, $c = 10$, $a = d$ and $B = \frac{d}{10}A$ where $A$ is the empirical covariance of the data. Approximate the posterior distribution of these variables with $q$ distributions factorized on $\pi$, and each $\mu_j$, $\Lambda_j$ and $c_i$ as discussed in class.

a) Implement the variational inference algorithm discussed in class and in the notes for $K = 2, 4, 10, 25$ and 100 iterations each.

b) For each $K$, plot the variational objective function over the 100 iterations. What pattern do you observe?

c) For the final iteration of each model, plot the data and indicate the most probable cluster of each observation according to $q(c_i)$ by a cluster-specific symbol. What do you notice about these plots as a function of $K$?

**Problem 3.** (35 points)

In this problem, you will implement a Bayesian nonparametric sampler for a marginalized version of the GMM. In contrast to Problem 2, in this problem we will use a joint prior on $(\mu_j, \Lambda_j)$. This is done for computational convenience in calculating the marginal distribution of the data. Specifically, we use the prior distribution

$$\mu_j \,|\, \Lambda_j \sim Normal(m, (c\Lambda)^{-1}), \qquad \Lambda_j \sim Wishart(a, B)$$

as well as the limit of the prior $\pi \sim Dirichlet(\alpha/K, \dots, \alpha/K)$ as $K \to \infty$.

In this problem you will implement the marginal sampler where $\pi$ is integrated out. For this problem, set $m$ to be the empirical mean of the data, $c = 1/10$, $a = d$ and $B = c \cdot d \cdot A$ where $A$ is the empirical covariance of the data. For the "cluster innovation parameter" set $\alpha = 1$.

a) Implement the above-mentioned Gibbs sampling algorithm discussed in class and described in the notes. Run your algorithm on the data provided for 500 iterations.

b) Plot the number of observations per cluster as a function of iteration for the six most probable clusters. These should be shown as lines that never cross; for example the $i$th value of the "second" line will be the number of observations in the second largest cluster after completing the $i$th iteration. If there are fewer than six clusters then set the remaining values to zero.

c) Plot of the total number of clusters that contain data as a function of iteration.