

Discriminative MCMC
Kai Puolamäki, Jarkko Salojärvi, Eerika Savia and Samuel Kaski
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, FINLAND
samuel.kaski@tkk.fi, <http://www.cis.hut.fi/projects/mi/prima>

In generative modeling tasks, it is well-known that usual Bayesian inference is not optimal for generalizing to new data if the model family is incorrect, that is, if the data does not come from any of the models within the model family. Arguably the best solution is to improve the model family by incorporating more prior knowledge. This is not always possible or feasible, however, and simplified models are being generally used, often with good results. There are good reasons for still applying Bayesian-style techniques [4] but the general problem of how to best do inference with incorrect model families is still open.

In discriminative modeling, here meaning inference on the distribution $p(y|x)$, the question of using discriminative vs. generative models has attracted a lot of interest. In essence, the question has been whether to model $p(y|x)$ directly or to build a generative model for the joint distribution $p(y, x)$ and compute the conditional distribution from that. It is easy to show that, for instance, the point estimates computed by maximizing the joint likelihood and the conditional likelihood differ. Maximum conditional likelihood works better asymptotically, and it can be optimized with expectation-maximization-type procedures [5]. Some other related point estimates have been proposed but while point estimates have been studied thoroughly, fewer results exist on extensions from point estimates to posterior distributions. The standard posterior distribution is optimal for discriminative modeling if the model family is correct, but is there an extension that would be analogous to standard Bayesian inference while working better for incorrect model families?

We are aware of only one suggestion, the so-called discriminative posterior [3], which has empirical support as well [1]. The posterior has however, as far as we know, only been justified more or less heuristically. We give an axiomatic justification, introduce Markov Chain Monte Carlo-type methods for computing with the posterior, and demonstrate empirically that it works as expected. The inference reduces to standard Bayesian inference if there are no covariates.

There exists another well-established line of research on using Bayesian methods for discriminative learning, namely Bayesian regression, where the x are considered to be co-variates of the model for y . From the generative modeling perspective such regression ignores any information about y supplied by x . This is justified if (i) the covariates are explicitly chosen when designing the experimental setting and hence are not noisy, or (ii) there is a separate set of parameters for generating x and $y|x$, and the sets are assumed to be independent in their prior distribution. Then the posterior factors out into two parts, and the parameters used for generating x are not needed or useful in the regression task. See [2] for more details.

For the purpose of regression, the discriminative posterior makes it possible to use more general model structures: in essence any generative model. The gained advantage, compared to using the standard non-discriminative posterior, should be that the predictions should be more accurate assuming the model family is incorrect. Compared to Bayesian regression the predictions should be better if the introduced generative model for x is informative.

References

- [1] J. Cerquides and R. L. Mántaras. Robust Bayesian linear classifier ensembles. In J. Gama, R. Camacho, P. Brazdil, A. Jorge, and L. Torgo, editors, *Machine Learning: ECML 2005*,

- Lecture Notes in Artificial Intelligence 3720, pages 72–83, Berlin, Germany, 2005. Springer-Verlag.
- [2] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, FL, 1995.
 - [3] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Roos, H. Tirri, and H. Wettig. Supervised posterior distributions. presentation at the Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain, 2002. <http://homepages.cwi.nl/~pdg/presentationpage.html>.
 - [4] L. K. Hansen. Bayesian averaging is well-tempered. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 265–271. MIT Press, 2000.
 - [5] J. Salojärvi, K. Puolamäki, and S. Kaski. Expectation maximization algorithms for conditional likelihoods. In L. D. Raedt and S. Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning (ICML-2005)*, pages 753–760, New York, USA, 2005. ACM press.

Topic: learning algorithms

Preference: poster