arXiv:1810.02678v1 [cs.LG] 5 Oct 2018

Local Interpretable Model-agnostic Explanations of Bayesian Predictive Models via Kullback–Leibler Projections

Tomi Peltola[1]
[1] Helsinki Institute for Information Technology HIIT, Department of Computer Science,
Aalto University, Helsinki, Finland
tomi.peltola@aalto.fi

Abstract

We introduce a method, KL-LIME, for explaining predictions of Bayesian predictive models by projecting the information in the predictive distribution locally to a simpler, interpretable explanation model. The proposed approach combines the recent Local Interpretable Model-agnostic Explanations (LIME) method with ideas from Bayesian projection predictive variable selection methods. The information theoretic basis helps in navigating the trade-off between explanation fidelity and complexity. We demonstrate the method in explaining MNIST digit classifications made by a Bayesian deep convolutional neural network.

1 Introduction

Interpreting why a machine learning model makes a certain prediction for given inputs can be difficult for complex models, but is important for evaluating the trustworthiness of the predictions and for model criticism and development. Explaining a prediction requires understanding the proper context of the prediction (inputs and what kind of changes the inputs could meaningfully have) and how the prediction would react to changes in the context (the mapping from inputs to the prediction). Understanding non-linear predictive models globally would require understanding the underlying prediction function at all locations of the input space. Local interpretations are more feasible and many methods have been proposed for this (e.g., [Baehrens *et al.*, 2010; Ribeiro *et al.*, 2016; Lundberg and Lee, 2017]). They rely on the assumption that non-linear functions can be approximated by simpler, more interpretable functions locally.

In addition to interpretability, properly accounting for uncertainty in the predictions is important in many applications, for example, in medicine. In Bayesian probabilistic modelling, the posterior distribution of the model parameters (the conditional distribution induced by the modelling assumptions and training data) and posterior predictive distributions naturally capture the uncertainty about the parameters and model structure (epistemic uncertainty) in addition to noise in prediction (aleatoric uncertainty). When explaining predictions, these uncertainties should also be accounted for.

We introduce a method for Kullback–Leibler divergence based local interpretable model-agnostic explanations, KL-LIME, that extends the recently proposed local interpretation method LIME [Ribeiro *et al.*, 2016] to Bayesian models (although it can also be used for non-Bayesian probabilistic models) and provides a principled way to handle different types of predictions (continuous valued, class labels, counts, censored and truncated data, etc.). The proposed method is based on combining the LIME approach with methods introduced for variable selection in Bayesian linear regression models [Goutis and Robert, 1998; Dupuis and Robert, 2003; Peltola *et al.*, 2014; Piironen and Vehtari, 2017]. The method works by fitting an interpretable explanatory model (e.g., a sparse linear model) to locally match the original prediction model via minimizing the Kullback–Leibler divergence between their predictive distributions.

The next section gives a more detailed background on LIME and Bayesian projection predictive variable selection. Section 3 describes the proposed method, KL-LIME, and Section 4 demonstrates it in explaining deep convolutional neural network predictions. Section 5 presents a summary and conclusions.

2 Background

In this section, we describe the LIME method and projection predictive variable selection. In the next section, they are combined to extend LIME to Bayesian predictive models.

2.1 Local Interpretable Model-agnostic Explanations – LIME

The Local Interpretable Model-agnostic Explanations (LIME) method of Ribeiro et al. [2016] provides local explanations of predictions of a classifier f by fitting a simpler, interpretable explanation model g locally around the data point x of which classification is to be explained. The explanation model is fit on an interpretable representation of the original data space. For example, let $x \in \mathbb{R}^d$ be a vector of the gray scale values of pixels in an image. An interpretable representation $x' \in \{0,1\}^{d'}$ might then be a vector of binary values representing the absence or presence of pixels in the image (absence meaning having the value of a background color, e.g., white). The LIME explanation $\hat{g}$

arises by solving the optimization problem

$$\hat{g} = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g),$$

where G is the explanation model family, $\mathcal{L}$ is a loss function, π_x defines the locality around x, and Ω is a complexity penalty.

In practice, G is taken to be the set of linear regression models, with Ω restricting that only some number of the explanatory features can have non-zero regression weights (although other types of explanation models could be used). The loss function is taken to be the weighted L2 distance

$$\mathcal{L}(f, g, \pi_x) = \sum_i \pi_x(z_i)(f(z_i) - g(z_i'))^2,$$

where the sum goes over a set of sampled perturbed points around x, $\{(z_i, z_i'), i = 1, \ldots, m\}$, where z_i is a perturbed data point in the original data space and z_i' the corresponding interpretable representation. $\pi_x(z_i)$ weights the samples based on their similarity to x, the point where the classification result is being explained.

2.2 Projection predictive variable selection

Variable selection is the problem of choosing a smaller set of covariates from among the full set available and is often used to simplify high-dimensional regression models. In Bayesian modelling, sparsity-inducing and shrinkage priors can be used to regularize high-dimensional regression models. However, they do not lead to truly sparse posterior distributions, since, with finite data, there will remain uncertainty about whether some covariates should or should not be included in the regression. Projection predictive variable selection is an approach to variable selection in Bayesian regression models, which removes covariates that do not considerably contribute to the explanatory power of the full model [Goutis and Robert, 1998; Dupuis and Robert, 2003; Peltola *et al.*, 2014; Piironen and Vehtari, 2017]. It works by projecting the information in the model encompassing all variables to a model that uses only a subset of the variables and has been empirically shown to have competitive performance compared to other variable selection approaches [Piironen and Vehtari, 2017].

For a prediction model M, defined on the full set of covariates, let $p(y \mid x, \theta, M)$ be the observation model of the target variable y given covariates x and model parameters θ and $p(\theta \mid D, M)$ the posterior distribution of the parameters given a training dataset D. Given the parameter θ of the full model M, the projection predictive variable selection approach fits a model M_s with a subset s of the covariates, denoted by x_s with $x = (x_s, x_{\setminus s})$, by minimizing the Kullback–Leibler divergence to the full model M

$$\hat{\theta}_s = \arg\min_{\theta_s} \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}[p(y \mid x_i, \theta, M) \| p(y \mid x_{i,s}, \theta_s, M_s)],$$

where i runs over the N training samples in the dataset D. In practice, the optimization is solved L times for L samples $\theta^{(l)}, l = 1, \ldots, L$, from the posterior distribution $p(\theta \mid D, M)$ to get a projected posterior distribution for the model M_s. The total information loss δ of using the subset x_s instead of full set x is then approximated as the average of the above loss over the samples from the posterior distribution:

$$\delta[M \| M_s] \approx \frac{1}{LN} \sum_{l=1}^{L} \sum_{i=1}^{N} \mathrm{KL}[p(y \mid x_i, \theta^{(l)}, M) \| p(y \mid x_{i,s}, \hat{\theta}_s^{(l)}, M_s)].$$

For generalized linear models, the optimization problems are related to the generalized linear model estimation equations [Goutis and Robert, 1998]. The projection method is, however, not limited to linear regression. For example, the approach has also been used for variable selection in Gaussian processes [Piironen and Vehtari, 2016]. A similar KL divergence minimization approach, with a further regularization penalty, is used in [Tran *et al.*, 2012] to define a predictive lasso method.

For variable selection, the projections are supplanted with a search process for finding a good subset of the covariates. This needs to weigh the benefits and costs of keeping a number of covariates. Dupuis and Robert [2003] introduced the relative explanatory power

$$1 - \frac{\delta[M \| M_s]}{\delta[M \| M_0]}, \tag{1}$$

where M_0 is a null model (e.g., the model without any covariates), to quantify the quality of the subset models and to help in determining the best subset model by choosing the smallest model that retains enough of the explanatory power of the full model. An alternative approach was suggested in [Peltola *et al.*, 2014], where cross-validation was combined with the projection predictive approach to estimate out-of-sample prediction performances at each model size along a forward selection path.

3 KL-LIME for explaining predictions of Bayesian models

We combine ideas from LIME and projection predictive variable selection to define the KL-LIME method for Kullback–Leibler divergence based local interpretable model-agnostic explanations of Bayesian predictive models (although the method can also be applied on non-Bayesian probabilistic models). Let $p(y \mid x, \theta, M)$ be the observation model of the predictive model M given input x and $p(\theta \mid D, M)$ the posterior distribution of its parameters θ given a dataset D. Similar to LIME, we define the explanation as an interpretable model $p(y \mid x', \phi, G)$ from (now a probabilistic) model family G with parameters ϕ, and possibly operating on a simplified representation x' of the original input x. Similar to projection predictive variable selection, the parameters of the explanation model are found by minimizing its Kullback–Leibler

divergence from the predictive model M:

$$\hat{\phi}^{(l)} = \arg\min_{\phi} \int \pi_x(z)\, \mathrm{KL}[p(y \mid z, \theta^{(l)}, M) \| p(y \mid z', \phi, G)] dz + \Omega(\phi)$$
$$\approx \arg\min_{\phi} \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}[p(y \mid z_i, \theta^{(l)}, M) \| p(y \mid z'_i, \phi, G)] + \Omega(\phi),$$

for $l = 1, \ldots, L$ posterior samples from $p(\theta \mid D, M)$ and where $\pi_x(z)$ is a probability distribution (we assume there is a mapping between z to z') defining the local input data space neighborhood around x, the data point which prediction is to be explained. Ω penalizes the complexity of the explanation model. In practice, the expectation over locality π_x is computed via a Monte Carlo approximation by sampling N points from $\pi_x(z)$.

To measure the fidelity of the explanation, we can compute the relative explanatory power between M and G, following Equation 1, with an appropriate definition of the null model. Alternatively, one could also compute more direct measures of the performance, such as mean squared error or classification accuracy by sampling a test set from the locality distribution $\pi_x(z)$.

In the demonstration of the approach below, we use linear models for the explanation model family and L1 regularization for the complexity penalty Ω. The KL minimization can then be solved with lasso regression (or generalized linear model variants of lasso regression). For non-Bayesian probabilistic models, one would not have posterior samples, but only a point estimate of the parameters θ which can be projected to the explanation parameters ϕ.

4 Demonstration

We demonstrate the proposed method in explaining deep convolutional neural network predictions in the MNIST dataset of images of digits, with the task of classifying between 3s and 8s. The neural network has two convolutional layers and two fully connected layers and uses ReLU activation functions[1]. Bayesian inference is approximated with the Bernoulli dropout method [Gal and Ghahramani, 2016b; Gal and Ghahramani, 2016a], with a dropout probability of 0.2 and 100 Monte Carlo samples at test time, which provides a rough approximation of the model uncertainty for prediction. The locality distribution around image x is defined by randomly zeroing pixels (i.e., setting them to the background value of white color) by first sampling the zeroing probability from a beta distribution and then sampling a binary mask with this probability in independent Bernoulli distributions. The simplified representation has 1 for pixels that are not at the background value and 0 otherwise. 1,000 samples are drawn from the locality distribution and used as data for fitting the explanation models with KL-LIME. The explanation model is a linear logistic regression model with L1 penalty.

The top row of Figure 1 shows an example of explaining an image of 8 that is misclassified by the classifier. The relative explanatory power curve can be used to determine the trade-off between explanation fidelity and complexity. In this case, the curve plateaus around 0.85 explanatory power, showing that it's not possible to attain perfect fidelity with the chosen explanation model. The mean explanation shows the posterior mean of the projected parameters of the explanation model. The pixels in the left hand side of the upper loop of the 8 steer the classification towards an 8 as expected for classifying between 8s and 3s. However, the explanation implies that the model has considered the right hand side parts of the image as pointing the classification to a 3. The variance of the explanation is largest in the left hand side of the upper loop of the digit.

The second row of Figure 1 shows mean explanations at different trade-offs between the fidelity and complexity. The lowest level of explanatory power does not capture the classification particularly well. In the case of images as here, even the most complex explanations (by the measure of how many pixels are active in the linear explanation model) are often readily interpretable, since humans are good at image perception. In many other cases, such as textual data or quantitative covariates, say, in personalized medicine, a proper trade-off between complexity and fidelity would be more important. Finally, the two bottom rows of Figure 1 show individual posterior samples of the explanation model. This allows getting a more complete picture of the model uncertainty reflected in the explanations.

Figure 2 shows another example of an explanation: in this case of an image of a 3 that is misclassified as an 8. The explanation implies that this happens because of the elongated lower left curve in the digit.

5 Conclusion

We presented a method, KL-LIME, to construct local interpretable explanations of Bayesian predictive models by projecting the information in the predictive distribution of the model to a simpler, interpretable probabilistic explanation model. The approach is based on combining ideas from the recent Local Interpretable Model-agnostic Explanations (LIME) method [Ribeiro *et al.*, 2016] and Bayesian projection predictive variable selection. This allows accounting for model uncertainty also in the explanations.

The approach gives a principled way of extending LIME to different types of predictions and explanations as long as we can compute the Kullback–Leibler divergence between the predictive distributions for the original model and the explanation model and minimize it to fit the explanation model. In particular, within this constraint, both the original task and the explanation model can be arbitrarily changed without losing the information theoretical interpretation of the projection for finding the explanation model. The value of the KL divergence can be used as a measure of the fidelity of the explanation.

[1] A slightly customized version of the PyTorch MNIST example is used, `https://github.com/pytorch/examples/tree/master/mnist/`, accessed on May 17th, 2018. This achieves about 99.2% accuracy on the test data.

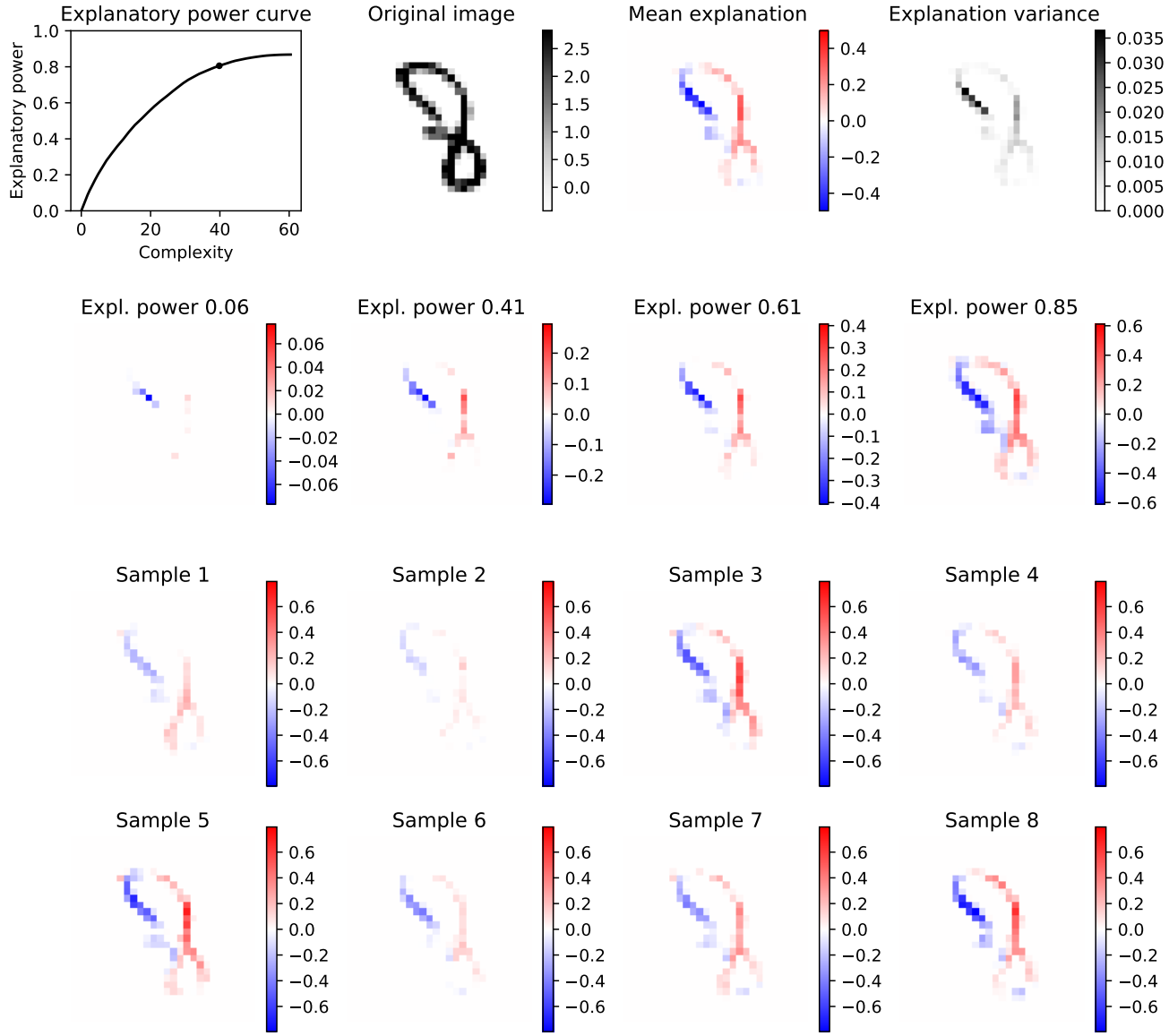

Figure 1: Example explanation of an image of an 8 that was misclassified as 3. Top row: Explanatory power curve shows the relative explanatory power as a function of the explanation complexity (here, the mean number of active pixels in the explanation). Dot shows the complexity chosen for the shown explanation (expl. power of 0.8). Mean explanation gives the posterior mean of the projected explanations and variance gives the uncertainty in the explanation. Second row: Explanations with different complexities and relative explanatory powers. Bottom two rows: Individual posterior samples of explanations.

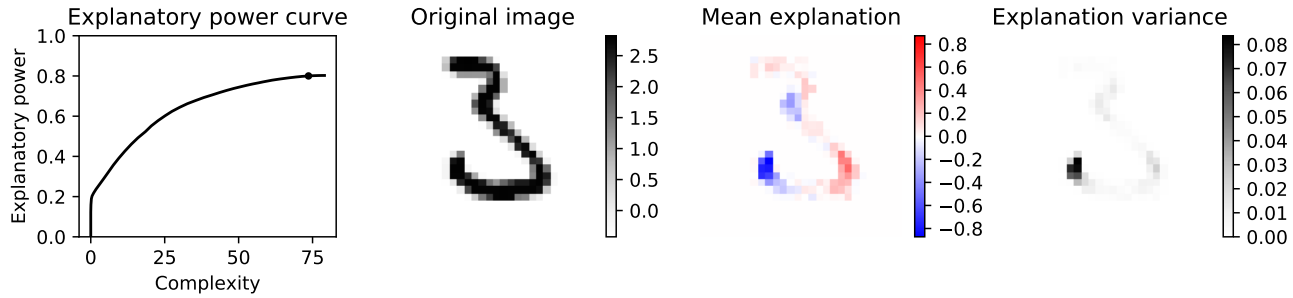

Figure 2: Example explanation of an image of a 3 that was misclassified as 8. See Figure 1 caption for the description of the panels.

References

[Baehrens *et al.*, 2010] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

[Dupuis and Robert, 2003] Jérome A Dupuis and Christian P Robert. Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111(1-2):77–94, 2003.

[Gal and Ghahramani, 2016a] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *4th International Conference on Learning Representations (ICLR) workshop track*, 2016.

[Gal and Ghahramani, 2016b] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059, 2016.

[Goutis and Robert, 1998] Constantinos Goutis and Christian P Robert. Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika*, 85(1):29–37, 1998.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777, 2017.

[Peltola *et al.*, 2014] Tomi Peltola, Aki S Havulinna, Veikko Salomaa, and Aki Vehtari. Hierarchical Bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop*, volume 1218, pages 79–88. CEUR Workshop Proceedings, 2014.

[Piironen and Vehtari, 2016] Juho Piironen and Aki Vehtari. Projection predictive model selection for Gaussian processes. In *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[Piironen and Vehtari, 2017] Juho Piironen and Aki Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[Tran *et al.*, 2012] Minh-Ngoc Tran, David J Nott, and Chenlei Leng. The predictive lasso. *Statistics and computing*, 22(5):1069–1084, 2012.