EECS E6720: Bayesian Models for Machine Learning

Columbia University, Fall 2016

Homework 3: Due Monday, November 14, 2016 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 1. (50 points)

We have a data set of the form $\{(x_i, y_i)\}_{i=1}^N$, where $y \in \mathbb{R}$ and $x \in \mathbb{R}^d$. We assume d is large and not all dimensions of x are informative in predicting y. Consider the following regression model for this problem:

$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T w, \lambda^{-1}), \quad w \sim \text{Normal}(0, \text{diag}(\alpha_1, \ldots, \alpha_d)^{-1}),$$

$$\alpha_k \overset{iid}{\sim} \text{Gamma}(a_0, b_0), \quad \lambda \sim \text{Gamma}(e_0, f_0).$$

Use the density function $\text{Gamma}(\eta|\tau_1, \tau_2) = \frac{\tau_2^{\tau_1}}{\Gamma(\tau_1)} \eta^{\tau_1 - 1} \text{e}^{-\tau_2 \eta}$. In this homework, you will derive a variational inference algorithm for approximating the posterior distribution with

$$q(w, \alpha_1, \ldots, \alpha_d, \lambda) \approx p(w, \alpha_1, \ldots, \alpha_d, \lambda | y, x)$$

a) Using the factorization $q(w, \alpha_1, \ldots, \alpha_d, \lambda) = q(w)q(\lambda)\prod_{k=1}^d q(\alpha_k)$, derive the optimal form of each q distribution. Use these optimal q distributions to derive a variational inference algorithm for approximating the posterior.

b) Summarize the algorithm derived in Part (a) using pseudo-code in a way similar to how algorithms are presented in the notes for the class.

c) Using these q distributions, calculate the variational objective function. You will need to evaluate this function in the next problem to show the convergence of your algorithm.

Problem 2. (50 points)

Implement the algorithm derived in Problem 1 and run it on the three data sets provided. Set the prior parameters $a_0 = b_0 = 10^{-16}$ and $e_0 = f_0 = 1$. We will not discuss sparsity-promoting "ARD" priors in detail in this course, but setting a_0 and b_0 in this way will encourage only a few dimensions of w to be significantly non-zero since many α_k should be extremely large according to $q(\alpha_k)$.

For each of the three data sets provided, show the following:

a) Run your algorithm for 500 iterations and plot the variational objective function.

b) Using the final iteration, plot $1/\mathbb{E}_q[\alpha_k]$ as a function of k.

c) Give the value of $1/\mathbb{E}_q[\lambda]$ for the final iteration.

d) Using $\hat{w} = \mathbb{E}_{q(w)}[w]$, calculate $\hat{y}_i = x_i^T \hat{w}$ for each data point. Using the z_i associated with y_i (see below), plot $\hat{y}_i$ vs z_i as a solid line. On the same plot show (z_i, y_i) as a scatter plot. Also show the function $(z_i, 10 * \text{sinc}(z_i))$ as a solid line in a different color.

   Hint about Part (d): z is the horizonal axis and y the vertical axis. Both solid lines should look like a function that smoothly passes through the data. The second line is ground truth.

Details about the data

The data was generated by sampling $z \sim \text{Uniform}(-5, 5)$ independently N times for $N = 100, 250, 500$ (giving a total of three data sets). For each z_n in a given data set, the response $y_n = 10 * \text{sinc}(z_n) + \epsilon_n$, where $\epsilon_n \sim N(0, 1)$.

We use z_n to construct a "kernel matrix" X. This is a mapping of z_n into a higher dimensional space (see Bishop for more details). For our purposes, it's just important to know that the nth row (or column, depending on which data set you use) of X corresponds to the location z_n. We let $X_{n,1} = 1$ and use the Gaussian kernel for the remaining dimensions, $X_{n,i+1} = \exp\{-(z_n - z_i)^2\}$ for $i = 1, \ldots, N$. Therefore, the dimensionality of each x_i is one greater than the number of data points. The sparse model picks out the relevant locations within the data for performing the regression.

Each data set contains the vector y, the matrix X and the vector of original locations z. This last vector will be useful for plotting.