EECS E6720: Bayesian Models for Machine Learning

Columbia University, Fall 2018

Homework 4: Due Wednesday, November 28, 2018 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, but your code should be your own. Do not submit code notebooks such as Jupyter. 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.* 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. You can resubmit as much as you like, but each time you resubmit be sure to upload **all** files you want graded! 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

You are given observations $X = \{x_1, \dots, x_n\}$ where each $x_i \in \{0, 1, 2, \dots, 20\}$. You model this as being generated from a Binomial mixture model of the basic form

$$x_i \,|\, c_i \sim \text{Binomial}(20, \theta_{c_i}), \qquad c_i \overset{iid}{\sim} \text{Discrete}(\pi).$$

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 where π is integrated out. Use the data provided for all experiments.

Problem 1. (30 points)

In this problem, you will derive and implement the EM algorithm for learning maximum likelihood values of π and each θ_k for $k = 1, \dots, K$. Use c as the model variable being integrated out.

a) Show your derivation of the E and M steps of the algorithm. You can present your solution for π at the same level of detail as in the notes. Summarize the algorithm in pseudo-code.

b) Implement the EM algorithm and run it for 50 iterations for $K = 3, 9, 15$. In one figure, plot the log marginal likelihood over iterations 2 to 50 for each K.

c) For the final iteration of each model, for integers 0 through 20 plot a scatter plot that has the integer on the x-axis and the index of the most probable cluster on the y-axis. (You will use $p(c|x, \theta)$ to find this.) Show this in three separate figures.

Problem 2. (35 points)

In this problem, you will implement a variational inference algorithm for approximating the posterior distribution of the Poisson mixture model. As additional priors, let

$$\pi \sim \text{Dirichlet}(\alpha), \qquad \theta_k \overset{iid}{\sim} \text{Beta}(a, b),$$

Set $\alpha = 1/10$, $a = 0.5$ and $b = 0.5$. Approximate the full posterior distribution of π, $\boldsymbol{\theta}$ and $\mathbf{c}$ with the factorized distribution $q(\pi, \boldsymbol{\theta}, \mathbf{c}) = q(\pi)\left[\prod_{k=1}^{K} q(\theta_k)\right]\left[\prod_{i=1}^{n} q(c_i)\right]$ using variational inference.

a) Derive the optimal $q(\pi)$, $q(\theta_k)$ and $q(c_i)$. Summarize the VI algorithm in pseudo-code.

b) Implement the resulting VI algorithm and run it for 1000 iterations for $K = 3, 15, 50$. In one figure, plot the variational objective function for each K over iterations 2 to 1000.

c) For the final iteration of each model, for integers 0 through 20 plot a scatter plot that has the integer on the x-axis and the index of the most probable cluster on the y-axis. (You will use $q(c)$ to find this.) Show this in three separate figures.

Problem 3. (35 points)

In this problem, you will implement a Bayesian nonparametric Gibbs sampler for a marginalized version of the model in Problem 2 using the ideas from Lecture 10. This means that in theory we use the limit of the prior $\pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$ for $K \to \infty$, but derive a marginal sampler for $p(\boldsymbol{\theta}, \mathbf{c}|\mathbf{x})$ with π integrated out. Set $\alpha = \frac{3}{4}$, $a = 0.5$ and $b = 0.5$.

a) Derive the conditional posteriors you will need: $p(\theta_k|\{x_i : c_i = k\})$ and $p(c_i|x_i, \boldsymbol{\theta}, c_{-i})$. Implement the resulting Gibbs sampling algorithm discussed in class and the notes, but applied to this specific data-generating distribution. Run your algorithm on the data provided for 1000 iterations with the initial number of clusters set to 30.

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 on the same plot that never cross; for example the m-th value of the "second" line will be the number of observations in the second largest cluster after completing the m-th iteration. If there are fewer than six clusters in an iteration then set the remaining values to zero.

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