MLTS Exercise 02 - Bayesian Linear Regression

The goal is to implement a Bayesian linear regression. You can find the necessary equations for implementing the algorithm here. For derivation of the equations refer to *Bishop, 2006 Book part "The evidence approximation"*.

We need to

1. Formulate the Bayesian linear regression model and its parameters.
2. Learn the model parameters based on training data using marginal likelihood.
3. Use the model parameters to make predictions on previously unseen test data.

1) Model Definition

We are interested in a Bayesian way of learning linear regression with degree M (sometimes called polynomial regression with degree M). Our data set is $\mathcal{D} = \{X, \mathbf{y}\}$ where we have N data points and $X \in \mathcal{R}^M, \mathbf{y} \in \mathcal{R}$.

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M + \epsilon = X^T \mathbf{w} + \boldsymbol{\epsilon},$$

where ϵ is a Gaussian distributed noise $\epsilon \sim \mathcal{N}(0, 1)$.

We are interested in a **Gaussian prior** over model parameters $\mathbf{w}$, which is as follows:

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1} I).$$

with zero mean and the parameter α which its inverse determines the variance of the distribution. Here, we are interested to learn the parameter of the prior α based on the data. This is called Empirical Bayes methods.

We are interested also in a **Gaussian likelihood** as follows:

$$p(\mathbf{y}|X, \mathbf{w}) = \mathcal{N}(\mathbf{y}|X^T \mathbf{w}, \beta^{-1} I),$$

Inference in the Bayesian linear model is based on the **posterior distribution** over the weights

$$p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w}|\alpha)}{p(\mathbf{y}|X)}$$

Due to the choice of a conjugate Gaussian prior distribution, the posterior will also be Gaussian.

2) Learning parameters using marginal likelihood

Marginal likelihood

In Bayesian modeling we can learn parameters by integrating over the parameters $\mathbf{w}$. The marginal likelihood can be obtained using:

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)d\mathbf{w},$$

If we take integration of the above equation and get log of the result we obtain log marginal likelihood as follows:

$$\ln p(\mathbf{y}|X) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|A| - \frac{N}{2}\ln(2\pi)$$

where $|A|$ is the determinant of matrix A and

$$A = \alpha I + \beta X^T X,$$

$$E(\mathbf{m}_N) = \frac{\beta}{2} \parallel \mathbf{y} - X\mathbf{m}_N \parallel^2 + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N,$$

$$\mathbf{m}_N = \beta\ A^{-1}\ X^T\mathbf{y}.$$

To easily get the inverse of A, we can compute its **Cholesky decomposition** U and invert this matrix:

$$U = cholesky(A)^T$$
$$U_i = U^{-1}$$
$$A^{-1} = U_i U_i^T$$

Maximizing marginal likelihood

The only thing remains in order to determine the marginal likelihood is to estimate the parameters α and β. We cannot find the maximum values of α and β analytically. Instead we take derivative of log marginal likelihood w.r.t. to each parameter and maximize it iteratively.
Maximization of $p(\mathbf{y}|X)$ w.r.t. to α satisfies

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T\mathbf{m}_N},$$

where

$$\gamma = M - \alpha\ tr(A^{-1}),$$

$tr(A)$ denotes trace of matrix A.

Maximization of $p(\mathbf{y}|X)$ w.r.t. to β satisfies

$$\beta = \frac{N - \gamma}{\parallel \mathbf{y} - X\mathbf{m}_N \parallel^2}.$$

Solution for α and β can be obtained by choosing an initial value for them and then using this to calculate m_N and γ and then re-estimate α and β, repeating until convergence (marginal likelihood does not change much) or after sufficiently large number of iterations. The values of α and β can be re-estimated together after each update of γ.

3) Prediction

Once we learned the model parameters α and β (including the intermediate parameters A, m_N and γ), we can use them to perform prediction. To make predictions y_* for the test data x_*, we average over all possible parameter values, weighted by their posterior probability as follows:

$$p(y_*|x_*, \mathbf{y}, X) = \int p(y_*|x_*, \mathbf{w}) p(\mathbf{w}|\mathbf{y}, X) d\mathbf{w} = \mathcal{N}(y_*, {\sigma_*}^2) = \mathcal{N}(\mathbf{m}_N^T \; x_*, \tfrac{1}{\beta} + x_*^T \; A^{-1} \; x_*)$$

where

$$A = \alpha I + \beta X^T X$$

and

$$m_N = \beta \; A^{-1} \; X^T \mathbf{y}.$$