Danmarks Tekniske Universitet

Written examination date: 16/8/2024

Pages: 7 (including this front page)

Course title: Bayesian Machine Learning

Course number: 02477

Aids allowed: All aids except internet

Exam duration: 4 hours

Weighting: 100%

02477 Bayesian Machine Learning Exam 2024R

Technical University of Denmark

- **Duration**: 4 hours
- **Aids**: All aids except internet
- **Student number**: Make sure you student number is visible on all pages.
- **Results:** Report all numeric results with 2 digits after the decimal point.
- **Explain** how you arrived at your results, document intermediate results when possible.
- **Hand-in**: Your solution must be handed in digitally as a PDF.

Contents

Part 1: Multi-class classification

Consider the following linear model for multi-class classification with $K = 3$ classes:

$$y_n|\mathbf{f}_n \sim \text{Categorical}\left[\text{softmax}(\mathbf{f}_n)\right], \tag{1}$$

$$\mathbf{f}_n = \mathbf{W}\phi(x_n), \tag{2}$$

$$\mathbf{W}_{ij} \sim \mathcal{N}(0, \alpha^{-1}), \tag{3}$$

where $y_n \in \{1, 2, 3\}$, $x_n \in \mathbb{R}$, $\alpha > 0$ is a hyperparameter, and $\boldsymbol{W}$ are the parameters of interest. The feature transformation $\phi(x)$ is given by $\phi(x) = \begin{bmatrix}1 & x\end{bmatrix}^T$ such that $\boldsymbol{W} \in \mathbb{R}^{K \times D}$ for $D = 2$.

Question 1.1: Identify the prior and likelihood of the model.

Let

$$\hat{\boldsymbol{W}}_{\text{MAP}} = \begin{bmatrix} -0.5 & -2.0 \\ 3.0 & 0.0 \\ 1.0 & 1. \end{bmatrix} \tag{4}$$

be a MAP-estimator for the model given in eq. (1)-(3) for some dataset $\mathcal{D}$ (not given).

Question 1.2: Use the plugin approximation with $\hat{\boldsymbol{W}}_{\text{MAP}}$ to compute the posterior predictive distribution for $x^* = -1$.

Let $\boldsymbol{W}^{(i)} \sim q(\boldsymbol{W})$ for $i = 1, 2, 3$ be samples from a variational approximation of the posterior, i.e. $p(\boldsymbol{W}|\mathcal{D}) \approx q(\boldsymbol{W})$:

$$\boldsymbol{W}^{(1)} = \begin{bmatrix} -0.15 & -1.92 \\ 3.2 & 0.45 \\ 1.37 & 0.8 \end{bmatrix}, \quad \boldsymbol{W}^{(2)} = \begin{bmatrix} -0.31 & -2.03 \\ 2.98 & 0.08 \\ 1.03 & 1.29 \end{bmatrix}, \quad \boldsymbol{W}^{(3)} = \begin{bmatrix} -0.35 & -1.98 \\ 3.09 & 0.07 \\ 1.3 & 0.96 \end{bmatrix}. \tag{5}$$

Question 1.3: Compute a Monte Carlo estimate of the posterior predictive distribution for $x^* = -1$ using samples given above.

The predictive distribution $p(y^*|\mathcal{D}, x^* = 3)$ is given in the table below:

k	$p(y^* = k\|\boldsymbol{y}, x^*)$
1	0.00
2	0.27
3	0.73

Question 1.4: Determine the entropy and confidence of the posterior predictive distribution for $\boldsymbol{x}^* = 3$ given in the table above.

Question 1.5: Suppose the value of the hyperparameter α is increased by a factor of 10. Explain in your own words how you would expect the MAP-estimate to change and why.

Part 2: Gaussian process regression

Let $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ be a dataset for regression, where $x_n \in \mathbb{R}$ and $y_n \in \mathbb{R}$ are the input and output for the n'th observation, respectively:

$$\boldsymbol{x} = \begin{bmatrix} -2.00 & 0.00 & 2.00 \end{bmatrix}$$
$$\boldsymbol{y} = \begin{bmatrix} -2.01 & 1.41 & 0.23 \end{bmatrix},$$

such that x_n and y_n are the n'th elements in $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, for $N = 3$.

Assume a Gaussian process regression model of the form

$$y_n = f(x_n) + \epsilon_n, \tag{6}$$

where $f \sim \mathcal{GP}(0, k_1(x, x'))$ and $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ is i.i.d additive Gaussian noise. Assume the following *squared exponential* kernel:

$$k_1(x, x') = 2 \exp\left(-\frac{1}{8}||x - x'||_2^2\right) \tag{7}$$

and assume the standard deviation of the noise is given by $\sigma = \frac{1}{2}$.

Question 2.1: Determine the value of the magnitude and lengthscale hyperparameters for the kernel k_1 in eq. (7).

Let $\boldsymbol{f} \in \mathbb{R}^N$ denote a vector containing the values of the function f evaluated at the training points, i.e. $\boldsymbol{f} = \begin{bmatrix} f(x_1) & f(x_2) & \dots & f(x_N) \end{bmatrix}$.

Question 2.2: Determine the analytical prior distribution $p(\boldsymbol{f}|\boldsymbol{x})$.

Question 2.3: Determine the analytical posterior distribution $p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{x})$.

Now consider a different kernel:

$$k_2(x, x') = \exp\left(-\frac{1}{2}||x - x'||_2\right) + 2. \tag{8}$$

Question 2.4: Determine the analytical prior variance of $f(x) \sim \mathcal{GP}(0, k_2(x, x'))$ for $x \in \mathbb{R}$ given by the kernel in eq. (8).

Part 3: A model with a mixture prior distribution model

Consider the following probabilistic model for an observed variable $\boldsymbol{y} \in \mathbb{R}^2$ and parameters of interest $\boldsymbol{\theta} \in \mathbb{R}^2$:

$$p(\boldsymbol{\theta}) = \frac{1}{2}\mathcal{N}(\boldsymbol{\theta}|-\boldsymbol{m}, \tau^2\boldsymbol{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}, \tau^2\boldsymbol{I}) \tag{9}$$

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I}), \tag{10}$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0.5 \\ -1 & 1 \end{bmatrix}, \qquad \boldsymbol{m} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \tau^2 = \sigma^2 = 1, \tag{11}$$

are constants and $\boldsymbol{I} \in \mathbb{R}^{2\times 2}$ is the identity matrix.

Question 3.1: Compute the maximum likelihood estimate for $\boldsymbol{\theta}$.

Question 3.2: Compute the value of the prior density for $\boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Question 3.3: Determine the analytical expression for the marginal likelihood.

If you did not solve Question 3.3, you can assume $p(\boldsymbol{y}) = 0.1$ when solving the next question.

Question 3.4: Compute the value of the posterior density for $\boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Part 4: A generalized linear model

Consider the generalized linear model

$$
\begin{aligned}
y_n | x_n, \boldsymbol{w} &\sim \text{Poisson}(\mu(x_n)),\\
\mu(x_n) &= \exp(3 + w_1 x_n + w_2 x_n^2)\\
\mathbf{w}|\alpha &\sim \mathcal{N}(0, \alpha^{-1}\mathbf{I}),
\end{aligned}
$$

where $x_n \in \mathbb{R}$ is an input with corresponding target $y_n \in \{0, 1, 2, \dots\}$, $\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ are the parameters of interest, $\alpha = 8$ is a fixed hyperparameter, and $\boldsymbol{I}$ is the identity matrix.

The plots below show the prior, the likelihood and the posterior, respectively, for the following dataset with $N = 3$ observations:

$$
\begin{aligned}
\boldsymbol{x} &= \begin{bmatrix} 1 & 2 & 3 \end{bmatrix},\\
\boldsymbol{y} &= \begin{bmatrix} 10 & 4 & 1 \end{bmatrix},
\end{aligned}
$$

such that x_n and y_n are the n'th elements in $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

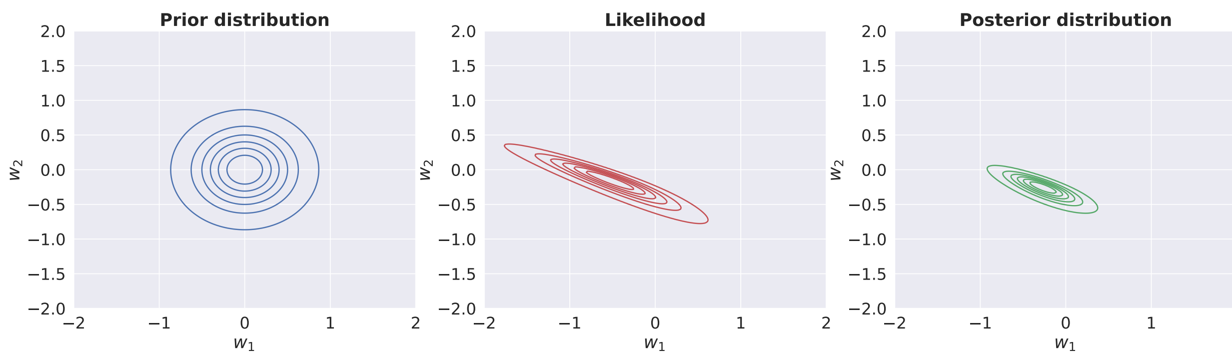

Figure 1: Contour plots of the prior, likelihood, and posterior, respectively.

Question 4.1: Use Figure 1 to visually identify (approximately) and report the maximum likelihood estimator and the MAP estimator for $\boldsymbol{w}$.

Question 4.2: Compute the prior mean of $\mu(x^*)$ for $x^* = 0$.

Question 4.3: Run a single MCMC chain using the Metropolis algorithm for 10^4 iterations using a standardized Gaussian as proposal distribution. Initialize the chain at $(w_1, w_2) = (0, 0)$. Discard 50% of the samples as warm up. Plot the resulting traces for both parameters.

Use the posterior samples of $\boldsymbol{w}$ from Question 4.4 to answer the next two questions. If you did not solve the previous question, you can draw 10^4 samples of $\boldsymbol{w}$ as

$$
\boldsymbol{w} \sim \mathcal{N}\left(\begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & -0.1 \\ -0.1 & 0.5 \end{bmatrix}\right) \tag{12}
$$

and assume these are samples from the correct posterior distribution when solving the next two questions.

Question 4.4: Estimate the posterior probability that w_1 is positive using the samples.

Consider now the test point $x^* = 1.5$.

Question 4.5: Estimate posterior probability that $\mu^* = \mu(x^*)$ is greater than 7 using the posterior samples.

Question 4.6: Compute an approximate 90% posterior credibility interval for $p(y^*|\boldsymbol{y}, x^*)$ using the posterior samples.

Part 5: A non-linear Gaussian model

Consider the following probabilistic model

$$\begin{aligned} y|w &\sim \mathcal{N}(e^w, 1) \\ w &\sim \mathcal{N}(0, 1) \end{aligned}$$

for a single observation $y \in \mathbb{R}$ and parameter $w \in \mathbb{R}$. The mode of the posterior distribution is

$$\hat{w}_{\text{MAP}} = \arg\max_w p(w|y) \approx 1.293404$$

for $y = 5$.

Question 5.1: Use ancestral sampling with $S = 1000$ to estimate the prior mean of y.

Question 5.2: Evaluate the logarithm of the joint density for $w = \hat{w}_{\text{MAP}}$ and $y = 5$.

We will now introduce a Laplace approximation of the posterior, i.e. $p(w|y) \approx q(w)$.

Question 5.3: Determine the approximate posterior mean and variance of the Laplace approximation q for $y = 5$.

Question 5.4: Use the Laplace approximation q to estimate the posterior probability of the event $w > w_{\text{MAP}}$ for $y = 5$.