Name: _____

Roll No.: _____ **Dept.**: _____

IIT Kanpur
CS772A (PML)
Quiz-2
Date: February 15, 2024

Instructions:

Total: **24 marks**

1. Please write your name, roll number, department on **all pages** of this question paper.
2. Write your answers clearly in the provided box. Keep your answer precise and concise.

Section 1 (8 short answer questions: 2+2+2+4+2+2+2+2+2+4 = 25 marks). .

1. Given data $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ for which the likelihood for each observation is an exponential family distribution, we can write the overall likelihood as $p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$. If we choose a prior $p(\theta|\tau_0, \nu_0) \propto \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, briefly describe the roles played by τ_0 and ν_0 when estimating θ?

   The prior's hyperparameter τ_0 represents the total sufficient statistics of the pseudo-observations and ν_0 represents the number of pseudo-observations.

2. What makes a model $p(y|\boldsymbol{x})$ a Generalized Linear Model (GLM)? Assuming other hyperparameters to be known, can the posterior distribution for the GLM's parameters be computed exactly for all GLMs?

   For a GLM, $p(y|\boldsymbol{x})$ is defined by an exponential family distribution whose natural parameters are defined by a linear model $\boldsymbol{w}^\top \boldsymbol{x}$. The posterior of GLM's parameters, which is $\boldsymbol{w}$ in this case, is not always computable exactly because the GLM likelihood $p(y|\boldsymbol{x})$ is not always conjugate to the prior on $\boldsymbol{w}$ (recall logistic regression).

3. Briefly explain (though precise equations) the difference between computing the posterior predictive distribution (PPD) for a single model m, and computing the PPD when we don't have one but a collection of models $m = 1, 2, \ldots, M$. You may answer it for a supervised or unsupervised learning setting.

   Consider the supervised learning setting. Here, the PPD for a single model m can be expressed as $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, m) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, m) p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, m) d\boldsymbol{w}$ which is averaging over the posterior of the weights $\boldsymbol{w}$. When we have multiple models $m = 1, 2, \ldots, M$, the PPD requires a second level of averaging over the posterior over the model choices as $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}) = \sum_{m=1}^{M} p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, m) p(m|\mathbf{X}, \boldsymbol{y})$.

4. Consider the linear regression model $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N)$ with priors $p(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|0, \lambda^{-1}\mathbf{I})$, $p(\lambda|a, b) = \text{Gamma}(\lambda|a, b), p(\beta|c, d) = \text{Gamma}(\beta|c, d)$. Assume a, b, c, d as known. Using the decomposition $p(\boldsymbol{y}, \boldsymbol{w}, \lambda, \beta|\mathbf{X}, a, b, c, d) = p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta) p(\boldsymbol{w}|\lambda) p(\lambda|a, b) p(\beta|c, d)$, derive CPs for λ and β. Only show key steps and final expressions. Note: For a non-negative scalar random variable x, $\text{Gamma}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$ and for an r.v. $\boldsymbol{x} \in \mathbb{R}^K$, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-K/2} |\boldsymbol{\Sigma}|^{-1/2} \exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})]$.

   To get the CP of λ, in the joint $p(\boldsymbol{y}, \boldsymbol{w}, \lambda, \beta|\mathbf{X}, a, b, c, d)$, we only need the terms that contain λ. Thus $p(\lambda|\boldsymbol{w}, a, b) \propto p(\boldsymbol{w}|\lambda) p(\lambda|a, b)$ which is a product of Gaussian and gamma. Because of conjugacy of this likelihood and prior pair, the posterior will be gamma and it is easy to see that it will be $p(\lambda|\boldsymbol{w}, a, b) = \text{gamma}(\lambda|a + \frac{D}{2}, b + \frac{\boldsymbol{w}^\top \boldsymbol{w}}{2})$. Likewise, to get the posterior for β, we only need the terms that contain β. Thus $p(\beta|\boldsymbol{w}, \mathbf{X}, \boldsymbol{y}, c, d) \propto p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta) p(\beta|c, d)$. Again, due to conjugacy, the posterior is gamma and it is easy to see that it will be $p(\beta|\boldsymbol{w}, \mathbf{X}, \boldsymbol{y}, c, d) = \text{gamma}(\beta|c + \frac{N}{2}, d + \frac{||\boldsymbol{y} - \mathbf{X}\boldsymbol{w}||^2}{2})$.

5. For a latent variable model of the form $p(\mathcal{D}|\mathbf{Z}, \Theta)$ where $\mathcal{D}$ denotes the data, $\mathbf{Z}$ denotes the latent variables, and Θ denotes the parameters, Expectation Maximization (EM) computes the point estimate of Θ by maximizing an objective of the form $\mathbb{E}_A[B]$. Name the quantities A and B and give their expressions.

   A is the CP of $\mathbf{Z}$, i.e., $p(\mathbf{Z}|\mathcal{D}, \Theta)$ and B is the complete data log-likelihood, i.e., $\log p(\mathcal{D}, \mathbf{Z}|\Theta)$.

6. Is GP based supervised learning a generative model for supervised learning? Briefly justify your answer.

   No, because GP only models the function relationship between input and output via a conditional distribution of output given input, and does not model the inputs. It is therefore a discriminative model.

7. Give two distinct advantages of GP based regression over a kernel method based regression method.

   (1) GP provides an estimate of the variance in the predictions. (2) When using GP, we can learn the optimal values of the hyperparameters of the kernel without using cross-validation.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-2
Date: February 15, 2024

8. Why estimating the hyperparameters by directly maximizing the marginal likelihood may not necessarily give the same solution as the one that we would get when using the EM algorithm?

   Sorry; this question was not precise and I apologize for the ambiguity. Upon convergence, they will yield the same solution (assuming careful initialization). I meant to ask if they yield similar mathematical expressions for the parameter updates (in EM, we get closed form expressions, which is one of the appealing aspects of EM, whereas for direct MLE-II, we usually don't get closed form solution). However, it didn't come across clearly in the question. We will treat all reasonable explanations for the question as correct.

9. Ignoring computational bottlenecks, will approximating a posterior $p(\mathbf{Z}|\mathcal{D})$ using Laplace's approximation and approximating it using VI with $q(\mathbf{Z}|\phi) = \mathcal{N}(\mathbf{Z}|\mu, \Sigma)$ give the same answer? Justify your answer briefly.

   No, they won't in general because they use very different methods for computing the approximate posterior. Laplace's approximation fits a Gaussian around the MAP estimate, whereas VI explicitly maximizes the ELBO, which is equivalent to minimizing the KL divergence between the true posterior and the proposed approximation $q(\mathbf{Z}|\phi) = \mathcal{N}(\mathbf{Z}|\mu, \Sigma)$. Consequently, both approaches may result in different solutions.

10. Consider a generative binary classification model with class-conditionals distributions $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_+, \mathbf{I})$ and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_-, \mathbf{I})$. Assume class marginals $p(y = +1) = p(y = -1) = 0.5$. Denoting the model's predictions as $y_* = \text{sign}[f(\boldsymbol{x}_*)]$, give the expression for $f(\boldsymbol{x})$. Based on the expression of $f(\boldsymbol{x}_*)$, is f a linear model?

    Note that we can predict $y_* = +1$ if $\mathcal{N}(\boldsymbol{x}_*|\boldsymbol{\mu}_+, \mathbf{I}) > \mathcal{N}(\boldsymbol{x}_*|\boldsymbol{\mu}_-, \mathbf{I})$, and $y_* = -1$, otherwise.

    Comparing the PDFs is equivalent to comparing their logs, which means

$$\begin{aligned} y_* &= \text{sign}[\log \mathcal{N}(\boldsymbol{x}_*|\boldsymbol{\mu}_+, \mathbf{I}) - \log \mathcal{N}(\boldsymbol{x}_*|\boldsymbol{\mu}_-, \mathbf{I})] \\ &= \text{sign}[-(\boldsymbol{x}_* - \mu_+)^\top(\boldsymbol{x}_* - \mu_+) + (\boldsymbol{x}_* - \mu_-)^\top(\boldsymbol{x}_* - \mu_-)] \\ &= \text{sign}[2\mu_+^\top \boldsymbol{x}_* - 2\mu_-^\top \boldsymbol{x}_* + ||\mu_-||^2 - ||\mu_+||^2] \\ &= \text{sign}[(\mu_+ - \mu_-)^\top \boldsymbol{x}_* + b] \quad \text{(ignoring the factors of 2)} \end{aligned}$$

    Thus we can write $f(\boldsymbol{x})$ as $f(\boldsymbol{x}) = (\mu_+ - \mu_-)\boldsymbol{w}^\top \boldsymbol{x}_* + b$ where $\boldsymbol{w} = \mu_+ - \mu_-$ and $b = ||\mu_-||^2 - ||\mu_+||$ which has the form of a linear model.