Probabilistic Machine Learning (CS772A)
Homework 4 (Due date: Nov 19, 2017, 11:59pm)

Instructions

- We will only accept electronic submissions and the main writeup must be as a PDF file. If you are handwriting your solutions, please scan the hard-copy and convert it into PDF. Your name and roll number should be clearly written at the top. In case you are submitting multiple files, all files must to be zipped and **submitted as a single file** (named: your-roll-number.zip). Please do not email us your submissions. Your submissions have to be uploaded at the following link: `https://tinyurl.com/yagzyd67`).
- Each late submission will receive a 10% penalty per day for up to 3 days. No submissions will be accepted after the 3rd late day.

Problem 1 (25 marks)

(Mean-field VB Inference for Univariate Gaussian) Assume N observations $\{x_1, \ldots, x_N\}$ drawn from a univariate Gaussian $\mathcal{N}(x|\mu, \tau^{-1})$ where μ denotes the mean of the Gaussian and τ denotes the precision. Assume the following priors on these parameters: $p(\mu) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1})$ and $p(\tau) = \text{Gamma}(\tau|a_0, b_0)$. Note that the precision of the Gaussian prior on μ depends on τ (which is the precision of the Gaussian likelihood).

Using the mean-field assumption $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$ with $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N)$ and $q_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$, write down the evidence lower bound (ELBO) $\mathcal{L}(q)$ for this model. Now by taking the derivatives of the ELBO with respect to each of the variational parameters $(\mu_N, \lambda_N, a_N, b_N)$, estimate $q_\mu(\mu)$ and $q_\tau(\tau)$.

Problem 2 (10+15 marks)

Part 1: GP Posterior

When discussing about GP regression, we saw that we can bypass the computation of GP posterior and can directly compute the posterior predictive $p(y_*|\boldsymbol{y})$ for a new input $\boldsymbol{x}_*$. Suppose we do care about the GP posterior and would like to derive its expression, given training data $(\mathbf{X}, \boldsymbol{y}) = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$.

Assume a zero mean GP prior $p(f) = \mathcal{GP}(0, \kappa)$ which, from the GP definition, is equivalent to $p(\boldsymbol{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$ where $\boldsymbol{f} = [f(\boldsymbol{x}_1,), \ldots, f(\boldsymbol{x}_N)]^\top$ is an $N \times 1$ vector and $\mathbf{K}$ is the $N \times N$ kernel matrix with $K_{nm} = \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m)$. Assume a likelihood model $p(y_n|\boldsymbol{x}_n, f) = \mathcal{N}(y_n|f(\boldsymbol{x}_n), \sigma^2)$, where $f \sim \mathcal{GP}(0, \kappa)$. Derive the expression for the GP posterior, i.e., $p(\boldsymbol{f}|\boldsymbol{y})$. You are free to use standard results for Gaussians.

Part 2: Visualizing GP Priors and Posteriors for Regression

Assume a GP prior $\mathcal{GP}(0, \kappa)$ where κ is the squared exponential (SE) kernel $\kappa(x, x') = \rho^2 \exp\left(-\frac{(x-x')^2}{\ell^2}\right)$. Note that this is the scalar input version of the SE kernel, and it can be generalized to the vector input case by replacing $(x - x')^2$ by $(\boldsymbol{x} - \boldsymbol{x}')^\top(\boldsymbol{x} - \boldsymbol{x}')$. Assume $\rho^2 = 1$.

Our data will consist of scalar inputs and will be generated using the model $y = \sin(x) + \epsilon_n$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = 0.05$. We will generate $N = 100$ uniformly spaced inputs $x_1, \ldots, x_N$ in the interval $[0, 4\pi]$ and generate the corresponding outputs $y_1, \ldots, y_N$ from the above model.

For each of the following 5 values of ℓ from $[0.2, 0.5, 1, 2, 10]$, your task will be the following

- Draw a random sample from the GP prior $p(\boldsymbol{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$ and plot it.
- Plot the *mean* of the GP posterior (from Part 1), again on the same figure but with a different color.
- On the same plot, also show the true function ($\sin(x)$) evaluated at the generated inputs (use a different color for this too). Note that this curve will be the same in all the 5 cases.

You have been provided some MATLAB skeleton code for this problem. The main file is `gp_demo.m`. I have also provided code to sample from a multivariate Gaussian and code to compute the SE kernel. You may use these or use your own code. Note that when computing the kernel matrix $\mathbf{K}$, you may need to add a small positive number to the diagonal entries to make it invertible (see the skeleton code `gp_demo.m`).

You need to submit the codes as well as plots generated. What difference do you see between the plots generated using $\ell = [0.2, 0.5, 1, 2, 10]$, in particular w.r.t. shapes of prior/posterior of GP vs the true function.

Problem 3 (20 marks)

(Adding Label Information to Topic Models) The Latent Dirichlet Allocation (LDA) model we saw in the class was an unsupervised model. Suppose that for each document d, we also have a binary label y_d (say, denoting the positive/negative sentiment if the documents are about product reviews).

Suggest at least two ways to use the label information by extending the standard LDA model. You only need to describe the generative model and do not have to derive the inference algorithm for the model.

Problem 4 (30 marks)

(Modeling Graphs) Suppose we are given a graph $\mathbf{A} \in \{0,1\}^{N \times N}$ in form of an adjacency matrix between N entities (say N individuals in a friendship network). Assume there are K "communities" of entities (akin to K clusters) and entity n's membership is denoted by a categorical latent variable $\boldsymbol{z}_n$. Let's assume that the probability of a connection between an entity n of community k and entity m from community ℓ is equal to $\eta_{k\ell} \in (0,1)$, i.e., $p(A_{nm} = 1 | \boldsymbol{z}_n = k, \boldsymbol{z}_m = \ell) = \eta_{k\ell}$. The overall generative model can be written as follows

$$
\begin{aligned}
\eta_{k\ell} &\sim \text{Beta}(a, b), \quad k = 1, \ldots, K, \ell = 1, \ldots, K \\
\boldsymbol{z}_n &\sim \text{multinoulli}(\pi_1, \ldots, \pi_K), \quad n = 1, \ldots, N \\
A_{nm} &\sim \text{Bernoulli}(\eta_{\boldsymbol{z}_n, \boldsymbol{z}_m})
\end{aligned}
$$

Also assume a Dirichlet$(\alpha, \ldots, \alpha)$ prior on $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$. Denote $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$.

- Derive the expression of $p(\mathbf{Z})$. Note that $p(\mathbf{Z})$ denotes the marginal prior of $\mathbf{Z}$ and must not depend on $\boldsymbol{\pi}$.
- Show that the marginal likelihood $p(\mathbf{A}|\mathbf{Z})$ of the network $\mathbf{A}$ as a function of $\mathbf{Z}$, with connection probabilities $\eta_{k\ell}$'s integrated out can be written as a product of K^2 ratios of beta functions. Derive and write down the final expression for $p(\mathbf{A}|\mathbf{Z})$ with each ratio term (of two beta functions) clearly shown.