Model-induced Regularization

Shinichi Nakajima  
Nikon Corporation, Tokyo 140-8601, Japan  
Nakajima.S@nikon.co.jp

Masashi Sugiyama  
Tokyo Institute of Technology and JST PRESTO, Tokyo 152-8552, Japan  
Sugi@cs.titech.ac.jp

Abstract

When the Bayesian estimation is applied to modern probabilistic models, an unintentional strong regularization is often observed. We explain the mechanism of this effect, and introduce relevant works.

Suppose we are given i.i.d. samples \( \{x_1, \ldots, x_n \in \mathbb{R} \} \) taken from Gaussian model with the mean parameter \( u \in \mathbb{R} \):

\[
p(x) = \mathcal{N}(x; u, 1^2).
\]

(1)

Assuming Gaussian prior, \( p_u(u) = \mathcal{N}(u; 0, c_u^2) \), where \( c_u^2 > 0 \) is a variance hyperparameter, we can perform Bayesian estimation, controlling \( c_u^2 \) as a regularization constant. What if we set \( c_u^2 \) to a large value (\( c_u^2 \to \infty \))? The answer may be trivial; we get an unregularized estimator. (More accurately, the mode of the Bayesian predictive distribution coincides to the maximum likelihood (ML) estimator.)

Suppose next the following model:

\[
p(x) = \mathcal{N}(x; ab, 1^2).
\]

(2)

Here, the parameters are \( a, b \in \mathbb{R} \), whose product corresponds to the parameter \( u \) in the original model (1). Let us assume Gaussian priors on \( a \) and \( b \): \( p_a(a) = \mathcal{N}(a; 0, c_a^2), p_b(b) = \mathcal{N}(b; 0, c_b^2) \). Will we similarly get an unregularized estimator of \( u = ab \) when \( c_a^2, c_b^2 \to \infty \)?

The answer is NO. The estimator tends to be strongly regularized. We call this effect model-induced regularization (MIR), since it is inherent in the model likelihood function.

Actually, Eq.(2) is a special case of the matrix factorization model, and therefore, MIR explains the empirically observed superiority (Salakhutdinov & Mnih, 2008) of full-Bayesian estimation over maximum a posteriori (MAP) estimation. Here, note that MIR is caused by density non-uniformity of distribution functions in the parameter space, and therefore observed only when at least one parameter is integrated out. (No parameter is integrated out in MAP.) Other popular models in machine learning, e.g., mixture models and hidden Markov models, also have a similar structure to Eq.(2), which induces MIR.

The origin of MIR can be explained in terms of the Jeffreys prior (Jeffreys, 1946), with which the two models, (1) and (2), give the equivalent estimation. Another explanation has been done in the context of visual recognition (Freeman, 1994). Although the idea of the Jeffreys prior is widely known, people seem to underestimate the strength of this effect.

In our poster, we will explain why MIR occurs. Then, works that relate MIR with singularities of probabilistic models are introduced. A powerful procedure for quantitative evaluation of MIR has been developed, and applied to various models (Watanabe, 2009). Theoretical analysis has been extended to the variational Bayesian (VB) approximation. We will also introduce works that clarified the strength of MIR when VB is applied (Nakajima & Sugiyama, 2010).

References


