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

from the Cesaro average and study universality of predictions that are based on some initial observations. This is exactly the framework used in Liang and Barron (2004) for a minimax study of density estimation and universal data compression. My first question is whether some of the results in this paper, such as Theorems 1 and 3, can be extended to the conditioning framework, and therefore can cover simple models like the normal mean problem or regression with an unbounded parameter space.

Note that the Cesaro consistency (3) does not imply that D_t goes zero, although the latter is more relevant in practice. For KL divergence, it turns out to be easy to modify p_w to achieve consistency in the sense that $\lim_t D_t = 0$ (Barron, 1987). For example, consider the following sample average of p_w,

$$\tilde{p}_w(z \mid \mathbf{Z}_1^{t-1}) = \frac{1}{t-1} \sum_{i=1}^{t-1} p_w(z \mid \mathbf{Z}_1^i).$$

By Jensen's inequality and the convexity of the KL divergence (with respect to the second argument), we have

$$D_t(p_\theta || \tilde{p}_w) \leq \frac{1}{t-1} \sum_{i=1}^{t-1} D_{i+1}(p_\theta || p_w),$$

which goes to zero as a consequence of (3) when w is information dense at θ. My second question is whether similar statements (on the limiting behavior of D_t) can be made for universal prediction.

The result in Section 3 on universality of Bayesian model averaging is interesting. I am wondering whether K, the number of models, has to be pre-fixed. In many modern statistical applications, the model space may increase with the sample size. So it will be nice if Theorem 4 can be extended to cover such scenario where $K = K(T)$.

References

Barron, A. R. (1987). *Problems in Communications and Computation*, chapter Are Bayes rules consistent in information?, 85–91. New York: Springer.

Liang, F. and Barron, A. R. (2004). "Exact Minimax Strategies for Predictive Density Estimation, Data Compression and Model Selection." *IEEE Transactions Information Theory*, 50: 2708–2726.