Approximate inference for the loss-calibrated Bayesian

Simon Lacoste-Julien, Zoubin Ghahramani
Department of Engineering
University of Cambridge, Cambridge, CB2 1PZ, UK
`(simon.lacoste-julien, zoubin)@eng.cam.ac.uk`

Executive summary

In this framework talk, we consider the impact of approximate inference in the context of Bayesian decision theory. We argue for the need to shift focus from the traditional approach of solely approximating the Bayesian posterior to developing *loss-calibrated* approximate Bayesian inference methods, in analogy to what is done in discriminative machine learning. We outline research questions which arise naturally from such a shift of focus.

Scope

Bayesian methods have enjoyed a surge of popularity in machine learning in the last decade. Even though it is often overlooked, the Bayesian paradigm is theoretically motivated from Bayesian decision theory, which provides a well-defined theoretical framework for rational decision making under uncertainty. Its ingredients are a loss $L(\theta, a)$ for an action $a \in \mathcal{A}$, a prior $p(\theta)$ and an observation model $p(x|\theta)$. Even if we assume that our subjective beliefs about $p(x, \theta)$ have been well-specified, we usually need to resort to approximations in order to use them in practice. Despite the central role of the loss in the decision theory formulation, most prevalent approximation methods seem to focus on approximating the posterior $p(\theta|x)$. For example, citing [2] in ch.6:

> We shall focus in this chapter on approximations to [the posterior] and integrals involving [the posterior] because this is the cornerstone of computational difficulties with Bayesian inference. In addition, if [the posterior] can be correctly approximated, it is usually possible to derive an approximation to [the posterior expected loss] for an arbitrary [action], and then to use a classical minimization method.

In contrast, our main point is to bring back in focus the need to *calibrate* the approximation methods to the loss under consideration. This philosophy has already been widely applied in the discriminative machine learning literature, as for example with the use of *surrogate loss functions* [1, 3]. In contrast, the "loss-calibrated" approximation approach seems to have been mainly limited in Bayesian approaches to simple settings and losses such as regression with quadratic loss or hypothesis testing with 0-1 loss. Our question is general and could apply to general losses, but for the sake of concreteness and motivated from applications in machine learning, we will focus on the predictive setting.

Category: learning theory, learning algorithms **Preference:** Oral

The predictive setting

Our goal is to estimate a function $h : \mathcal{X} \to \mathcal{Y}$ where the output space $\mathcal{Y}$ is discrete. A domain-specific loss $\ell(y, y')$ is given. In the context of *structured prediction* (such as machine translation where $\mathcal{X}$ are sentences in one language and $\mathcal{Y}$ are possible translations in another language), the loss $\ell(y, y')$ is highly informative about $\mathcal{Y}$ by providing some kind of distance on discrete objects, in contrast to the 0-1 loss for classification or quadratic loss for regression. We obtain the standard statistical decision theory setting by defining a suitable prediction loss: $L(\theta, h) \doteq \mathbb{E}_{(x,y)\sim p(x,y|\theta)}[\ell(y, h(x))]$, where an action in this case is a function h. Given a dataset $\mathcal{D} = \{(x_i, y_i)_{i=1}^n\}$ of iid observations from $p(x, y|\theta)$ for unknown θ, our (Bayesian) goal is to find a function h which minimizes the *expected posterior loss* $R(h|\mathcal{D}) \doteq \int_\Theta L(\theta, h)p(\theta|\mathcal{D})d\theta$. We consider in our framework the loss ℓ to be known and the expected posterior loss to be the ultimate evaluation metric for our possible actions, though frequentist notions of risk could also be studied. The standard approach to solve this problem in practice when the quantities are not analytic is to get an approximation for the posterior $p(\theta|\mathcal{D})$. This can be done in a deterministic fashion, as for example using a variational approach of selecting a suitable $q(\theta) \in \mathcal{Q}$ with minimum KL divergence to the posterior; or in a stochastic fashion, such as using MCMC sampling from the posterior (using only a finite number of samples yields an approximate representation of the posterior). Given the approximation q to the posterior, an (approximate) optimal action is chosen by minimizing the expected posterior loss R_q which assumes that q is the true posterior.

On the other hand, this traditional approach can yield actions which are suboptimal (in term of expected posterior loss) compared to either using a different criterion to choose q and minimizing R_q, or use a different approximation framework altogether (as for example approximating the predictive distribution $p(y|x, \mathcal{D})$ directly rather than approximating the posterior). We consider the following settings to illustrate the relevance of this point.

non-trivial ℓ: Given an asymmetric classification loss or a structured loss ℓ, the best approximation to the posterior doesn't necessarily translate to yielding good decisions for the predictive loss. For example, a model $p(x, y|\theta)$ could yield the biggest losses on y for θ in the tails of the posterior, which are not necessarily well approximated by a generic variational approach or a MCMC approximation.

test distribution mismatch: We can consider the setting where $p(x, y|\theta) = p(x|\alpha)p(y|x, \theta)$ and α is known but different for the generating process for the dataset $\mathcal{D}$ and the test set loss $L(\theta, h)$. Different approximations q to the posterior could yield better performance in different regions of $\mathcal{X}$, and the test set distribution $p(x|\alpha)$ should thus be included to pick the best approximation tradeoff.

parametric decision boundary: Instead of supposing that $h(x)$ could be arbitrary, practical considerations could make us consider h to belong to a parametric family $\mathcal{H}$ (e.g. linear decision boundaries). Again, this will induce tradeoffs in the expected posterior loss which won't be captured by only looking at the quality of the approximation to the posterior.

These examples suggest that we need approximate Bayesian inference methods (both deterministic or stochastic) that are *calibrated to the loss function $L(\theta, h)$*. Having outlined these limitations of the standard Bayesian setup, we propose to explore the space of such loss-calibrated inference methods.

References

[1] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[2] Christian P. Robert. *The Bayesian Choice*. Springer, New York, 2001.

[3] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

This is basically the generalization error. It has unfortunately been called the 'risk' by Vapnik, clashing with the terminology already used in statistical decision theory.