Bayesian Analysis (2012) **7**, Number 1, pp. 45–46

Comment on Article by Sancetta

Feng Liang*

I will start my discussion with some clarification on the difference between prediction consistency (in the Cesaro sense) and universality. Suppose we are given observations $Z_1, Z_2, \ldots$ sequentially, which, without loss of generality, are assumed to be i.i.d. samples from some distribution P_θ with density p_θ, where $\theta \in \Theta$. Of interest is to estimate p_θ sequentially based on previous observations. A natural Bayes estimator at time t, based on $\mathbf{Z}_1^{t-1} = (Z_1, \ldots, Z_{t-1})$, is given by

$$p_w(z \mid \mathbf{Z}_1^{t-1}) = \int p_\theta(z) w(\theta \mid \mathbf{Z}_1^{t-1}) d\theta, \tag{1}$$

where $w(\theta \mid \mathbf{z}_1^{t-1}) \propto w(\theta) \prod_{i=1}^{t-1} p_\theta(z_i)$ is the posterior distribution of θ updated by data $(z_1, \ldots, z_{t-1})$ and $w(\theta)$ is the prior distribution. At time t, we measure the error/risk of the Bayes estimator p_w by its Kullback-Leibler divergence with respect to the true density p_θ, namely,

$$D_t(p_\theta || p_w) = E_{Z_1,\ldots,Z_{t-1}|\theta} \int p_\theta(z) \log \frac{p_\theta(z)}{p_w(z \mid \mathbf{Z}_1^{t-1})} dz. \tag{2}$$

An interesting question is under what conditions p_w is a consistent estimator of p_θ. That's the question studied in Barron (1987). His answer relevant to this paper is that if prior w is information dense at θ (see Section 1 of Sancetta's paper), then p_w is consistent in the Cesaro sense, i.e., the Cesaro average of D_t goes to zero,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} D_t(p_\theta || p_w) = 0. \tag{3}$$

Universality of prediction, studied in this paper, requires the supremum of the Cesaro average go to zero,

$$\lim_{T\to\infty} \sup_{\theta\in\Theta} \frac{1}{T} \sum_{t=1}^{T} D_t(p_\theta || p_w) = 0, \tag{4}$$

and therefore is stronger than Cesaro consistency (3). For example, consider a simple normal mean problem with Z_t i.i.d. $\sim \mathsf{N}(\theta, 1)$. No Bayes procedures p_w are universal, unless θ is in a compact set; on the other hand, many priors that are information dense lead to consistent Bayes prediction (in the Cesaro sense). Here no estimators (not just Bayes) are universal because our maximum error at $t = 1$ is infinity: without conditioning on any data, our estimate at $t = 1$ is just a fixed density, whose KL divergence with respect to $\mathsf{N}(\theta, 1)$ can be made arbitrarily large (unless the parameter space Θ is bounded), therefore $\sup_\theta D_1 = \infty$. However, in most real applications, what happens at $t = 1$ is of little interest. So we could drop D_1 (or the first couple of D_i's)

*Dept. of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, liangf@uiuc.edu

 DOI:10.1214/12-BA701B