Homework 1 (Lecture: Bayesian Learning)

Solve Exercises 1–8, and submit an answer sheet at the end of the lecture on 7.1.2016.

1 Linear regression model

Consider a linear regression model with unknown parameter $\boldsymbol{w} = \boldsymbol{a} \in \mathbb{R}^M$:

$$p(y|\boldsymbol{x}, \boldsymbol{a}) = \mathrm{Norm}_1(y; \boldsymbol{a}^\top \boldsymbol{x}, \sigma^2) = \frac{\exp\left(-\frac{(y-\boldsymbol{a}^\top \boldsymbol{x})^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}. \tag{1}$$

$\mathrm{Norm}_M(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the M-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$. We treat the noise variance σ^2 as a fixed constant.

Assume that we observed N samples $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})\}$, and that, for each given input $\boldsymbol{x}^{(n)}$, the output $y^{(n)}$ was independently and identically (i.i.d.) drawn from $\mathrm{Norm}_1(y; \boldsymbol{a}^{*\top}\boldsymbol{x}, \sigma^2)$ with unknown $\boldsymbol{a}^*$.

Define

$$\boldsymbol{y} = (y^{(1)}, \ldots, y^{(N)})^\top \in \mathbb{R}^N, \qquad \boldsymbol{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)})^\top \in \mathbb{R}^{N \times M}.$$

Then, the model likelihood is written as

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a}) = \mathrm{Norm}_N(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{a}, \sigma^2 \boldsymbol{I}_N) = \frac{\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{N/2}}, \tag{2}$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. We adopt Gaussian prior with mean $\boldsymbol{0}$ and covariance $\boldsymbol{C}$.

$$p(\boldsymbol{a}|\boldsymbol{C}) = \mathrm{Norm}_M(\boldsymbol{a}; \boldsymbol{0}, \boldsymbol{C}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{a}^\top \boldsymbol{C}^{-1}\boldsymbol{a}\right)}{(2\pi)^{M/2}|\boldsymbol{C}|^{1/2}}. \tag{3}$$

2 Posterior and Predictive distributions

Exercise 1: Derive the posterior distribution $p(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C})$ on $\boldsymbol{a}$.

Exercise 2: Derive the predictive distribution $p(y^*|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C})$ on the output y^* for a new input $\boldsymbol{x}^*$.

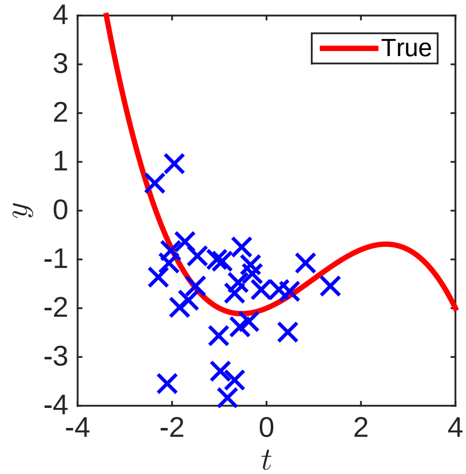

Figure 1: $N = 30$ samples from the linear regression model $y = \boldsymbol{a}^{*\top}\boldsymbol{x} + \varepsilon$, where $\boldsymbol{a}^* = (-2, 0.4, 0.3, -0.1)^\top$, $\boldsymbol{x} = (1, t, t^2, t^3)^\top$, and $\varepsilon \sim \mathrm{Norm}_1(0, 1^2)$.

Exercise 3: A set of training samples (shown as crosses in Fig.1) are given in a separate file "data.txt". The upper row corresponds to t and the lower row corresponds to y. Draw the same figure as Fig. 1. Numerically compute the mean $\widehat{y}$ and the covariance $\widehat{\sigma}_y^2$ of the predictive distribution $p(y^*|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C})$ as a function of t^* (on a grid, for example, $t^* = -4.00, -3.99, -3.98, \ldots, 4.00$). Here, set $\boldsymbol{C} = 10000\boldsymbol{I}_M$ and $\sigma^2 = 1$. Then, overlap the three curves corresponding to $\widehat{y}$ and $\widehat{y} \pm \widehat{\sigma}_y$ in the previous figure. Any programming language or software can be used. Only submit a printed figure.

3 Marginal Likelihood and Empirical Bayesian Learning

Exercise 4: Compute the marginal likelihood $p(\mathcal{D}|\boldsymbol{C})$ (without omitting any constant factor).

Estimating the hyperparameter $\boldsymbol{C}$ by maximizing the marginal likelihood is called *empirical Bayesian learning*. Equivalently, we minimize the negative log of the marginal likelihood

$$F^* = -\log p(\mathcal{D}|\boldsymbol{C}), \tag{4}$$

which is called the *Bayes free energy* or *stochastic complexity*. In addition, $\log p(\mathcal{D}|\boldsymbol{C})$ is called the *log marginal likelihood* or *evidence*.

Assume that

$$\boldsymbol{C} = \mathbf{Diag}(c_1^2, \ldots, c_M^2) \in \mathbb{D}_{++}^M, \qquad \boldsymbol{X} = \boldsymbol{I}_M, \tag{5}$$

where $\mathbb{D}_{++}^M$ denotes the set of positive definite diagonal matrices (i.e., $c_m^2 > 0, \forall m$). With the diagonal covariance to be estimated via empirical Bayesian learning, the prior

$$p(\boldsymbol{a}|\boldsymbol{C}) = \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi c_m^2}} \exp\left(-\frac{a_m^2}{2c_m^2}\right)$$

is called *automatic relevance determination* (ARD) prior.

Under the assumption (5), the Bayes free energy can be decomposed as

$$2F^* = \sum_{m=1}^{M} 2F_m^* + \text{const.}, \tag{6}$$

where each F_m^* depends on c_m^2 but not on $c_{m'}^2$ for $m' \neq m$.

Note that the latter assumption in (5) is only for simplifying the subsequent computation. In typical applications, $\boldsymbol{X}$ is not diagonal.

Exercise 5: Compute F_m^*.

Exercise 6: Draw F_m^* as a function of c_m^2 for $y_m^2 = 0, 1, 1.5, 2$ and $\sigma^2 = 1$. (Submit a printed figure.)

Exercise 7: Prove that the solution to $\min_{\boldsymbol{C}} F^*$ is given by

$$\widehat{c}_m^2 = \begin{cases} y_m^2 - \sigma^2 & \text{if } y_m^2 > \sigma^2, \\ +0 & \text{otherwise.} \end{cases} \tag{7}$$

Exercise 8: Derive the empirical Bayesian estimator $\widehat{\boldsymbol{a}}^{\mathrm{EB}}$ (the posterior mean of $\boldsymbol{a}$ with the hyperparameter $\boldsymbol{C}$ replaced with its estimator).