Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-1
Date: February 2, 2022

Instructions:

Total: **30 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 (10 very short answer questions: 10 x 3 = 30 marks). .

1. Suppose we know that the weight vector $\boldsymbol{w} \in \mathbb{R}^D$ for a regression problem is close to some value $\boldsymbol{w}_0 \in \mathbb{R}^D$. Suggest a prior distribution that incorporates this assumption and mention its parameters. Also write down the expression for the regularizer this prior corresponds to.

   Can use the following Gaussian prior: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_0, \lambda^{-1}\mathbf{I})$. Note that we may also use a Gaussian prior with full covariance matrix Σ (although spherical covariance , $\lambda^{-1}\mathbf{I}$ usually suffices unless you have some strong prior belief about the weights being correlated). This Gaussian prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_0, \lambda^{-1}\mathbf{I})$ corresponds to the following regularlizer: $\frac{\lambda}{2}(\boldsymbol{w} - \boldsymbol{w}_0)^\top(\boldsymbol{w} - \boldsymbol{w}_0)$.

2. Why variance of the posterior predictive is larger than the variance of a plug-in predictive?

   It is because the posterior predictive takes into account the full posterior distribution of the parameters, thereby incorporating the uncertainty (i.e., variance) of the parameters, whereas plug-in predictive only uses a point estimate of the parameters when making the prediction and thus ignores the parameter uncertainty.

3. Write down, in terms of the marginalization based definition, the general expression for the posterior predictive distribution (PPD) of a supervised learning model with likelihood of the form $p(y|\boldsymbol{x}, \boldsymbol{w})$ where $\boldsymbol{w}$ denotes the model weights, training data is $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, and the test input is $\boldsymbol{x}_*$.

$$p(y_*|\boldsymbol{x}_*, \mathcal{D}) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w})p(\boldsymbol{w}|\mathcal{D})d\boldsymbol{w}$$

4. Suppose the above PPD is intractable. Write down the expression for its Monte-Carlo approximation.

   Assuming we have drawn M i.i.d. samples of $\boldsymbol{w}$ from the posterior $p(\boldsymbol{w}|\mathcal{D})$ and denoting them as $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(M)}$, the Monte-Carlo approximation can be written as $p(y_*|\boldsymbol{x}_*, \mathcal{D}) \approx \frac{1}{M}\sum_{i=1}^M p(y_*|\boldsymbol{x}_*, \boldsymbol{w}^{(i)})$.

5. For a model with likelihood being an exponential family distribution $p(x|\theta) = h(x)\exp[\theta^\top\phi(x) - A(\theta)]$, will the MLE solution for θ be unique? Briefly justify your answer.

   The log-likelihood, ignoring the terms that don't depend on θ, will be of the form $\theta^\top\phi(x) - A(\theta)$ Since $A(\theta)$ is convex and $\theta^\top\phi(x)$ is linear, the log-likelihood will be a concave function and thus will have a unique maxima (likewise, the NLL will be a convex function and will have a unique minima).

6. MAP estimation and fully Bayesian inference both incorporate the prior distribution over the model parameters. What additional benefits does the fully Bayesian approach offer when it comes to making predictions for test data?

   Both incorporate the prior distribution and thus both combat overfitting (recall that MAP performs regularization). However, the fully Bayesian approach also takes into account the uncertainty in the parameters, so instead of relying on a single best estimate, it computes a posterior weighted average prediction. This yields more robust predictions as compared to MAP based prediction.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-1
Date: February 2, 2022

7. Taking the example of a model with two unknowns θ_1 and θ_2, and training data $\mathcal{D}$, briefly explain the difference between joint posterior, conditional posterior, and marginal posterior.

   Joint posterior ($p(\theta_1, \theta_2|\mathcal{D})$ in this case) is the joint distribution of all unknowns conditioned on the training data. Conditional posterior is the distribution of one (or some) unknowns conditioned on the data *and* the rest of the unknowns (e.g., $p(\theta_1|\theta_2, \mathcal{D})$ and $p(\theta_1|\theta_2, \mathcal{D})$ in this case). Marginal posterior is the distribution of each unknown conditioned on the data and all other unknowns integrated out (e.g., $p(\theta_1|\mathcal{D})$ and $p(\theta_2|\mathcal{D})$ in this case).

8. Consider two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for some data $\mathbf{X}$. Suppose you want to decide which of the two is the better model by (1) Comparing their posterior probabilities; and (2) Comparing their marginal likelihoods. Will (1) and (2) give the same result? Briefly justify your answer (maximum 1-2 sentences).
   Comparing posterior probabilities of the two models means comparing $p(\mathcal{M}_1|\mathbf{X}) \propto p(\mathcal{M}_1)p(\mathbf{X}|\mathcal{M}_1)$ and $p(\mathcal{M}_2|\mathbf{X}) \propto p(\mathcal{M}_2)p(\mathbf{X}|\mathcal{M}_2)$. Comparing the marginal likelihoods of the two models means comparing $p(\mathbf{X}|\mathcal{M}_1)$ and $p(\mathbf{X}|\mathcal{M}_2)$, which ignores the prior probabilities of choosing the models, i.e., $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$.

9. Briefly explain what is MLE-II (maximum 1-2 sentences) and what is it used for? What is its advantage over cross-validation?
   MLE-II is the approach for estimating the best value of the hyperparameters by maximizing the marginal likelihood, which is the probability of the data as a function of the hyperparameters. The advantage of MLE-II over cross-validation is that it does not require a separate validation-set to find the best value of the hyperparameters because the marginal likelihood is computed using the training data itself.
10. Given an already computed posterior distribution $p(\theta|\mathcal{D})$, assuming it is some known parametric distribution (e.g., a Beta, or a Gaussian), can you obtain the MAP estimate of θ from this posterior? Can you obtain MLE of θ from this posterior? Briefly justify your answers.
    We can obtain the MAP estimate from the posterior easily since it is simply the mode of the posterior. MLE can't be obtained easily because it does not correspond to any property of the posterior (e.g., mean, median, mode, etc).