EECS E6720: Bayesian Models for Machine Learning

Columbia University, Fall 2016

Homework 2: Due Friday, October 21, 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. (40 points)

Assume we have $\{x_1, \ldots, x_N\}$ with $x \in \mathbb{R}^d$. We model them as independent random variables with $x_n \sim N(Wz_n, \sigma^2 I)$. The matrix $W \in \mathbb{R}^{d \times k}$ is unknown and we model $z_n \sim N(0, I)$. The columns of W have zero-mean spherical Gaussian priors with variance λ^{-1}, which can be written according to a matrix Gaussian prior as

$$p(W) = \left(\frac{\lambda}{2\pi}\right)^{dk/2} \exp\left\{ -\frac{\lambda}{2}\text{trace}(W^T W)\right\}.$$

Our goal is to find

$$W' = \arg\max_W \ \ln p(x_1, \ldots, x_N, W).$$

Derive an EM algorithm for doing this where $z_1, \ldots, z_N$ function as the "hidden variables" we integrate out. As a reminder, at iteration t this involves calculating posteriors for each z_n given W_{t-1}, taking expectations of the log joint likelihood using each $q(z_n)$, and maximizing over the result with respect to W. Clearly describe every step in your derivation of the algorithm for full credit, both mathematically and in words. Also, summarize the steps of the algorithm at the end using pseudo-code.

(Side comment: There is no correct pseudo-code language we are looking for. We simply want to see a clear picture of the algorithm you would implement without having to search for the pieces in your response. This should be in addition to your detailed derivation.)

Problem 2. (35 points)

In this problem, you will implement an EM algorithm for the probit regression model using the same digits data set from the last homework. A detailed discussin of EM for probit regression can be found in the class notes. For this problem do the following:

a) Implement the EM algorithm for probit regression described in the class notes. Use the parameter setting $\sigma = 1.5$ and $\lambda = 1$. Run your algorithm on the data set provided for $T = 100$ iterations.

b) Plot $\ln p(\vec{y}, w_t | X)$ as a function of t.

c) Make predictions for all data in the testing set by assigning the most probable label to each feature vector. In a 2×2 table, list the total number of 4's classified as 4's, 9's classified as 9's, 4's classified as 9's, and 9's classified as 4's (i.e., a confusion matrix). Use the provided ground truth for this evaluation.

d) Pick three misclassified digits and reconstruct the images as described in the readme file. Show these three images and their predictive probabilities.

e) Pick the three most ambiguous predictions, i.e., the digits whose predictive probabilities are the closest to 0.5. Reconstruct the three images as described in the readme file and show them and their predictive probabilities.

f) Treat the vector w_t as if it were a digit and reconstruct it as an image for $t = 1, 5, 10, 25, 50, 100$. Show these images and comment on what you observe.