02477 Bayesian Machine Learning 2024: Assignment 2

This is the second assignment out of three in the Bayesian machine learning course 2024. The assignment is a group work of 3-5 students (please use the same groups as in assignment 1 if possible) and hand in via DTU Learn). The assignment is **mandatory**. The deadline is **31st of March 23:59**.

Part 1: Objective functions for regression modelling

In this part we will study the duality between common *objective/cost/loss functions* used in supervised machine learning (e.g. sum-of-squares and cross-entropy) and their corresponding probabilistic interpretations. In the supervised learning paradigm, we often have a set of input features $\boldsymbol{x} \in \mathcal{X}$ and a set of outputs $y \in \mathcal{Y}$. For regression tasks, the input and output spaces are often taken to be $\mathcal{X} = \mathbb{R}^D$ and $\mathcal{Y} = \mathbb{R}$, respectively. Given a dataset $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, we now assume noisy observations of the form

$$y_n = f(\boldsymbol{x}_n) + \epsilon_n, \tag{1}$$

where $f(\boldsymbol{x}_n)$ is the model prediction for $\boldsymbol{x}_n$ and ϵ_n is independent and additive noise. The model $f(\boldsymbol{x}_n)$ can be any model, e.g. a linear model, neural network or Gaussian process, and the goal is to learn f using the dataset $\mathcal{D}$. Assuming the model f is parametrized by a vector $\mathbf{w} \in \mathbb{R}^D$, a very common approach for learning $\boldsymbol{w}$ is to minimize the *sum-of-squares objective function*:

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - f(\boldsymbol{x}_n))^2, \tag{2}$$

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} J(\boldsymbol{w}). \tag{3}$$

To minimize overfitting, we often add a regularization term $R(\boldsymbol{w})$ to control the complexity of the model

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - f(\boldsymbol{x}_n))^2 + \lambda \cdot R(\boldsymbol{w}). \tag{4}$$

Choosing $R(\boldsymbol{w}) = \frac{1}{2}\sum_i w_i^2$ leads to the popular *ridge regularizer* and $R(\boldsymbol{w}) = \sum_i |w_i|$ leads to the so-called *LASSO regularizer*, which is useful for promoting sparsity. In this assignment, we will show that minimizing the sum-of-squares loss function in eq. (2) is equivalent to the maximum likelihood estimator for a probabilistic model, where the likelihood is a Gaussian distribution of the form:

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(\boldsymbol{x}_n), \beta^{-1}), \tag{5}$$

where $\beta > 0$ is the *precision* of the noise. Furthermore, we will also show that minimizing the regularized objective in eq. (4) is equivalent to the MAP estimator for a probabilistic model with Gaussian likelihood and a prior distribution that depends on the specific choice of regularization term $R(\boldsymbol{w})$, e.g.

$$p(\boldsymbol{w}) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha^{-1}), \quad \alpha > 0 \tag{6}$$

$$p(\boldsymbol{w}) = \prod_{i=1}^{M} \text{Laplace}(w_i|0, b), \quad b > 0 \tag{7}$$

where the Gaussian prior corresponds to the ridge regularizer and the Laplace prior corresponds to the LASSO regularizer. The density for the Laplace distribution is given by

$$\text{Laplace}(w_i|0, b) = \frac{1}{2b}\exp\left(-\frac{|w_i|}{b}\right). \tag{8}$$

Task 1.1: Show that the maximum likelihood solution of $\boldsymbol{w}$ for eq. (5) is equivalent to the mean square error solution from eq. (2) .

Hint: First show that $\ln p(\boldsymbol{y}|\boldsymbol{w}) = -\frac{\beta}{2}\sum_{n=1}^{N}(y_n - f(\boldsymbol{x}_n))^2 + K$*, where* K *is a constant independent of the parameters* $\boldsymbol{w}$*. Next, consider how additive and multiplicative constants affect the solution to minimization problems.*

Task 1.2: Show that the MAP (maximum a posteriori) estimator for $\boldsymbol{w}$ for the Gaussian likelihood in eq. (5) with the Gaussian prior in eq. (6) is equivalent to the solution for ridge regression when the regularization parameter is $\lambda = \frac{\alpha}{\beta}$.

Hint: Write the expression for $\ln p(\boldsymbol{w}|\boldsymbol{y})$*, apply Bayes theorem, simplify, and ignore constants independent of* $\boldsymbol{w}$*.*

Task 1.3: Show that the MAP (maximum a posteriori) estimator for $\boldsymbol{w}$ for the Gaussian likelihood in eq. (5) with the Laplace prior in eq. (7) is equivalent to the solution for LASSO regression objective: sum-of-squares loss with L1-penalty for $\lambda = \frac{1}{b\beta}$.

Part 2: Objective functions for classification

We will now turn our attention to objective functions for classification. For binary classification, the input and output spaces are taken to be $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0, 1\}$, and one of the most common objective function for this task is the binary cross-entropy given by

$$J(\boldsymbol{w}) = -\sum_{n=1}^{N}\left[y_n \ln \pi(\boldsymbol{x}_n) + (1 - y_n)\ln(1 - \pi(\boldsymbol{x}_n))\right], \tag{9}$$

where $\pi(\boldsymbol{x}_n)$ denotes the predicted probability for $y_n = 1$, which is parametrized by $\boldsymbol{w}$. The loss can be generalized to multi-class problems as follows

$$J(\boldsymbol{w}) = -\sum_{n=1}^{N}\sum_{i=1}^{K} y_{n,i} \ln \pi_i(\boldsymbol{x}_n), \tag{10}$$

where K is the number of classes and $y_n \in \{0, 1\}^K$ is a binary vector such that $y_{n,i}$ is 1 if the n'th example belongs to the i'th class and zero otherwise, and $\pi_i(\boldsymbol{x}_n)$ is the predicted probability for observation n belong the i'th class.

The equivalent probabilistic counterparts for binary and multi-class classification are models where the observations are assumed to be described Bernoulli distributions or Categorical distributions, respectively. That is, for binary classification we use the Bernoulli distribution as likelihood

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{n=1}^{N} \mathrm{Ber}(y_n|\pi(\boldsymbol{x}_n)), \tag{11}$$

where the function $\pi(\boldsymbol{x}_n) = \sigma(f(\boldsymbol{x}_n))$ is typically defined using a so-called inverse link function $\sigma(\cdot) : \mathbb{R} \to (0, 1)$ for some model output $f(\boldsymbol{x}_n)$. For multi-class classification problems, we use the *Categorical distribution* as likelihood for the data

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{n=1}^{N} \mathrm{Cat}(y_n|\boldsymbol{\pi}(\boldsymbol{x}_n)), \tag{12}$$

where $\boldsymbol{\pi}(\boldsymbol{x}_n) = [\pi_1(\boldsymbol{x}_n), \pi_2(\boldsymbol{x}_n), \ldots, \pi_K(\boldsymbol{x}_n)]$ and the probability mass function of the Categorical distribution is given by

$$\mathrm{Cat}(y_n|\boldsymbol{\pi}(\boldsymbol{x}_n)) = \prod_{i=1}^{K} \pi_i(\boldsymbol{x}_n)^{y_{n,i}}, \tag{13}$$

where the function $\boldsymbol{\pi}(\boldsymbol{x}_n)$ is K-dimensional and typically defined via the softmax function, i.e. $\boldsymbol{\pi}(\boldsymbol{x}) = \text{softmax}(f(\boldsymbol{x}))$ for some model $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^K$.

Task 2.1: Show that the maximum likelihood estimator for (11) **is the same solution as the solution to the binary cross entropy in eq.** (9)**.**

Task 2.2: Show that the maximum likelihood estimator for (12) **is the same solution as the solution to the general cross entropy in eq.** (10)**.**

Part 3: Gaussian processes and covariance functions

In this part, we will study covariance functions for Gaussian process models. Consider the following six covariance functions

$$k_1(x, x') = 2\exp\left(-\frac{(x-x')^2}{2 \cdot 0.3^2}\right) \tag{14}$$

$$k_2(x, x') = \exp\left(-\frac{(x-x')^2}{2 \cdot 0.1^2}\right) \tag{15}$$

$$k_3(x, x') = 4 + 2xx' \tag{16}$$

$$k_4(x, x') = \exp\left(-2\sin\left(3\pi \cdot |x-x'|\right)^2\right) \tag{17}$$

$$k_5(x, x') = \exp\left(-2\sin\left(3\pi \cdot |x-x'|\right)^2\right) + 4xx' \tag{18}$$

$$k_6(x, x') = \frac{1}{5} + \min(x, x') \tag{19}$$

Some of them should be familiar and some of them might be new to you.

Task 3.1: Compute the marginal prior variance of a Gaussian process, $f_i(x) \sim \mathcal{GP}(0, k_i(x, x'))$, i.e. compute $\mathbb{V}[f_i(x)]$ for each of the six covariance functions (for $i = 1, 2, \ldots, 6$).

Hint: What is the relation between variance and covariance?

Task 3.2: Which of the six covariance functions are stationary covariance functions?

Task 3.3: Let $\boldsymbol{X} = \{x_i\}_{i=1}^{100}$ be a sorted set of equidistant points in the interval $[0, 2]$. Figure 1 shows the covariance matrices for function values evaluated at $\boldsymbol{X}$ as well as samples from the corresponding Gaussian process prior for each of the six covariance functions. Match the plots to each of the six covariance functions.

Task 3.4: Now consider the following covariance function $k(\mathbf{x}, \mathbf{x}') = (\kappa^2 + \lambda^2 \mathbf{x}^T \mathbf{x}')^2$ for a problem in 2D, i.e. $\mathbf{x} \in \mathbb{R}^2$. The hyperparameters $\kappa, \lambda > 0$ are considered fixed and known. Compute the equivalent feature expansion for a linear model, i.e. determine the function $\phi(\mathbf{x})$ such that $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$.

Assume you are working on a regression problem. Instead of removing the empirical mean of the target $\{y_n\}$ prior to the analysis, you want to model the mean of the data explicitly using the following model $y_n = f(x_n) + b + \epsilon_n$, where $x_n \in \mathbb{R}$ is the input, $y_n \in \mathbb{R}$ is the output, $f \sim \mathcal{GP}(0, k(x, x'))$ and $e_n \sim \mathcal{N}(0, \sigma^2)$. The parameter $b \in \mathbf{R}$ is an intercept with prior distribution $p(b) = \mathcal{N}(b|0, \tau^2)$.

$$p(\mathbf{y}, \mathbf{f}, b) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|b)p(b) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{f}|b \cdot \mathbf{1}, \mathbf{K})\mathcal{N}(b|0, \tau^2) \tag{20}$$

where $\mathbf{K} \in \mathbb{R}^{N \times N}$ is a covariance matrix and $\mathbf{1} \in \mathbf{R}^N$ is a vector of ones.

Task 3.5: Compute the marginal prior distribution of $\mathbf{f}$, i.e. $p(\mathbf{f})$ and the marginal likelihood of the data $p(\mathbf{y})$ from eq. (20)**. Inspect your result for $p(\mathbf{y})$ and write down the equivalent kernel function.**

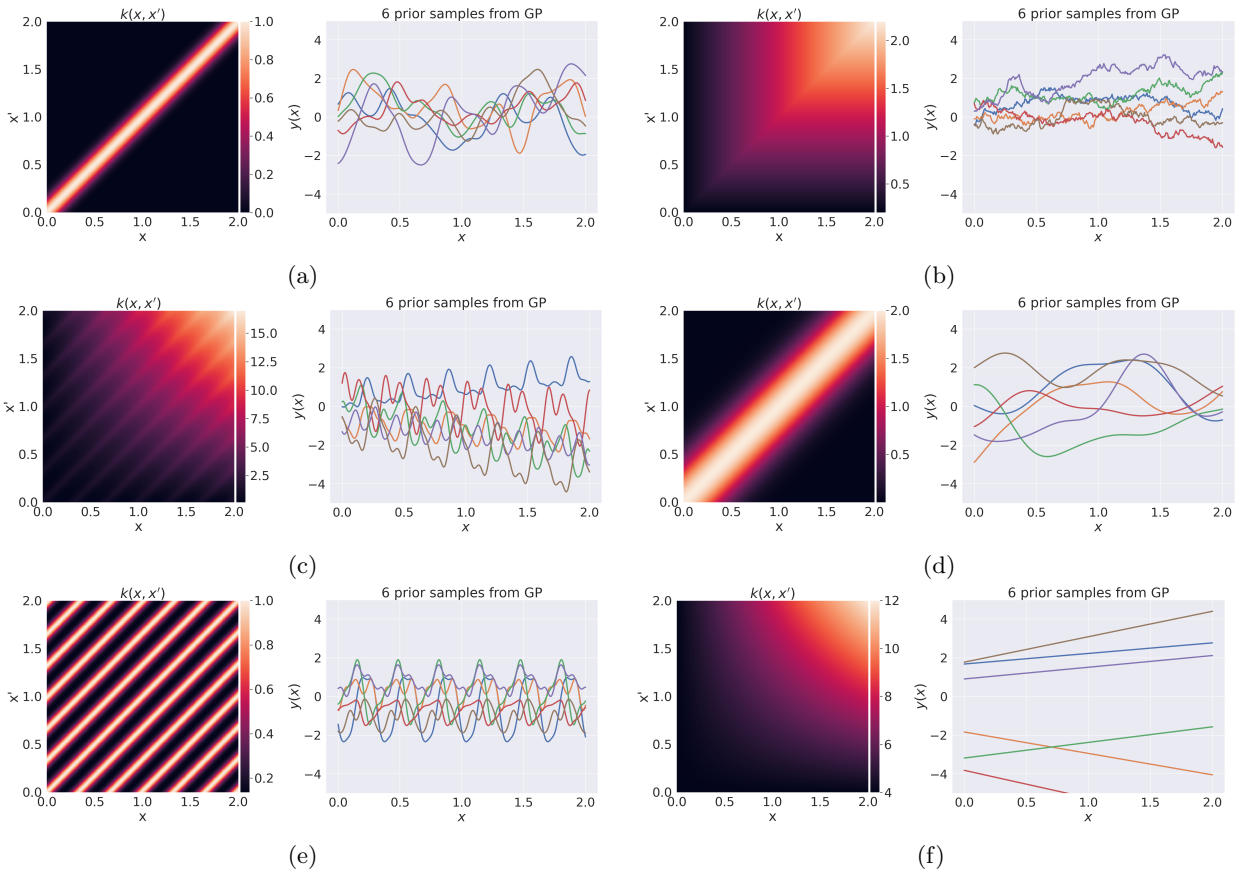

Figure 1: Covariance matrices and samples from the corresponding Gaussian process prior distribution for the six different covariance functions in randomized order.

Hint: Section 3.3 in Murphy1 might be handy.

Now we will re-analyze the bike sharing dataset from exercise 5 using the model in eq. (20). This time we will **not standardize** the targets before the analysis, so make sure to **remove that part**. We will still use the log transform, though. For the model in eq. (20), let $k(x, x')$ be the squared exponential covariance function.

Task 3.6: For the bike sharing dataset, plot the prior predictive distribution as well as 30 samples from the prior distribution in eq. (20) using the marginal prior $p(\mathbf{f})$ defined above. Use $\tau = 5, \ell = 50, \sigma = 0.1, \kappa = 1$. Use the same intervals for x as in exercise 5.

Task 3.7: Estimate the hyperparameters using marginal likelihood and report the estimates

Hint: You need to adapt the existing code from week 5 to handle the new hyperparameter.

Task 3.8: Plot the posterior distribution for f as well as 30 samples from the posterior distribution.

Task 3.9: In exercise 6, we studied linear models for a multi-class classification problem. Suppose we would like to use Gaussian processes instead of linear models to represent the latent functions for the same dataset. Write up the joint distribution for a multi-class classification model with 4 classes, where each latent function is modelled using a Gaussian process.

Hints:

1. *H*ow many latent functions do we have? What are the prior for each function? What is the likelihood?
2. See exercise 6 if you need to refresh the typical likelihood for multi-class classification problems.
3. *N*ote that you do need to any calculation here, but simply identify the component of the model and write up the joint distribution in terms of the relevant distribution, e.g. Categorical distribution and Gaussian process.