Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-1
Date: January 29, 2024

Instructions:

Total: **25 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 (12 short answer questions: 11 x 2 + 1 x 3 = 25 marks). .

1. Given likelihood $p(x|\theta)$ and prior $p(\theta|\lambda)$, start with the joint distribution of x and θ and show the basic steps for how the marginal likelihood $p(x|\lambda)$ is obtained using the product and sum rules of probability.

   The marginal likelihood is $p(x|\lambda) = \int p(x, \theta|\lambda)d\theta = \int p(x|\theta)p(\theta|\lambda)d\theta$. The integral refers to the sum rule to integrate out θ and the factorization $p(x, \theta|\lambda) = p(x|\theta)p(\theta|\lambda)$ is the product rule.

2. Epistemic uncertainty refers to the uncertainty in the model parameters; e.g., the variance of the posterior $p(\theta|\mathcal{D})$. Can epistemic uncertainty be reduced? If yes, how? If no, why not?

   Yes, the epistemic uncertainty is because of not having enough training data and can be reduced by increasing the training data.

3. We can use MLE-II to find optimal value of a hyperparameter λ as $\hat{\lambda} = \text{argmax}_\lambda \log p(\mathcal{D}|\lambda)$. If you suspect overfitting when using $\hat{\lambda}$, suggest a way to prevent overfitting (not allowed to use more training data).

   We can add a regularizer for λ (or introduce a prior $p(\lambda)$ which is equivalent to using a regularizer) which gives rise to a MAP-II procedure: $\hat{\lambda} = \text{argmax}_\lambda \log p(\mathcal{D}|\lambda) + \log p(\lambda)$

4. Briefly explain how the posterior predictive distribution $p(y_*|\boldsymbol{y})$ of a test observation y_*, given training observations $\boldsymbol{y} = \{y_i\}_{i=1}^N$, can be written as a ratio of two quantities without needing the posterior distribution of the model parameters. What are these quantities?

   The PPD $p(y_*|\boldsymbol{y})$ can be written as the ratio of two marginal distribution: $p(y_*|\boldsymbol{y}) = \frac{p(y_*, \boldsymbol{y})}{p(\boldsymbol{y})}$ where the numerator is the marginal likelihood of training and test data combined, and the denominator is the marginal likelihood of just the training data.

5. Briefly explain why the posterior $p(\theta|\mathcal{D})$ of parameters θ given data $\mathcal{D}$ can be thought of as an ensemble.

   We can generate M samples $\{\theta^{(i)}\}_{i=1}^M$ drawn i.i.d. from the posterior $p(\theta|\mathcal{D})$. This set of parameters $\{\theta^{(i)}\}_{i=1}^M$ represents an ensemble of size M.

6. Consider a posterior predictive distribution $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}) = \int p(y_*|\boldsymbol{w}, \boldsymbol{x}_*)p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})d\boldsymbol{w}$. Briefly explain the role of multiplying $p(y_*|\boldsymbol{w}, \boldsymbol{x}_*)$ by $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})$, and intuitively why it is the right thing to do?

   The integral says that we are using all possible values of the weights $\boldsymbol{w}$ to make the prediction $p(y_*|\boldsymbol{w}, \boldsymbol{x}_*)$ and taking the average. However, we are weighing each prediction $p(y_*|\boldsymbol{w}, \boldsymbol{x}_*)$ by $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})$. The latter represents the "goodness" or importance of $\boldsymbol{w}$ under the posterior. All possible weights contribute but good weights contribute more and not-so-good weights contribute less. Thus the PPD is an importance weighted average prediction.

7. Given N observations $\boldsymbol{y} = \{y_i\}_{i=1}^N$ from a univariate Gaussian $\mathcal{N}(y|\mu, \sigma^2)$ with its variance σ^2 parameter known, suppose the posterior distribution of the Gaussian's mean is $p(\mu|\boldsymbol{y}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$. Derive the posterior predictive distribution $p(y_*|\boldsymbol{y})$ of a new observation y_* without integrating over the posterior.

   We can write y_* as $y_* = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma_N^2)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Thus y_* is the sum of two independent Gaussian random variables μ and ϵ and therefore $p(y_*|\boldsymbol{y}) = \mathcal{N}(y_*|\mu_N, \sigma_N^2 + \sigma^2)$.

8. Given a test input $\boldsymbol{x}_*$, write the expression for Monte Carlo approximation of $p(y_* = 1|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ for the logistic regression model for binary classification. Here $(\mathbf{X}, \boldsymbol{y})$ denotes the training data and assume that we have sampled M i.i.d. weight vectors $\{\boldsymbol{w}^{(i)}\}_{i=1}^M$ from the posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})$.

   $p(y_* = 1|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}) = \frac{1}{M}\sum_{i=1}^M p(y_* = 1|\boldsymbol{x}_*, \boldsymbol{w}^{(i)}) = \frac{1}{M}\sum_{i=1}^M \sigma({\boldsymbol{w}^{(i)}}^\top \boldsymbol{x}_*)$

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-1
Date: January 29, 2024

---

9. Consider a model with weights $\boldsymbol{w} \in \mathbb{R}^D$ having prior $\mathcal{N}(\boldsymbol{w}|\mathbf{0}, \boldsymbol{\Sigma})$. Consider 2 cases for $\boldsymbol{\Sigma}$: (1) diagonal covariance matrix, and (2) spherical covariance matrix. Briefly state what's the advantage of (1) over (2)?
Using (1) is equivalent to using a different variance/precision hyperparameter for the Gaussian prior $p(w_d) = \mathcal{N}(w_d|0, \sigma_d^2)$ each component w_d of the vector $\boldsymbol{w}$ and thus imposes different amount of regularization on each feature of the input. This is more advantageous as compared to option (2) which uses a spherical covariance $\sigma^2\mathbf{I}$ and therefore regularizes each feature equally.

10. For a linear regression model with Gaussian likelihood and Gaussian prior over the weights, mention two advantages of the PPD $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ over the plug-in predictive $p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MAP})$.
(1) PPD is more robust because it doesn't rely on a single best estimate of the weight vector $\boldsymbol{w}$ to make the prediction; (2) PPD for the linear regression has a test input-dependent variance whereas the plug-in predictive has the same variance for all test inputs.

11. Briefly describe what is Laplace's approximation and what is the main computational bottleneck when using Laplace's approximation?
Laplace's approximation approximates an intractable posterior $p(\theta|\mathcal{D})$ using a Gaussian whose mean is the MAP estimate of θ and whose precision matrix is the negative of the Hessian of the log-posterior. It's main computational bottleneck is the computation of the inverse of the Hessian (e.g., when doing operations such as sampling from the posterior)

12. Show that the expectation $\mathbb{E}_{p(y|\theta)}[s(\theta)]$ of the score function $s(\theta) = \nabla_\theta \log p(y|\theta)$ is equal to zero.
$\mathbb{E}_{p(y|\theta)}[s(\theta)] = \int \nabla_\theta \log p(y|\theta) p(y|\theta) dy = \int \frac{\nabla_\theta p(y|\theta)}{p(y|\theta)} p(y|\theta) dy = \nabla_\theta \int p(y|\theta) dy = \nabla_\theta 1 = 0$