Bayesian approach for Neural Networks.

García, Salvador
s1655274

January 2016

Contents

Abstract

The concept of *uncertainty* in statistics is an important topic that is covered widely in books and papers. However, in the machine learning literature, this uncertainty only is considered implicitly when the final model is selected, but it is neither reported nor measured. To give a solution to this problem, a Bayesian approach for Neural Networks will be given. The basic idea is to use a Gaussian prior distribution over each weight in the model, then the posterior distribution of the weights and the predictive distribution will be computed.

As it is known, the Bayesian computations are not easy and specific methods are required for each problem. In this review, three general methodologies for estimate the Bayesian integrals will be considered: *Variational Inference (VI)*, *Laplace Approximation (LA)* and *Markov Chain Monte Carlo (MCMC)*. After that, a novel approach will be introduced, this is based on the ideas of the VI methods, but some approximations will be made with MCMC. Then, two of these algorithms will be explained briefly [8] [16]. To test the models, examples with the public MNIST dataset and in deep reinforcement learning will be made; while, for the second one, the a Gaussian bivariate posterior distribution will be [12] will be used.

1 Introduction

This review will be centered around the topic of Bayesian Neural Networks (BNNs) and in a recent approach to give a solution to its inherent computation problems, especially those that are a consequence of the Bayesian framework. As the first part of this introduction, a summary of empirical models in natural and social sciences will be given. Similarly, these models will be classified by with two approaches in the machine learning literature: statistical models and neural networks models. In the second part, the importance of uncertainty in the model selection stage will be explained, primordially with two different examples in the AI field. Also, an analysis of uncertainty under statistical and neural networks models will be given, indicating deficiencies on each of the methods. In the last part of this introduction, a solution to the uncertainty calculation in the context of neural networks will be proposed [8] [16] [4].

All the natural and social sciences rely on empirical stimulus acquired through different situations or phenomena. As a consequence, questions began to arise to understand and explain the relations and interactions in our world. As one way to give rational answers to the questions, a range of models is proposed in each field of study. These are different in functionalities, some of them just look for a good forecast; others, on the other hand, seek for explanation and causal relations. It is important to have a robust theory framework supporting each family of models and, also, to understand the intrinsic properties and advantages of each one of them deeply. This importance relies on the fact that many sciences like biology, chemistry, medicine, economics and politics use every day this theory to explain the phenomena corresponding to each field of study.

The families of models are diverse, depending primarily on the area of study where they were created. For example, the ones that are born in the field of statistics are mainly based on measure theory, specifically in probability theory. Then, most of the family have a simple structure and can be easily interpretable, but relies heavily on assumptions of the data [7]. The fulfillment of the assumptions comes with a big reward: the calculus of uncertainty is straightforward. For example, in the case of the ordinary linear regression, it is assumed that the errors of the model are normally distributed, in addition to the correct model specification, the independence of the errors and the homoscedasticity of the errors. Then, the theoretical distribution of the response (associated

with the uncertainty) can be described immediately [7]. If the assumptions are not accomplished, then the distribution of the response does not follow the theoretical one. Other examples of this statistical approach are the complete family of Generalized Linear Models (Logistic regression, Poisson regression and log-linear regression to mention some).

On the other hand, models born in the field of computer science have, in general, better predictions, but with a complex structure that cannot be interpreted so quickly [2] [1]. A big problem treated in this review is that in these models the uncertainty is not modeled directly. A solution for the neural networks models is to place Gaussian prior distribution over each weight in the hidden layers. These networks, depending on the number of parameters on the distribution, receive different names. Neural networks with a finite number of parameters are called Bayesian Neural Networks [3]. On the other hand, the ones that have infinite parameters and Gaussian prior distributions can be approximated through Gaussian processes [[14]] [2].

Now, given the families of alternative families of models, how to select the best model that considers the uncertainty? Typically, a model is preferred because of its accuracy [3]. For example, it is common that a classification model is preferred over another if it classifies more items correctly. However, the precision of the model is not considered. This precision is what is called *Uncertainty* and two example will be given to remark the important of calculating it. It is critical to mention that, if an approach of train-test-validation sets is used, then the uncertainty of the prediction is considered to select the best model, but it does not directly calculate how *certain* is the given solution. Methods to calculate precise estimates are the main topic in this review.

The importance of uncertainty in the models is presented with different examples in the papers and books considered. For example, [8] explains how the unawareness of the uncertainty can yield to endanger human life, specifically with the *Automated cancer detection based on MRI* [8]. As a brief explanation, it states that if the selected model just give a probability to the person expert in the topic (without a measure of uncertainty), it can bias his judgment of

[1] With a mathematical and optimization basis [13]

[2] In practice, just is modeled a network with a large number of parameters

[3] this accuracy can be measured with different metrics, represented by a loss function or a reward function, like the quadratic loss or the 0-1 loss

the case, because he will trust equally all the probabilities given by the model; although, some of them were made almost at random. On the other hand, if the probability and the variance are given, then the expert can say if the prediction is reliable and can use it for the diagnosis. The second example is given by the same author, but now in the context of *autonomous cars*. This is related to the failure of an autonomous car in 2016, where the car failed to differentiate between a white side of a bus and the sky [1]. If the system in the vehicles detects an ambiguous decision in a certain scenario, then it would be useful to ask the user of the car to take explicit control of the problem [8].

In this review, the feasibility of different approaches to calculating the uncertainty in the context of Bayesian neural networks will be examined. Additionally, the advantages and disadvantages of each one of them will be discussed. As an introduction to the problem, three Bayesian difficulties associated with the calculation of integrals will be given. After that a brief explanation of the common solutions for this integration problem will be given: Laplace Approximation (LA) [5], Variational Inference (VI) [4] and Markov Chain Monte Carlo Methods (MCMC) [4]. Then a fourth approach introduced in recent papers [16] [8] will be proposed; which, are based on VI models, but use stochastic steps (based on MCMC approach). [4] This new family of solutions claims to bring together the accuracy of the MCMC models and the speed of the VI methods [16], but there is still much work to do.

The structure of this review is divided into four parts. The first part studies the Bayesian neural networks models and the computational problem associated with the topic. The second explains the three classic approaches to give a solution to the problem: LA, VI, and MCMC. These procedures have been expanded to create a large amount of papers and books related to them, so just a brief review will be given. The third will include the center part of this review, the variational inference with stochastic steps. The fourth states the results of the new models based on the studied papers. At last critic ideas and ideas for further research will be given.

[4] This is the main reason to include the explanation of the LA, VI and MCMC techniques.

2 Bayesian neural networks

As claimed in [8], the Bayesian neural networks are based on Bayesian modeling; this ha the advantage to learn even if there is few data, trusting in the priors. Besides, these Bayesian neural networks models can also have an implicit regularization, like all the Bayesian models. Also, as suggested in [5], is a very versatile method, mainly because of the specification of the prior distributions. In fact, some of them have been studied; for example, taking a gamma prior distribution [6]. These two topics are significant areas inside the Bayesian framework, but will not be covered in this review.

To understand the problem of the Bayesian framework, a brief review of the theory associated with this methodology will be given.

2.1 The problem

First, the widely known Bayes theorem is defined:

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{1}$$

With θ the parameters of the model, X the data, $P(\theta)$ the prior probability of θ, $P(\theta|X)$ the posterior probability of θ, and $P(X|\theta)$ as the likelihood of X given θ.

When taking a Bayesian approach to any problem, the computations can become quite difficult and often intractable with analytical methods [4]. As an example, three difficulties will be shown, the first is associated with the calculus of the posterior distribution of θ, the second with the computation of the predictive distribution of X, and at last the calculation of an integral that arises when making the model selection. This is also the case in the computation of the Bayesian Neural Networks.

2.1.1 Calculation of the posterior distribution of θ

As seen in eq. (1), the denominator contains the expression $P(X)$. Using a basic probability property, it is equal to:

$$P(X) = \int_{\Theta} P(X, \theta) d\theta \tag{2}$$

2.1.2 Calculation of the predictive distribution of X^*

The predictive distribution of X^* (a new observation) includes in its formula the posterior distribution of θ and is as follows [4]:

$$P(X^*|X) = \int_{\Theta} P(X^*|\theta, X) P(\theta|X) d\theta \tag{3}$$

2.1.3 Model selection

The last example of the integration difficulties arises in the model selection stage:

$$P(M|X) = \frac{P(X|M)P(M)}{P(X)} \tag{4}$$

With M the selected model. The denominator of the eq. (4) is like the equation eq. (1), thus it can be written as follows:

$$P(X) = \int_{\theta} P(X|\theta, M) P(\theta|M) d\theta \tag{5}$$

The eq. (2), eq. (3) and eq. (5) are known as marginalization and, sometimes, are difficult to calculate. Hence, it is important to find a method to estimate them.

3 Traditional methods

As mentioned before, three main approaches arose to give a solution to these integrals; the first is a straightforward method known as the Laplace Approximation (LA). This approximates the original distribution with a Gaussian distribution [5]. The second approach is a family of methods under the name of Variational Inference (VI) [4] which tries to approximate the function with an optimization framework. The third approach is the Markov Chain Monte Carlo

methods (MCMC) [4] which are subsampling methods that try to approximate the original integral with stochastic simulations.

Although these approaches are the traditional ones, the area is in constant development. For example, some books that were published ten years ago [4] does not mention the interaction between these ideas, and considerate each one of them as entirely different, one deterministic and the other stochastic. As a contrast, papers from last two years have begun to give a solution involving the Variational inference with stochastics steps (MCMC) [16] [8]. This is the final approach that will be presented in this review.

In this section the notation used by [4] will be used. In general, for the following techniques, it is supposed that the original distribution $p(\theta|X)$ (the posterior distribution) is equal to:

$$p(\theta|X) = \frac{1}{Z} f(\theta|X) \tag{6}$$

Where the function $f(\theta|X)$ can be evaluated [5], however, the normalization constant Z is unknown.

3.1 Laplace approximation

Laplace approximation is naive solution for the problem. The main idea is to fit a Gaussian distribution to the original one $p(\theta|X)$ at the same time that it preserves the same mode of the original. As [4] proposed, find a gaussian $q(\theta|X)$ that is centred on a maximum value of the original $p(\theta|x)$. The method is very fast, and straightforward, but if the distribution is bimodal, or have multiple maximum values, then the approximation could not be good (as the normal is unimodal). The next equations express this idea (remember that the value of the posterior distribution $p(\theta|X)$ cannot be evaluated, but it is possible to evaluate $f(\theta|X)$). Remembering that Taylor expansion of a function $h(x)$ around the point a is equal to [4]:

$$\sum_{n=0}^{\infty} \frac{h^{(n)}(a)}{n!}(x-a)^n \tag{7}$$

[5] Because it can be evaluated, then it can be *sampled*, so this is the intuition of the MCMC.

To simplify the notation for this section, lets define $f(z) = f(\theta|X)$ and $q(z) = q(\theta|X)$ Then, making the second order Taylor expansion of $\ln(f(z))$ around z_0 (consider that, as z_0 is the mode of the posterior distribution, then $f^{(1)}(z_0)$ is equal to 0) [4]:

$$\ln(f(z)) \approx ln(f(z_0)) + \frac{1}{2}(z - z_0)^T H(z - z_0) \tag{8}$$

With H the Hessian matrix. Then, is easy to see that applying the exp function, we can get the kernel of a Gaussian. Now, with this distribution is easy to calculate the normalization constant; therefore, the posterior probability distribution is completely defined. This is a quick and easy approach to solving this solution (For the calculation of the Hessian, direct or iterative procedures can be done, although for big datasets can be problematic).

3.2 Variational Inference (VI)

This family of methods is related to an approximation with a probability distribution $q(\theta)$ of the original one $p(\theta|X)$. The difference with the Laplace approximation is that this estimate uses the Jensen inequality and then tries to maximize the variational lower bound, or equivalently, minimize the KullBack-Leibler divergence of the two distributions. [8]. The variety of methods included in this family is wide, and a chapter included in the book [4].

The main advantage of this family of methods is that have an explicit objective function and sometimes have analytical solutions [16]. The method is very attractive because the original problem now is transformed into an optimization problem, where there are many ways to solve it. For example, the book [15] contains a vast amount of methods to work with optimization problems. Also, there are optimization tools that have been intensely studied for problems when the quantity of data is enormous.

One of most common method of variational inference is the one that minimizes the Kullback-Leibler divergence of two functions:

$$KL(q(x)|p(x)) = \int_X q(x) log \frac{q(x)}{p(x)} \tag{9}$$

Applied to this problem:

$$\begin{aligned} KL(q(\theta)|p(\theta|X)) &= \int_X q(\theta) log \frac{q(\theta)}{p(\theta|X)} \\ &= \int_X q(\theta) log \frac{q(\theta)}{p(\theta, X)p(X)} \\ &= \int_X q(\theta) log \frac{q(\theta)}{p(\theta, X)} + log(p(X)) \end{aligned} \tag{10}$$

Then, this issue is equivalent to minimize the first term on the right side of the above equation. (because the second does not depend on $q(\theta)$). Also, rearranging the terms, a maximization problem can be stated.

$$\begin{aligned} log(P(X)) =& KL(q(\theta)|p(\theta|X)) - \int_X q(\theta) log \frac{q(\theta)}{p(\theta, X)} \\ =& KL(q(\theta)|p(\theta|X)) + \mathcal{L}(q) \end{aligned} \tag{11}$$

With the term $\mathcal{L}(q) = -\int_X q(\theta) log \frac{q(\theta)}{p(\theta,X)}$ often known as *Variational lower bound*. This way, the problem can also be treated as a maximization one.

$$\begin{aligned} \mathcal{L}(q) =& log(P(X)) - KL(q(\theta)|p(\theta|X)) \\ =& -\int_X q(\theta) log \frac{q(\theta)}{p(\theta, X)} \end{aligned} \tag{12}$$

If the Jensen inequality is applied to the above equations, then it is easy to find the boundaries. Remembering, the Jensen inequality consist in one function and the expectation, in this example the function is log, and the integral is used as the expectation. One way to find a solution to these optimization problems is the idea to restrict the family of the approximate distributions $q(\theta)$. A common technique is named *mean-field variational Bayes* [4]:

$$q(\theta) = \prod_1^M q_i(\theta_i) \tag{13}$$

This will be discussed intensely in [4], but the above formulation is useful for the methods proposed in the next section.

3.3 Markov Chain Monte Carlo (MCMC)

This family of methods, along with the Variational Inference family are the most common approaches. [16]. The difference with the VI is that MCMC is a subsampling method, so it approximates the distribution stochastically. The method starts with a random variable z and applies a *stochastic transition operator* that just depends on the value of the same variable, but one moment in the past (Markov property) [16]:

$$z_t = q(z_t|z_{t-1}, x) \tag{14}$$

Then, if correct stochastic transition operator is selected, the distribution of z will converge to the original distribution $p(z|x)$. [16]. The main advantage of MCMC is the fact that it guarantees that it converges to the original distribution, although there is not a specific number of iterations to guarantee a convergence [16]. This topic still in investigation.

4 Stochastic Variational Inference methods

The optimization problem related to the VI methods is complicated and only is easy to find an analytical solution in some cases [8]. Although an approximation is found, this can be bad due to the correlation structure between the weights [8]. Solutions began to arise to solve this problem. In 1993, a diagonal covariance matrix between the weights was considered for this issue [11], but some years later an approach with a full covariance matrix was achieved [6]. It is important to mention that now, this approach is almost impossible for neural networks that have millions of neurons because the covariance matrix calculation could be huge.

In 2011, [10] began to use another solution to the huge amount of data. With more data, calculate the likelihoods in the problems started to be very expensive. Therefore, he proposed subsampling methods to estimate it and improve the VI method. Now, the model could use a large quantity of data. [8]. Two recent approaches following this idea are described.

The key idea of [16] is to make a stochastic approximation of the function $q(\theta)$

used in the variational inference framework. This way, the approximation can be calculated like:

$$q(\theta_T|x) = q(\theta_0|x)\prod_{t=1}^{T} q(\theta_t|\theta_{t-1}, x) \tag{15}$$

All the variables θ_i with $i = 1, ..., n-1$ are named *auxiliary random variables*. The set of all of this variables will be called γ. now, the distribution can be formulated as follows [16]:

$$q(\theta_T|x) = \int q(\gamma, \theta_T|x)d\gamma \tag{16}$$

Therefore, the distribution now can be interpreted as a mixture of distributions [16]. Also he introduces an auxiliary function $r(\gamma|x, z_T)$ (the inverse model) that is used to approximate the transition operator $q(\theta_T|x)$. The way that he proposes is to let be the function $r(\gamma|x, z_T)$ flexible, then optimize it on its parameters. To finish this approach, he proposes that this auxiliary function also follows a Markov property.

On the other hand, [8] proposes another technique that also uses a stochastic process. The central idea is not to evaluate the complete expression of the likelihood with all the dataset, but just approximate the expression $\mathcal{L}(\theta)$. This way, instead of computing the complete $\mathcal{L}(\theta)$ in each iteration of the optimization procedure, just an approximation $\widehat{\mathcal{L}}(\theta)$ is used. This idea is usual in the optimization field, reduce the cost of computation of the function in each iteration, with the cost of more iterations.

The ideas proposed in the papers just try to find an intermediate step between the MCMC and the Variational inference methods. The idea of [8] is very common in optimization field; for example, when the Newton method is used for big datasets, it is not common to use the second derivative in the descent method. Instead of that, they make each iteration faster, at the cost of more iterations. Adapting this example to the trade-off of the approximations, now they want to get some of the speed of VI, but with an accuracy given stochastic approximation provided by the MCMC.

For testing the first method, the author proposed to fit a bivariate Gaussian. In the example, as the posterior is known, the explicit lower bound is calculated.

Then a standard MCMC is compared with an approximation proposed, and it is shown that converges faster. On the other hand, for the second one, the MNIST dataset was used. Then an approximation regarding a mixture of Bernoulli is implemented. The best error achieved was a little less than 1.7%, that is close to the ranges reported in the MNIST web page.

5 Conclusions and further research

The purposes of the methods presented in the review are to find a technique that can scale with more data, and that can adapt to more complex models [8]. With the MCMC approach, many posterior distributions can be found regardless the analytical form. This is a very recent topic. Consequently, there are just a few papers that talk about the problem and leave many questions open. Some of them, as commented at the beginning, involve the specification of different prior distribution and techniques about regularization.

The second most important conclusion, as [8] suggest, is that it is necessary to stop treating the Bayesian approach entirely different from the one given by the neural networks. Both can be used together and create models that work better. Not only the neural networks benefit from the Bayesian framework, but also Bayesian modeling can help from neural networks models, using the transformed data that the neural networks make. For example, there are statistical techniques like PCA or LDA that contributes to reducing dimensionality, but there are neural networks approaches can help to find better representations of the data for the problem or the task.

Another approach that worth to explore, but it was not investigated, was the approach of neural networks with infinite weights. As [14] says, this approach can be like the one given by a Gaussian process. So, it worth further investigation on this topic. Lastly, one weakness of the papers is that there are no universal criteria to compare the models. For example, in the classical Bayesian literature, the BIC is usually used to select the model [9] (Although it is stated that is just a convention). However, here, each paper evaluates the model with different metrics.

References

[1] AP and REUTERS. Tesla working on 'improvements' to its autopilot radar changes after model s owner became the first selfdriving fatality. `http://www.dailymail.co.uk/news/article-3693494/Tesla-working-autopilot-radar-changes-self-driving-fatality-men-watching-Harry-Potter-video.html`, 2016.

[2] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[3] Christopher M Bishop. Bayesian neural networks. *Journal of the Brazilian Computer Society*, 4(1), 1997.

[4] Christopher M Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[5] Wray L Buntine and Andreas S Weigend. Bayesian back-propagation. *Complex systems*, 5(6):603–643, 1991.

[6] Christopher Bishop D. Barber. Ensemble learning in bayesian neural networks. In *Generalization in Neural Networks and Machine Learning*, page 215–237. Springer Verlag, January 1998.

[7] Norman Richard Draper, Harry Smith, and Elizabeth Pownell. *Applied regression analysis*, volume 3. Wiley New York, 1966.

[8] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

[9] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[10] Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pages 2348–2356, 2011.

[11] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. ACM, 1993.

[12] Yann LeCun, Corinna Cortes, and Christopher JC Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2010.

[13] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[14] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[15] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[16] Tim Salimans, Diederik P Kingma, Max Welling, et al. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226, 2015.