ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Sensitivity Analysis for Predictive Uncertainty in Bayesian Neural Networks

Stefan Depeweg[1,2], José Miguel Hernández-Lobato[3], Steffen Udluft[2], Thomas Runkler[1,2]

1 - Technical Unversity of Munich, Germany. 2 - Siemens AG, Germany.
3 - University of Cambridge, United Kingdom.

Abstract. We derive a novel sensitivity analysis of input variables for predictive epistemic and aleatoric uncertainty. We use Bayesian neural networks with latent variables as a model class and illustrate the usefulness of our sensitivity analysis on real-world datasets. Our method increases the interpretability of complex black-box probabilistic models.

1 Introduction

Extracting human-understandable knowledge out of black-box machine learning methods is an important topic of research. One aspect of this is to figure out how sensitive the model response is to which input variables. This can be useful both as a sanity check, if the approximated function is reasonable, but also to gain new insights about the problem at hand. For neural networks this kind of model inspection can be performed by a sensitivity analysis [1, 2], a simple method that works by considering the gradient of the network output with respect to the input variables.

Our key contribution is to transfer this idea towards predictive uncertainty: What features impact the uncertainty in the predictions of our model? To that end we use Bayesian neural networks with latent variables [3, 4], a recently introduced probabilistic model that can describe complex stochastic patterns while at the same time account for model uncertainty. From their predictive distributions we can extract epistemic and aleatoric uncertainties [5, 4]. The former uncertainty originates from our lack of knowledge of model parameter values and is determined by the amount of available data, while aleatoric uncertainty consists of irreducible stochasticity originating from unobserved (latent) variables. By combining the sensitivity analysis with a decomposition of predictive uncertainty into its epistemic and aleatoric components, we can analyze which features influence each type of uncertainty. The resulting sensitivities can provide useful insights into the model at hand. On one hand, a feature with high epistemic sensitivity suggests that careful monitoring or safety mechanisms are required to keep the values of this feature in regions where the model is confident. On the other hand, a feature with high aleatoric uncertainty indicates a dependence of that feature with other unobserved/latent variables.

2 Bayesian Neural Networks with Latent Variables

In this section we review a recent family of flexible probabilistic models for multi-output regression. These models were previously introduced by [3] and we refer to them as Bayesian Neural Networks with latent variables (BNN+LVs).

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Given data $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N$, formed by feature vectors $\mathbf{x}_n \in \mathbb{R}^D$ and targets $\mathbf{y}_n \in \mathbb{R}^K$, we assume that $\mathbf{y}_n = f(\mathbf{x}_n, z_n; \mathcal{W}) + \boldsymbol{\epsilon}_n$, where $f(\cdot, \cdot; \mathcal{W})$ is the output of a neural network with weights $\mathcal{W}$ and K output units. The network receives as input the feature vector $\mathbf{x}_n$ and the latent variable $z_n \sim \mathcal{N}(0, \gamma)$. We choose rectifiers, $\varphi(x) = \max(x, 0)$, as activation functions for the hidden layers and and the identity function, $\varphi(x) = x$, for the output layer. The network output is corrupted by the additive noise variable $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with diagonal covariance matrix $\boldsymbol{\Sigma}$. The role of the latent variable z_n is to capture unobserved stochastic features that can affect the network's output in complex ways. The network has L layers, with V_l hidden units in layer l, and $\mathcal{W} = \{\mathbf{W}_l\}_{l=1}^L$ is the collection of $V_l \times (V_{l-1} + 1)$ weight matrices. The $+1$ is introduced here to account for the additional per-layer biases. We approximate the exact posterior $p(\mathcal{W}, \mathbf{z} \,|\, \mathcal{D})$ with:

$$q(\mathcal{W}, \mathbf{z}) = \underbrace{\left[\prod_{l=1}^{L}\prod_{i=1}^{V_l}\prod_{j=1}^{V_{l-1}+1} \mathcal{N}(w_{ij,l} | m_{ij,l}^w, v_{ij,l}^w)\right]}_{q(\mathcal{W})} \times \underbrace{\left[\prod_{n=1}^{N} \mathcal{N}(z_n \,|\, m_n^z, v_n^z)\right]}_{q(\mathbf{z})} . \quad (1)$$

The parameters $m_{ij,l}^w$, $v_{ij,l}^w$ and m_n^z, v_n^z are determined by minimizing a divergence between $p(\mathcal{W}, \mathbf{z} \,|\, \mathcal{D})$ and the approximation q. The reader is referred to the work of [6, 3] for more details on this. In our experiments, we tune q using black-box α-divergence minimization with $\alpha = 1.0$.

2.1 Uncertainty Decomposition

BNNs+LVs can capture complex stochastic patterns, while at the same time account for model uncertainty. They achieve this by jointly learning $q(\mathbf{z})$, which describes the values of the latent variables that were used to generate the training data, and $q(\mathcal{W})$, which represents uncertainty about model parameters. The result is a flexible Bayesian approach for learning conditional distributions with complex stochasticity, e.g. bimodal or heteroscedastic noise [3].

The predictive distribution of a BNN+LVs for the target variable $y_\star$ associated with the test data point $\mathbf{x}_\star$ is

$$p(y_\star | \mathbf{x}_\star) = \int p(y_\star | \mathcal{W}, \mathbf{x}_\star, z_\star) p(z_\star) q(\mathcal{W}) \, dz_\star \, d\mathcal{W} . \quad (2)$$

where $p(y_\star | \mathcal{W}, \mathbf{x}_\star, z_\star) = \mathcal{N}(y_\star | f(\mathbf{x}_\star, z_\star; \mathcal{W}), \boldsymbol{\Sigma})$ is the likelihood function, $p(z_\star) = \mathcal{N}(z_\star | 0, \gamma)$ is the prior on the latent variables and $q(\mathcal{W})$ is the approximate posterior for $\mathcal{W}$ given $\mathcal{D}$. In this expression the integration with respect to $z_\star$ must be done using the prior $p(z_\star)$. The reason for this is that the $y_\star$ associated with $\mathbf{x}_\star$ is unknown and consequently, there is no other evidence on $z_\star$ than the one coming from $p(z_\star)$.

In Eq. (2), the randomness or uncertainty on $y_\star$ has its origin in $\mathcal{W} \sim q(\mathcal{W})$, $z_\star \sim p(z_\star)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This means that there are two types of uncertainties entangled in our predictons for $y_\star$: aleatoric and epistemic [7, 5]. The aleatoric uncertainty originates from the randomness of $z_\star$ and ϵ and cannot be

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

reduced by collecting more data. By contrast, the epistemic uncertainty originates from the randomness of $\mathcal{W}$ and can be reduced by collecting more data, which will typically shrink the approximate posterior $q(\mathcal{W})$.

We can use the variance $\sigma^2(y_k^\star|\mathbf{x}^\star)$ as a measure of predictive uncertainty for the k-th component of $\mathbf{y}^\star$. The variance can be decomposed into an epistemic and aleatoric term using the law of total variance:

$$\sigma^2(y_k^\star|\mathbf{x}^\star) = \sigma^2_{q(\mathcal{W})}(\mathbf{E}_{p(z^\star)}[y_k^\star|\mathcal{W},\mathbf{x}^\star]) + \mathbf{E}_{q(\mathcal{W})}[\sigma^2_{p(z^\star)}(y_k^\star|\mathcal{W},\mathbf{x}^\star)] \tag{3}$$

The first term, that is $\sigma^2_{q(\mathcal{W})}(\mathbf{E}_{p(z^\star)}[y_k^\star|\mathcal{W},\mathbf{x}^\star])$ is the variability of $y_k^\star$, when we integrate out $z^\star$ but not $\mathcal{W}$. Because $q(\mathcal{W})$ represents our belief over model parameters, this is a measure of the *epistemic* uncertainty. The second term, $\mathbf{E}_{q(\mathcal{W})}[\sigma^2_{p(z^\star)}(y_k^\star|\mathcal{W},\mathbf{x}^\star)]$ represents the average variability of $y_k^\star$ not originating from the distribution over model parameters $\mathcal{W}$. This measures *aleatoric* uncertainty, as the variability can only come from the latent variable $z^\star$.

3 Sensitivity Analysis of Predictive Uncertainty

We want to extend the method of sensitivity analysis toward predictive uncertainty: how much does each feature affect each type of uncertainty? Answers to this question can provide useful insights about a model at hand. For instance, a feature with high aleatoric sensitivity indicates a strong interaction with other unobserved/latent features. If a practitioner can expand the set of features by taking more refined measurements, it may be advisable to look into variables which may exhibit dependence with that feature and which may explain the stochasticity in the data. Furthermore, a feature with high epistemic sensitivity, suggests careful monitoring or extended safety mechanisms are required to keep this feature values in regions where the model is confident.

We start by briefly reviewing the technique of sensitivity analysis [1, 2], a simple method that can provides insight into how changes in the input affect the network's prediction. Let $\mathbf{y} = f(\mathbf{x};\mathcal{W})$ be a neural network fitted on a training set $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N$, formed by feature vectors $\mathbf{x}_n \in \mathbb{R}^D$ and targets $\mathbf{y}_n \in \mathbb{R}^K$. We want to understand how each feature i influences the output dimension k. Given some test data $\mathcal{D}_{\text{test}} = \{\mathbf{x}_n^\star, \mathbf{y}_n^\star\}_{n=1}^{N_{\text{test}}}$, we use the partial derivate of the output dimension k w.r.t. feature i:

$$I_{i,k} = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \Big| \frac{\partial f(\mathbf{x}_n^\star)_k}{\partial x_{i,n}^\star} \Big| . \tag{4}$$

In Section 2.1 we saw that we can decompose the variance of the predictive distribution of a BNN with latent variables into its epistemic and aleatoric components. Our goal is to obtain sensitivities of these components with respect to the input variables. For this we use a sampling based approach to approximate the two uncertainty components [4] and then calculate the partial derivative of these w.r.t. to the input variables. For each test data point $\mathbf{x}_n^\star$, we perform $N_w \times N_z$ forward passes through the BNN. We first sample $w \sim q(\mathcal{W})$ a total of N_w times and then, for each of these samples of $q(\mathcal{W})$, performing N_z forward

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

passes in which w is fixed and we only sample the latent variable z. Then we can do an empirical estimation of the expected predictive value and of the two components on the right-hand-side of Eq. (3):

$$\mathbf{E}[y^\star_{n,k}|\mathbf{x}^\star_n] \approx \frac{1}{N_w}\frac{1}{N_z}\sum_{n_w=1}^{N_w}\sum_{n_z=1}^{N_z} y^\star_{n_w,n_z}(\mathbf{x}^\star_n)_k \tag{5}$$

$$\sigma_{q(\mathcal{W})}(\mathbf{E}_{p(z^\star)}[y^\star_{n,k}|\mathcal{W},\mathbf{x}^\star_n]) \approx \hat{\sigma}_{N_w}(\frac{1}{N_z}\sum_{n_z=1}^{N_z} y^\star_{n_w,n_z}(\mathbf{x}^\star_n)_k) \tag{6}$$

$$\mathbf{E}_{q(\mathcal{W})}[\sigma^2_{p(z^\star)}(y^\star_{n,k}|\mathcal{W},\mathbf{x}^\star_n)]^{\frac{1}{2}} \approx \big(\frac{1}{N_w}\sum_{n_w=1}^{N_w}\hat{\sigma}^2_{N_z}(y^\star_{n_w,n_z}(\mathbf{x}^\star_n)_k)\big)^{\frac{1}{2}}. \tag{7}$$

where $y^\star_{n_w,n_z}(\mathbf{x}^\star_n)_k = f(\mathbf{x}^\star_n, z^{n_w,n_z}; \mathcal{W}^{n_w})_k$ and $\hat{\sigma}^2_{N_z}$ ($\hat{\sigma}^2_{N_w}$) is an empirical estimate of the variance over N_z (N_w) samples of z ($\mathcal{W}$). We have used the square root of each component so all terms share the same unit of $y^\star_{n,k}$. Now we can calculate the sensitivities:

$$I_{i,k} = \frac{1}{N_{\text{test}}}\sum_{n=1}^{N_{\text{test}}}\Big|\frac{\partial \mathbf{E}[y^\star_{n,k}|\mathbf{x}^\star_n])}{\partial x^\star_{i,n}}\Big| \tag{8}$$

$$I^{\text{epistemic}}_{i,k} = \frac{1}{N_{\text{test}}}\sum_{n=1}^{N_{\text{test}}}\Big|\frac{\partial \sigma_{q(\mathcal{W})}(\mathbf{E}_{p(z^\star)}[y^\star_{n,k}|\mathcal{W},\mathbf{x}^\star_n])}{\partial x^\star_{i,n}}\Big| \tag{9}$$

$$I^{\text{aleatoric}}_{i,k} = \frac{1}{N_{\text{test}}}\sum_{n=1}^{N_{\text{test}}}\Big|\frac{\partial \mathbf{E}_{q(\mathcal{W})}[\sigma^2_{p(z^\star)}(y^\star_{n,k}|\mathcal{W},\mathbf{x}^\star_n)]^{\frac{1}{2}}}{\partial x^\star_{i,n}}\Big|, \tag{10}$$

where Eq. (8) is the standard sensitivity term. We also note that the general drawbacks [2] of the sensitivity analysis, such as considering every variable in isolation, arise due to its simplicity. These will also apply when focussing on the uncertainty components.

4 Experiments

In this section we want to do an exploratory study. For that we will first use an artifical toy dataset and then use 8 datasets from the UCI repository [8] in varying domains and dataset sizes. For all experiments, we use a BNN with 2 hidden layer. We first perform model selection on the number of hidden units per layer from $\{20, 40, 60, 80\}$ on the available data. We train for 3000 epochs with a learning rate of 0.001 using Adam as optimizer. For the sensitivity analysis we will sample $N_w = 200$ $w \sim q(\mathcal{W})$ and and $N_z = 200$ samples from $z \sim \mathcal{N}(0, \gamma)$. All experiments were repeated 5 times and we report average results.

4.1 Toy Data

We consider a regression task for a stochastic function with heteroskedastic noise: $y = 7\sin(x_1) + 3|\cos(x_2/2)|\epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$. The first input variable x_1 is responsible for the shape of the function whereas the second variable x_2 determines

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

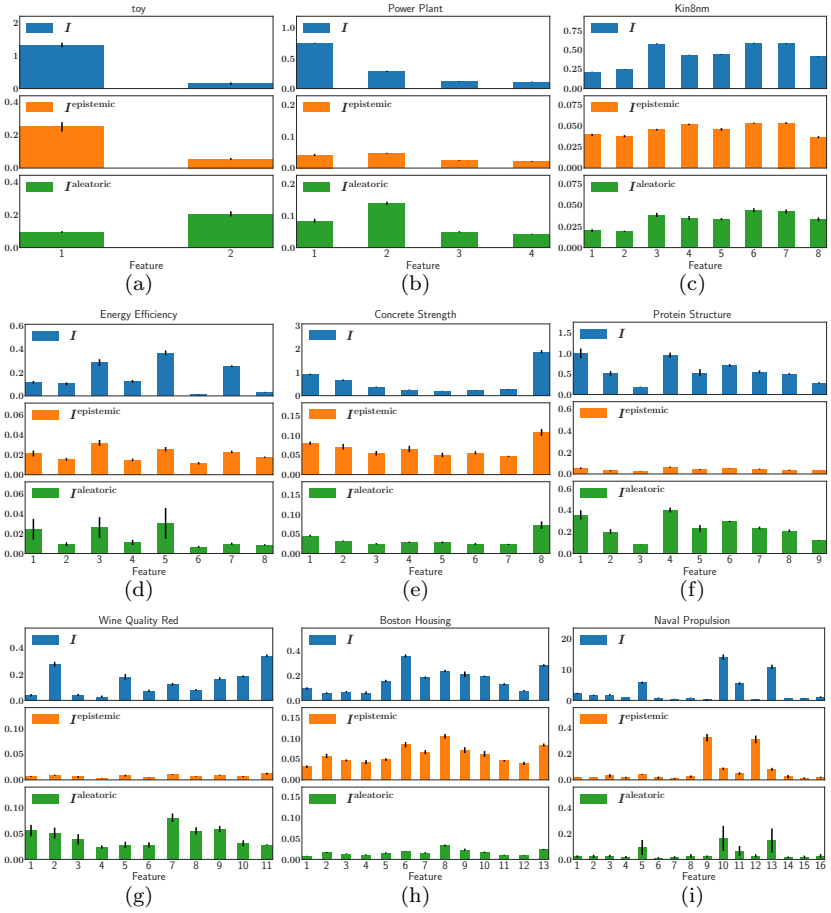

Fig. 1: Sensitivity analysis for the predictive expectation and uncertainty on toy data (a) and UCI datasets (b)-(i). Top row shows sensitivities w.r.t. expectation (Eq. (8)). Middle and bottom row show sensitivities for epistemic and aleatoric uncertainty (Eq. (9) and Eq. (10)). Error bars indicate standard errors.

the noise level. We sample 500 data points with $x_1 \sim \text{exponential}(\lambda = 0.5) - 4$ and $x_2 \sim \mathcal{U}(-4, 4)$. Fig. 1a shows the sensitivities. The first variable x_1 is responsible for the epistemic uncertainty whereas x_2 is responsible for the aleatoric uncertainty which corresponds with the generative model for the data.

4.2 UCI Datasets

We consider several real-world regression datasets from the UCI data repository [8]. Detailed descriptions can be found on the respective website. For evaluation we use the same training and test data splits as in [6]. In Fig. 1 we show the results of all experiments. For some problems the aleatoric sensitivity is most prominent (Fig. 1f,1g), while in others we have predominately epistemic

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

sensitivity (Fig. 1e,1h) and a mixture in others. This makes sense, because we have variable dataset sizes (e.g. Boston Housing with 506 data points and 13 features, compared to Protein Structure with 45730 points and 9 features) and also likely different heterogeneity in the datasets.

In the power-plant example feature 1 (temperature) and 2 (ambient pressure) are the main sources of aleatoric uncertainty of the target, the net hourly electrical energy output. The data in this problems originates from a combined cycle power plant consisting of gas and steam turbines. The provided features likely provide only limited information of the energy output, which is subject to complex combustion processes. We can expect that a change in temperature and pressure will influence this process in a complex way, which can explain the high sensitivities we see. The task in the naval-propulsion-plant example, shown in Fig. 1i, is to predict the compressor decay state coefficient, of a gas turbine operated on a naval vessel. Here we see that two features, the compressor inlet air temperature and air pressure have high epistemic uncertainty, but do not influence the overall sensitivity much. This makes sense, because we only have a single value of both features in the complete dataset. The model has learned no influence of this feature on the output (because it is constant) but any change from this constant will make the system highly uncertain.

5 Conclusion

In this paper we provided a new way of sensitivity analysis for predictive epistemic and aleatoric uncertainty. Experiments indicate useful insights of this method on real-world datasets.

References

[1] Li Fu and Tinghuai Chen. Sensitivity analysis for input vector in multilayer feedforward neural networks. In *Neural Networks, 1993., IEEE International Conference on*, pages 215–218. IEEE, 1993.

[2] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 2017.

[3] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Learning and policy search in stochastic dynamical systems with bayesian neural networks. *arXiv preprint arXiv:1605.07127*, 2016.

[4] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems. *arXiv preprint arXiv:1710.07283*, 2017.

[5] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[6] José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang Bui, and Richard E Turner. Black-box α-divergence minimization. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016.

[7] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105 – 112, 2009. Risk Acceptance and Risk Communication.

[8] Moshe Lichman. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2013.