Machine Learning from a Bayesian Perspective

Mohammad Emtiyaz Khan
RIKEN Center for AI Project
Tokyo, Japan
emtiyaz.khan@riken.jp

November 8, 2021

Abstract

I summarize a Bayesian perspective of machine learning. We view Bayes as an optimization problem whose solutions use the information-geometry of the posterior. Using this perspective, we can show that many machine-learning methods have a (more general) Bayesian side to them. I believe this perspective to be essential for bridging the gap between 'artificial' and 'natural' learning systems.

1 A note about the note

For now, this note deliberately lacks details. One way to read this is to use the accompanying slides. My hope is to add some equations, figures and illustration in the future. Many technical details discussed here can be found in Khan and Rue [2021]

2 Machine learning and Bayes

A main goal of machine-learning is to design AI systems that can learn like us. We humans, and other animals, collect experiences throughout our lives to learn and adapt. Machines currently are extremely bad at this. Majority of successful machine-learning paradigms are the ones that use 'bulk' learning in a 'static' world, where all the information is assumed to be available at once and the world stands still while we learn about it. This is far from the reality of the world we live in, and it is not surprising to see such systems fail. How can we bridge this gap between machines and living-beings? Taking a Bayesian perspective seems to be one way to go, but we argue that this is perhaps the only way forward.

Why? Information processing systems, both artificial and natural, are often Bayesian, or at least of that nature. This is true in many fields, such as psychology, economics, neuroscience, physics, and information theory, where processing of information/evidence naturally follows a Bayesian course. Yet, Bayes appears as a specialized technique in ML, and being a Bayesian is considered to be a matter of choice. Our goal here is to demonstrate that this is not a question of choice, and that most successful ideas used in machine-learning today are in fact of a Bayesian nature. The goal however is not to just point out the Bayesian nature, rather show the gains obtained by this perspective. The hope is that the reader will see benefits arising from the Bayesian perspective, and use it to further the goals of machine learning.

3 Benefits and challenges of Bayes

What are some of the benefits of a Bayesian perspective, one may ask. The main benefit comes from a probabilistic viewpoint where the model is not just one parameter (vector), rather we have many such models (parameter vectors) sampled from a 'posterior' distribution. Think of it as a jury made up of a diverse set of powerful people, rather than a mighty king. The decisions made by the jury are more robust and less biased towards one type of errors, and a diverse jury can be more accommodating to new evidence that may come in the future. In technique terms, this translates to decisions that 'generalize' to more tasks, 'overfit' less to just one task, and 'adapt' to new data easily. The decisions of the jury when averaged also tell us about the quality of the jury, which enables 'model selection' using 'marginal likelihoods'. These are just some of the benefits of the Bayesian perspective.

But then why are we not using Bayes all the time? Well, this is where most turn their backs to Mr. Bayes. It turns out that collecting decisions of *all* the jury members is painful (as expected). To preserve the benefits, each jury member must be weighted according to their 'posterior' probability, which requires a sum over all members, a beast to compute for large juries. For example, for an infinite number of them, the sum could be a high-dimensional intergral with no close-form expression. Only in simple problems this is possible, for example, in models with 'conjugate' priors. A large number of Bayesians devote their time to this pure goal of posterior computation, but this is also where many turn their backs and take to anti-Bayesianism.

4 Why Bayes?

Why must we compute the posterior accurately, if we can get similar benefits with a simpler approach? This is a common questions among non-Bayesians, and a sensible one too, which is why we will mainly focus on this question in what follows. The answer is simple: we do not have to use Bayesian techniques all the time, but often we end up with a Bayesian solution anyways, irrespective of the route we take. Therefore a knowledge of the underlying Bayesian principle can save a lot of trouble. This is true even in non-obvious cases where the problems are inherently non-Bayesians, and have nothing to do with the sums or integrals.

In what follows, our attempt will be to convince the reader of this point. We will unravel the underlying Bayesian principle of many ML algorithms. We will do so by showing that they all can be seen as specific instances of a single Bayesian algorithm. The generality is due to an optimization reformulation of Bayes, whose solutions have roots in information geometry. This will then connect the Bayesian solutions to non-Bayesian ones.

5 Bayes as optimization

The posterior computation is usually done via Bayes rule, but the computation can also be viewed as a minimization problem over the space of all probability distributions. In this formulation, we minimize the sum of the expected 'loss', taken with respect to the distibution, and the negative-entropy of the distribution.

$$\min_{p \in \mathcal{P}} \mathbb{E}_{p(\boldsymbol{\theta})} \left[\sum_i \ell_i(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}) \right] - \mathcal{H}(p) \tag{1}$$

At first, it might appear overcomplicated because one may ask "why not just minimize the loss directly?" The reformulation here is takes a much more principled approach, and there are important reasons to do so.

One way to understand the reformulation is to see the problem as a 'sequential information processing'. Without any data, we may know nothing (assuming no prior knowledge). Therefore, we can accept any explanation, which in technical terms means that all model parameters are assigned equal probability, and the entropy is the highest it can be. Given additional information, the entropy must reduce, more and more so, as we observe more data. The optimization reformulation captures this sequential view of Bayesian principles, originally proposed by E. T. Jaynes under the maximum-entropy principle [Jaynes, 1982], which he later realized to be a more general Bayesian principle.

6 Bayes is everywhere

Despite its Bayesian nature, the above formulation appears in many different non-Bayesian settings. For example, in evolution strategy (random search), a similar formulation appears with a 'search' distribution which helps us search for better minima. For global-optimization, the distribution is used to obtain 'stochastic relaxation'. In homotopy methods, the loss is convolved with a distribution to 'smooth' to avoid spurious minima, and similar ideas are used in continuation methods, and optimization with perturbation. Finally, policy distributions used in reinforcement learning also use similar objectives. Interested readers can see Lin et al. [2021, Sec. 1] for some more details and Khan et al. [2018, Sec. 1.1] for references.

These non-Bayesian settings often employ entropy regularization, with the hope that the diversity introduced due to the distribution will help us explore the space better, and avoid getting stuck at a local minima. From a Bayesian perspective, on the other hand, the entropy is a must-have, not an add-on feature. If one is really unhappy with the idea, a compromise can be made with a modification where a temperature parameter is included in front of the entropy. Zero termperature then corresponds to a non-Bayesian formulation, while a temperature of 1 corresponds to a pure Bayesian strategy. The Bayesian perspective can be accommodating in this respect.

7 Bayesian vs Non-Bayesian Solutions

Due to the entropy term, the Bayesian solutions fundamentally differ from non-Bayesian ones that do not include the entropy. But, the good news is that the non-Bayesian solutions can be obtained as a special case, which makes the Bayesian approach strictly more general (and again accommodating). To appreciate this viewpoint, we will now resort to exponential-family approximations of the posterior, by restricting the optimization to take place only over a subclass of distributions.

$$\min_{q \in \mathcal{Q}} \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_i \ell_i(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}) \right] - \mathcal{H}(q), \text{ where } q(\boldsymbol{\theta}) \propto \exp\left[\langle \boldsymbol{\lambda}, \boldsymbol{T}(\boldsymbol{\theta}) \rangle - A(\boldsymbol{\lambda})\right] \tag{2}$$

The audience unfamiliar with exponential-family may simply think of a Gaussian distribution with a mean and covariance.

Every exponential-family distribution is associated with a 'natural' parameterization and an 'expectation' parametrization (the latter is also known as the moment). The spaces of two parameterizations are 'duals', in this case, connected through 'Legendre Transform'. This simply mean that, one parameterization is equal to the derivative of a convex function taken with respect to another parameterization, and this operation is invertible.

$$\boldsymbol{\mu} = \nabla_{\boldsymbol{\lambda}} A(\boldsymbol{\lambda}), \qquad \boldsymbol{\lambda} = \nabla_{\boldsymbol{\mu}} A^*(\boldsymbol{\mu}), \tag{3}$$

The Bayesian solution always consist of such maps between the two spaces.

We can now state the result: the optimal posterior-approximation is the one with natural-parameter equal to the gradient of the expected loss, taken with respect to the corresponding expectation-parameter,

$$\boldsymbol{\lambda}_* = \nabla_{\boldsymbol{\mu}} \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_i \ell_i(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}) \right] \Bigg|_{\boldsymbol{\mu}=\boldsymbol{\mu}_*}. \tag{4}$$

The gradient here is computed in the dual space (expectation parameter space), which also turns out to be equal to a special type of gradient in the primal space, called the *natural-gradient*. Natural gradients are simply the vanilla gradients scaled by Fisher-information matrix,

$$\nabla_{\boldsymbol{\mu}} = \boldsymbol{F}(\boldsymbol{\lambda})^{-1} \nabla_{\boldsymbol{\lambda}} \tag{5}$$

The scaling adjust the gradient according to the 'information-geometry' of the distribution. It may not seem important right away, and also far from obvious, but this relationship arises in virtually *all* Bayesian computations. For example, in exact posterior computation, the 'exponential-weighting' is due to the dual of the entropy.

$$p^*(\boldsymbol{\theta}) \propto e^{-\left[\sum_i \ell_i(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta})\right]} \qquad \xrightarrow{\text{Posterior Approximation}} q^*(\boldsymbol{\theta}) \propto e^{\langle -\nabla_{\boldsymbol{\mu}} \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_i \ell_i(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta})\right]\big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_*}, \boldsymbol{T}(\boldsymbol{\theta}) \rangle} \tag{6}$$

The result shown in Eq. 4 is the main takeaway of this note. All the connections follow due to this equation.

How? We will give an intuitive explanation here; the details and illustrative examples are in Khan and Rue [2021]. Eq. 4 has an 'information-matching' interpretation,

1. Natural-gradients in Eq. 4 contain essential 'higher-order' derivatives of the loss.

2. This information is assigned to the appropriate natural parameters.

This information matching is the main result to connect Bayesian and non-Bayesian solutions. Its power is demonstrated in Khan and Rue [2021] by deriving many ML algorithms as a special instance of a Bayesian algorithm.

References

E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982. doi: 10.1109/PROC.1982.12425.

M. E. Khan and H. Rue. Bayesian learning rule. *arXiv preprint arXiv:2107.04562*, 2021.

M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2611–2620, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/khan18a.html`.

W. Lin, F. Nielsen, M. E. Khan, and M. Schmidt. Tractable structured natural-gradient descent using local parameterizations. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.