Exercise 1: Bayesian Linear Model

In the Bayesian linear model, we assume that the data follows the following law:

$$y = f(\mathbf{x}) + \epsilon = \boldsymbol{\theta}^T \mathbf{x} + \epsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and independent of $\mathbf{x}$. On the data-level this corresponds to

$$y^{(i)} \quad = \quad f\left(\mathbf{x}^{(i)}\right) + \epsilon^{(i)} = \boldsymbol{\theta}^T \mathbf{x}^{(i)} + \epsilon^{(i)}, \quad \text{for } i \in \{1, \ldots, n\}$$

where $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ are iid and all independent of the $\mathbf{x}^{(i)}$'s. In the Bayesian perspective it is assumed that the parameter vector $\boldsymbol{\theta}$ is stochastic and follows a distribution.

Assume we are interested in the so-called maximum a posteriori estimate of $\boldsymbol{\theta}$, which is defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}).$$

(a) Show that if we choose a uniform distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto 1,$$

then the maximum a posteriori estimate coincides with the empirical risk minimizer for the L2-loss (over the linear models).

(b) Show that if we choose a Gaussian distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2\tau^2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right], \qquad \tau > 0,$$

then the maximum a posteriori estimate coincides for a specific choice of τ with the regularized empirical risk minimizer for the L2-loss with L2 penalty (over the linear models), i.e., the Ridge regression.

(c) Show that if we choose a Laplace distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{\sum_{i=1}^p |\boldsymbol{\theta}_i|}{\tau}\right], \qquad \tau > 0,$$

then the maximum a posteriori estimate coincides for a specific choice of τ with the regularized empirical risk minimizer for the L2-loss with L1 penalty (over the linear models), i.e., the Lasso regression.

Exercise 2: Gaussian Posterior Process

Assume your data follows the following law:

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

with $\boldsymbol{f} = f(\boldsymbol{x}) \in \mathbb{R}^n$ being a realization of a Gaussian process (GP), for which we a priori assume

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')).$$

$\boldsymbol{x}$ here only consists of 1 feature that is observed for n data points.

(a) Derive / define the prior distribution of $\boldsymbol{f}$.

(b) Derive the posterior distribution $\boldsymbol{f}|\boldsymbol{y}$.

(c) Derive the posterior predictive distribution $y_*|x_*, \boldsymbol{x}, \boldsymbol{y}$ for a new sample x_* from the same data-generating process.

(d) Implement the GP with squared exponential kernel, zero mean function and $\ell = 1$ from scratch for $n = 2$ observations $(\boldsymbol{y}, \boldsymbol{x})$. Do this as efficiently as possible by explicitly calculating all expensive computations by hand. Do the same for the posterior predictive distribution of y_*. Test your implementation using simulated data.