Modified Bayesian Statistical Inference and Renormalization Group

Toshiaki Aida
Faculty of Engineering, Okayama University, Japan

Abstract

Bayesian statistical inference plays a central role as a fundamental framework of a large variety of probabilistic information processing problems. For example, in system identification problems such as learning, regression, $\cdots$, it enables us to determine the parameters of a system in a probabilistic sense, even when we have no enough number of outputs from it. This suggests that Bayesian framework has a mechanism to reduce the number of degrees of freedom to be considered according to the number of data, so that we may determine those effective degrees of freedom. It also supports the existence of the mechanism that Bayesian framework derives information criteria in a natural way.

On the other hand, statistical mechanics has a long history in the research of the response of a system under the change of its effective number of degrees of freedom. Renormalization group is one of the most representative theories to deal with the subject, and has succeeded in predicting such scaling behavior of a system as is seen in critical phenomena.

In our talk, our first purpose is to make it clear how the notion of renormalization group is naturally involved in Bayesian framework of statistical inference. Next, we extend the framework to obtain better predictive performance. Finally, we report the result of a non-perturbative analysis applied to the extended framework.

The connection between Bayesian framework and renormalization group is easier to see in the context of non-parametric models. For this purpose, we start with the following two models for density estimation [1, 2, 3]:

$$S[\phi] \simeq \frac{l}{2}\int dx\ (\partial_x\phi)^2 + N\left[\frac{1}{l}\int dx\ e^{-\phi(x)} - 1\right] + N\int dx\ P_N(x)\phi(x)\ , \tag{1}$$

and for regression [4]:

$$S[\phi] = \frac{l}{2}\int_0^L dx\ (\partial_x\phi)^2 + \frac{N}{2\sigma_a^2}\int_0^L dx\ P_N(x)\ [\phi(x) - y(x)]^2\ . \tag{2}$$

Here, $\phi(x)$ is a function to be inferred, and $P_N(x)$ is a distribution of given data $P_N(x) \equiv \frac{1}{N}\sum_{i=1}^N \delta(x - x_i)$. Bayes theorem enables us to infer the function through the posterior distribution given by $exp(-S[\phi])/Z$ with Z a normalization constant.

In these models, the terms with no derivative are proportional to the number of examples N, because they are the contribution from likelihoods or from a constraint. Among them, ϕ^1-term has the effect to shift the expectation value of the function ϕ, and the inverse of the coefficient of ϕ^2-term defines a length scale within which a datum can affect the function. In this way, the length scale divides a space into bins, which lead to effective degrees of freedom.

This is because Bayesian setting has a mechanism to reduce effectively the number of degrees of freedom of a system according to the number of observed data N: the length scale sets a large bin size when N is small, while a small bin size when N is large.

Motivated by the above observation, let us discuss the problem in more general setting which extends the one in ref. [5] to a functional case.

After we have observed data $y^N \equiv \{y_1, \ldots, y_N\}$, which are generated independently of each other for $x^N \equiv \{x_1, \ldots, x_N\}$ according to a probability $P(y|\phi^*, x)$ parameterized by a function ϕ^*, we want to predict an output y_{N+1} for a new input x_{N+1}.

In Bayesian framework, it is natural to predict the output, based on the average of $P(y_{N+1}|\phi, x_{N+1})$ over the posterior distribution $P(\phi|y^N, x^N)$ for the function ϕ.

$$\begin{aligned} P(y_{N+1}|y^N, x^{N+1}) &= \int \mathcal{D}\phi \; P(\phi|y^N, x^N) \; P(y_{N+1}|\phi, x_{N+1}) \; , \\ &= \frac{\int \mathcal{D}\phi \; P(\phi) \; \prod_{i=1}^{N+1} P(y_i|\phi, x_i)}{\int \mathcal{D}\phi \; P(\phi) \; \prod_{i=1}^{N} P(y_i|\phi, x_i)} \; , \end{aligned} \tag{3}$$

where $P(\phi)$ is a prior distribution and takes the form of the exponential of a derivative term.

The performance for prediction is measured by the Kullback-Leibler divergence between the predictive distribution (3) and the true one, averaged over all the data and the parameter which is also a probabilistic variable now [5]. In our case, the error ΔI_{N+1} is given by the K.L. divergence averaged over the data $\{y^N, x^{N+1}\}$ and the true function ϕ^*.

$$\Delta I_{N+1} \equiv < -\ln P(y_{N+1}|y^N, x^{N+1}) \; + \; \ln P(y_{N+1}|\phi^*, x_{N+1}) >_{y^{N+1}, \; x^{N+1}, \; \phi^*} \; , \tag{4}$$

$$= < \Delta\Gamma_{N+1}[\hat{\phi}_{N+1}] \; + \; \ln P(y_{N+1}|\phi^*, x_{N+1}) >_{y^{N+1}, \; x^{N+1}, \; \phi^*} \; . \tag{5}$$

Here, $\Delta\Gamma_{N+1}[\hat{\phi}_{N+1}] \equiv \Gamma_{N+1}[\hat{\phi}_{N+1}] - \Gamma_N[\hat{\phi}_N]$ is a difference equation for *effective action* defined by

$$\Gamma_N[\phi] \; = \; -\ln \int \mathcal{D}\chi \; P(\phi + \chi) \prod_{i=1}^{N} P(y_i|\phi + \chi, x_i) \exp\Big(\frac{\delta\Gamma_N}{\delta\phi}\chi \Big) \; . \tag{6}$$

$\hat{\phi}_N$ is a solution of the equation $\delta\Gamma_N[\phi]/\delta\phi = 0$, and is known to be equal to the expectation value of ϕ with Z a normalization constant.

$$\hat{\phi}_N(x) \; = \; \frac{1}{Z} \int \mathcal{D}\phi \; \phi(x) \; P(\phi) \; \prod_{i=1}^{N} P(y_i|\phi, x_i) \; . \tag{7}$$

$\Delta\Gamma_{N+1}[\hat{\phi}_{N+1}]$ is nothing but *renormalization group equation*, because the increase of N causes the decrease of a length scale of the model, as is discussed previously. Thus, we have made clear the connection between prediction performance and renormalization group equation.

The error (4) shows well-known universal asymptotic behavior $\Delta I_{N+1} \sim D/2N$, where D is the number of effective degrees of freedom and simply equal to the number of parameters in parametric cases [5].

However in (3), apart from Bayesian framework, if we introduce a scaling part $F_N(\phi)$ to the prior distribution $P(\phi)$:

$$P(\phi) \; \longrightarrow \; P_N(\phi) = (Z_N)^{-1} P(\phi) \exp[-F_N(\phi)], \; P_{N+1}(\phi) = (Z_{N+1})^{-1} P(\phi) \exp[-F_{N+1}(\phi)] \tag{8}$$

we are able to obtain better predictive performance.

It is naturally achieved through the renormalization group equation $\Delta\Gamma_{N+1}[\hat{\phi}_{N+1}]$.

We will report the result of a non-perturbative analysis developed for exact renormalization group equations.

References

[1] W. Bialek, C.G. Callan and S.P. Strong, Phys.Rev.Lett. **77** (1996) 4693.
[2] T.E. Holy, Phys. Rev. Lett. **79** (1997) 3545.
[3] T. Aida, Phys. Rev. Lett. **83** (1999) 3554.
[4] T. Aida, Proc. IEEE Int. Workshop on Neural Networks for Signal Processing (2002) 179.
[5] M. Opper and D. Haussler, Phys. Rev. Lett. **75** (1995) 3772.