Danmarks Tekniske Universitet

Written examination date: 15/5/2024

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 2024

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: Linear Gaussian systems

Consider the following Markov Chain

$$\boldsymbol{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{1}$$
$$\boldsymbol{x}_2 | \boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x_1}, \mathbf{\Sigma}) \tag{2}$$
$$\boldsymbol{x}_3 | \boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x_2}, \mathbf{\Sigma}) \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ and $\mathbf{\Sigma} \in \mathbb{R}^{2\times 2}$ are constants.

Question 1.1: Determine the conditional distribution of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$.

Question 1.2: Determine the marginal distribution of $\boldsymbol{x}_3$, i.e. $p(\boldsymbol{x}_3)$.

Question 1.3: Determine the conditional distribution of $\boldsymbol{x}_3$ given $\boldsymbol{x}_1$.

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.

Assume a Gaussian process regression model of the form

$$y_n = f(x_n) + \epsilon_n, \tag{4}$$

where $f \sim \mathcal{GP}(0, k(x, x'))$ and $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ is i.i.d additive Gaussian noise.

Assume the following kernel:

$$k_1(x, x') = \kappa^2 \exp\left(-\frac{1}{2\ell^2}||x - x'||_2^2\right) \tag{5}$$

and the following dataset with $N = 8$ observations

$$\begin{aligned} \boldsymbol{x} &= \begin{bmatrix} -2.17 & 1.99 & 0.57 & -3.01 & -1.16 & 3.30 & -4.85 & -0.86 \end{bmatrix} \\ \boldsymbol{y} &= \begin{bmatrix} 0.88 & 0.46 & -0.06 & 0.98 & 0.45 & 0.88 & -0.66 & 0.05 \end{bmatrix} \end{aligned}$$

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

Question 2.1: Determine the likelihood for the regression model in eq. (4)**.**

Assume the following values for the hyperparameters:

$$\kappa = 0.7, \quad \ell = \frac{1}{2}\sqrt{2}, \quad \sigma = \frac{1}{5}. \tag{6}$$

Question 2.2: Determine the analytical prior predictive distribution $p(y^*|x^* = 1)$ **for** $y^* = y(x^*)$**.**

Question 2.3: Determine the analytical posterior predictive distribution $p(y^*|\boldsymbol{y}, x^* = 227)$ **for** $y^* = y(x^*)$**.**

Question 2.4: Determine the analytical expression for the marginal likelihood $p(\boldsymbol{y}|\kappa, \ell, \sigma)$ **and compute the numerical value of** $\log p(\boldsymbol{y}|\kappa, \ell, \sigma)$**.**

Question 2.5: Determine the analytical posterior distribution $p(f^*|\boldsymbol{y}, x^* = 1)$ **for** $f^* = f(x^*)$ **as well as the analytical posterior predictive distribution for** $p(y^*|\boldsymbol{y}, x^* = 1)$ **for** $y^* = y(x^*)$**.**

Consider now a different kernel

$$k_2(x, x') = c_1 \left(1 + \frac{||x - x'||}{2\ell^2}\right)^{-1} + c_2 x x',$$

with hyperparameters $c_1, c_2, \ell > 0$.

Question 2.6: Determine whether k_2 **is a stationary kernel and determine whether** k_2 **is an isotropic kernel.**

Question 2.7: Compute the prior covariance between $\boldsymbol{f} = [f(x_1), f(x_2), \ldots, f(x_8)]$ **and** $f^* = f(x^*)$ **for** $x^* = -1$ **for** $c_1 = c_2 = 1$ **and** $\ell = \frac{1}{\sqrt{2}}$**.**

Part 3: Binary classification

Consider the following generalized linear model for binary classification:

$$\begin{aligned} y_n|\boldsymbol{w}, x_n &\sim \text{Ber}(\sigma(f(x_n))) \\ f(x) &= w_0 + w_1 x + w_2 x^2 \\ \boldsymbol{w} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \end{aligned}$$

where $\boldsymbol{w} = \begin{bmatrix} w_0 & w_1 & w_2 \end{bmatrix} \in \mathbb{R}^3$, $\boldsymbol{I} \in \mathbb{R}^{3\times 3}$ is the identity matrix, $\sigma(\cdot)$ is the logistic sigmoid function, $x_n \in \mathbb{R}$ and $y_n \in \{0, 1\}$.

Assume $\hat{\boldsymbol{w}}_{\text{MAP}} = \begin{bmatrix} 2.647 \\ -1.688 \\ -0.596 \end{bmatrix}$ is the MAP-estimator for $\boldsymbol{w}$ given some dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, where $\boldsymbol{y} \in \mathbb{R}^N$ denotes the targets.

Question 3.1: Suppose you suspect a bug in the code for computing the MAP estimator. Explain how you could verify that $\hat{w}_{\text{MAP}}$ is indeed the correct MAP estimator.

Question 3.2: Determine the posterior predictive distribution for $p(y^*|\boldsymbol{y}, x^* = -3)$ using the plug-in approximation based on the MAP-estimator.

Consider now the following Gaussian approximation of the posterior distribution for the weights

$$p(\boldsymbol{w}|\boldsymbol{y}) \approx \mathcal{N}\left(\boldsymbol{w}|\boldsymbol{w}_{\text{MAP}}, \boldsymbol{S}\right) \quad \text{for} \quad \boldsymbol{S} = \begin{bmatrix} 3. & -0.39 & -0.3 \\ -0.39 & 1.55 & 0.37 \\ -0.3 & 0.37 & 0.14 \end{bmatrix}. \tag{7}$$

Question 3.3: Use the Gaussian approximation to compute a 90% posterior credibility interval for w_0.

Question 3.4: Use the Gaussian approximation to compute the approximate posterior distribution $p(f^*|\boldsymbol{y}, x^* = -3)$ and the approximate posterior predictive distribution for $p(y^*|\boldsymbol{y}, x^* = -3)$ using the probit approximation.

Consider the following utility matrix for a decision (i.e. the prediction) $\hat{y} \in \{0, 1\}$ and the true value $y \in \{0, 1\}$:

$\mathcal{U}(y, \hat{y})$	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	2	1
$y = 1$	1	2

Suppose the predictive posterior distribution for a specific x^* is given by $p(y^* = 1|\boldsymbol{y}, x^*) = 0.129$.

Question 3.5: Compute the expected utility for each decision $\hat{y} \in \{0, 1\}$ and determine the optimal decision wrt. the utility matrix above.

Part 4: Variational inference

Consider the following probabilistic model

$$\begin{aligned} y|w_1, w_2 &\sim \mathcal{N}(w_1 w_2, \sigma^2) \\ w_1 &\sim \mathcal{N}(0, 1) \\ w_2 &\sim \mathcal{N}(0, 1), \end{aligned}$$

for a single observation $y \in \mathbb{R}$ and parameters $w_1, w_2 \in \mathbb{R}$.

Consider a variational family $\mathcal{Q}$ consisting of distributions of the form:

$$q(w_1, w_2) = \mathcal{N}(w_1|m_1, v_1)\mathcal{N}(w_2|m_2, v_2)$$

with variational means $m_1, m_2 \in \mathbb{R}$ and variational variances $v_1, v_2 > 0$.

Assume $y = 1$ and $\sigma^2 = 1$ and assume the variational parameters are initialized as follows: $m_1 = -1$, $m_2 = 1$ and $v_1 = v_2 = 1$.

Question 4.1: Determine the analytical expression of the entropy of $q(w_1, w_2)$ and evaluate it for the initial variational parameter values given above.

The evidence lowerbound (ELBO) for this model can be written as

$$\mathcal{L}[q] = \mathbb{E}_{q(w_1,w_2)}[\log p(y|w_1, w_2)] - \mathrm{KL}[q(w_1, w_2)||p(w_1, w_2)], \tag{8}$$

where the first term is the expected log likelihood and the second term is the KL-divergence is calculated between the posterior approximation $q(w_1, w_2)$ and the prior $p(w_1, w_2)$.

Question 4.2: Determine the analytical expression for the expected log likelihood and evaluate it for the initial variational parameter values given above.

Let $q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}[q(w_1, w_2)||p(w_1, w_2|y)]$ denote the optimal variational approximation.

Question 4.3: Determine the approximate posterior covariance between w_1 and w_2 with respect to the optimal approximation q^*.

Part 5: Markov Chain Monte Carlo

Assume the posterior density for a model with parameters $z_1, z_2 \in \mathbb{R}$ and data $\mathcal{D}$ are given by

$$\log p(z_1, z_2|\mathcal{D}) = -(1 - z_1)^2 - 20(z_2 - z_1^2)^2 - z_1^2 - z_2^2 + \text{constant} \tag{9}$$

Question 5.1: Plot the contours of the posterior density for the ranges $z_1 \in [-2, 2]$ and $z_2 \in [-1, 3]$ with 100 equidistant points for both z_1 and z_2.

Let $\boldsymbol{z} = \begin{bmatrix} z_1 & z_2 \end{bmatrix}$ and consider a Metropolis sampler with an isotropic Gaussian proposal distribution $q(\boldsymbol{z}^*|\boldsymbol{z}^{(k-1)}) = \mathcal{N}(\boldsymbol{z}^*|\boldsymbol{z}^{(k-1)}, \boldsymbol{I})$ for $\boldsymbol{z}^{(k)} = \left[z_1^{(k)}, z_2^{(k)}\right]$, where $k \in \mathbb{N}$ denotes the iteration number.

Question 5.2: Run a single MCMC chain using the Metropolis algorithm for 10^4 iterations using the proposal distribution given above. Initialize the chain at $(z_1, z_2) = (0, 1.5)$. Discard 10% of the samples as warm up. Plot the resulting traces for both parameters.

Use the posterior samples of $\boldsymbol{z}$ from Question 5.2 to answer the next two questions. If you did not solve the previous question, you can draw 10^4 samples of $\boldsymbol{z}$ as

$$\boldsymbol{z} \sim \mathcal{N}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}\right) \tag{10}$$

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

Question 5.3: Estimate the posterior mean of $\sin(z_1 z_2)$ using the samples.

Question 5.4: Estimate the posterior probability $p(z_1 > z_2|\mathcal{D})$ using the samples.