JMLR: Workshop and Conference Proceedings 1:1–7, 2016 ICML 2016 AutoML Workshop

Bayesian optimization for automated model selection *

Gustavo Malkomes† LUIZGUSTAVO@WUSTL.EDU
Chip Schaff† CBSCHAFF@WUSTL.EDU
Roman Garnett GARNETT@WUSTL.EDU
Department of Computer Science and Engineering
Washington University in St. Louis, St. Louis, MO 63130

Abstract

Despite the success of kernel-based nonparametric methods, kernel selection still requires considerable expertise, and is often described as a "black art." We present a sophisticated method for automatically searching for an appropriate kernel from an infinite space of potential choices. Previous efforts in this direction have focused on traversing a kernel grammar, only examining the data via computation of marginal likelihood. Our proposed search method is based on Bayesian optimization in model space, where we reason about model evidence as a function to be maximized. We explicitly reason about the data distribution and how it induces similarity between potential model choices in terms of the explanations they can offer for observed data. In this light, we construct a novel kernel between models to explain a given dataset. Our method is capable of finding a model that explains a given dataset well without any human assistance, often with fewer computations of model evidence than previous approaches, a claim we demonstrate empirically.

1. Introduction

Over the past decades, enormous human effort has been devoted to machine learning; preprocessing data, model selection, and hyperparameter optimization are some examples of critical and often expert-dependent tasks. The complexity of these tasks has in some cases relegated them to the realm of "black art." In kernel methods in particular, the selection of an appropriate kernel to explain a given dataset is critical to success in terms of the fidelity of predictions, but the vast space of potential kernels renders the problem nontrivial.

Recent work has begun to tackle the kernel-selection problem in a systematic way. Duvenaud et al. (2013) described generative grammars for enumerating a countably infinite space of arbitrarily complex kernels via exploiting the closure of kernels under additive and multiplicative composition. We adopt this kernel grammar in this work as well. Given a dataset, those authors proposed searching this infinite space of models using a greedy search mechanism. Beginning at the root of the grammar, their approach traverse the tree greedily attempting to maximize the (approximate) evidence for the data given by a GP model incorporating the kernel.

In this work, we develop a more sophisticated mechanism for searching through this space. The greedy search described above only considers a given dataset by querying a model's evidence. Our search performs a *metalearning* procedure, which, conditional on a dataset,

† An extended version of this paper will appear at Neural Information Processing Systems (NIPS), 2016.
† These authors contributed equally to this work.

establishes similarities among the models in terms of the space of explanations they can offer for the data. With this viewpoint, we construct a novel kernel between models (a "kernel kernel"). We then approach the model-search problem via Bayesian optimization, treating the model evidence as an expensive black-box function to be optimized as a function of the kernel. The dependence of our kernel between models on the distribution of the data is critical; depending on a given dataset, the kernels generated by a compositional grammar could be especially rich or deceptively so.

We propose an automatic framework for exploring a set of potential models, seeking the model that best explains a given dataset. Although we focus on Gaussian process models defined by a grammar, our method could be easily extended to any probabilistic model with a parametric or structured model space. Our search appears to perform competitively with other baselines across a variety of datasets, including the greedy method from Duvenaud et al. (2013), especially in terms of the number of models for which we must compute the (expensive) evidence, which typically scales cubically for kernel methods.

2. Bayesian optimization for model search

Consider a supervised learning problem defined on an input space $\mathcal{X}$ and output space $\mathcal{Y}$ accompanied by a set of training observations $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ represents the design matrix $\mathbf{x}_i \in \mathcal{X}$, and $y_i \in \mathcal{Y}$ is the respective value or label to be predicted. Ultimately, we want to use $\mathcal{D}$ to predict the value y_* associated with an unseen point $\mathbf{x}_*$. Given a probabilistic model $\mathcal{M}$, we may accomplish this via formation of the predictive distribution.

Suppose, however, that we are given a collection of probabilistic models $\mathbb{M}$ that could have plausibly generated the data. Ideally, finding the source of $\mathcal{D}$ would let us solve our prediction task with the highest fidelity. Let $\mathcal{M} \in \mathbb{M}$ be a probabilistic model, and let $\Theta_{\mathcal{M}}$ be the corresponding parameter space. These models are typically parametric families of distributions, each of which encodes a *structural* assumption about the data, for example, that the data can be described by a linear, quadratic, or periodic trend. Further, the member distributions ($\mathcal{M}_\theta \in \mathcal{M}$, $\theta \in \Theta_{\mathcal{M}}$) of $\mathcal{M}$ differ from each other by a particular value of some properties—represented by the *hyperprameters* θ—related to the data such as amplitude, characteristic length scales, etc.

We wish to select one model from this collection of models $\mathbb{M}$ to explain $\mathcal{D}$. From a Bayesian perspective, the principled approach for solving this problem is *Bayesian model selection*.[1] The critical value is *model evidence*, the probability of generating the observed data given a model $\mathcal{M}$:

$$p(\mathbf{y} \mid \mathbf{X}, \mathcal{M}) = \int_{\Theta_{\mathcal{M}}} p(\mathbf{y} \mid \mathbf{X}, \theta, \mathcal{M})\, p(\theta \mid \mathcal{M})\, \mathrm{d}\theta. \tag{1}$$

The evidence integrates over θ to account for all possible explanations of the data offered by the model, under a prior $p(\theta \mid \mathcal{M})$ associated with that model.

Our goal is to automatically explore a space of models $\mathbb{M}$ to select a model[2] $\mathcal{M}^* \in \mathbb{M}$ that explains a given dataset $\mathcal{D}$ as well as possible, according to the model evidence. The essence of our method, which we call *Bayesian optimization for model search* (BOMS), is

1. "Model selection" is unfortunately sometimes also used in GP literature for the process of hyperparameter learning (selecting some $\mathcal{M}_\theta \in \mathcal{M}$), rather than selecting a model class $\mathcal{M}$, the focus of our work.
2. We could also select a *set* of models but, for simplicity, we assume that there is one model that best explains that data with overwhelming probability, which would imply that there is not benefit in considering more than one model, e.g., via Bayesian model averaging.

viewing the evidence as a function $g\colon \mathbb{M} \to \mathbb{R}$ to be optimized. We note two important aspects of g. First, for large datasets and/or complex models, g is an expensive function, for example growing cubically with $|\mathcal{D}|$ for GP models. Further, gradient information about g is impossible to compute due to the discrete nature of $\mathbb{M}$. We can, however, query a model's evidence as a black-box function. For these reasons, we propose to optimize evidence over $\mathbb{M}$ using *Bayesian optimization,* a technique well-suited for optimizing expensive, gradient-free, black-box objectives (Brochu et al., 2010). In this framework, we seek an optimal model $\mathcal{M}^*$ that maximizes the (log) model evidence: $g(\mathcal{M}; \mathcal{D}) = \log p(\mathbf{y} \mid \mathbf{X}, \mathcal{M})$.

We begin by placing a Gaussian process (GP) prior on g, $p(g) = \mathcal{GP}(g; \mu_g, K_g)$, where $\mu_g\colon \mathbb{M} \to \mathbb{R}$ is a mean function and $K_g\colon \mathbb{M}^2 \to \mathbb{R}$ is a covariance function appropriately defined over the model space $\mathbb{M}$. This is a nontrivial task due to the discrete and potentially complex nature of $\mathbb{M}$. We will suggest useful choices for μ_g and K_g when $\mathbb{M}$ is a space of Gaussian process models below. Now, given observations of the evidence of a selected set of models, $\mathcal{D}_g = \{(\mathcal{M}_i, g(\mathcal{M}_i; \mathcal{D}))\}$, we may compute the posterior distribution on g conditioned on $\mathcal{D}_g$, which will be an updated Gaussian process (Rasmussen and Williams, 2006). Bayesian optimization uses this probabilistic belief about g to induce an inexpensive acquisition function to select which model we should select to evaluate next. Here we use the classical *expected improvement* (EI) (Jones et al., 1998) acquisition function, or a slight variation described in (Snoek et al., 2012), because it naturally considers the trade off between exploration and exploitation. The exact choice of acquisition function, however, is not critical to our proposal. In each round of our model search, we will evaluate the acquisition function in the optimal model evidence for a number of candidate models $\mathcal{C}(\mathcal{D}_g) = \{\mathcal{M}_i\}$, and compute the evidence of the candidate where this is maximized:

$$\mathcal{M}' = \arg\max_{\mathcal{M} \in \mathcal{C}} \alpha_{\text{EI}}(\mathcal{M}; \mathcal{D}_g).$$

We then incorporate the chosen model $\mathcal{M}'$ and the observed model evidence $g(\mathcal{M}'; \mathcal{D})$ into our model evidence training set $\mathcal{D}_g$, update the posterior on g, select a new set of candidates, and continue. We repeat this iterative procedure until a budget is expended, typically measured in terms of the number of models considered.

The acquisition function allows us to quickly determine which models are more promising than others, given the evidence we have observed so far. Since $\mathbb{M}$ is an infinite set of models, we cannot consider every model in every round. Instead, we define a heuristic to evaluate the acquisition function at a smaller set of *active candidate models* $\mathcal{C}$. We construct and maintain this set by balancing exploration (diversity) against exploitation (models likely to have higher evidence). We begin each round with a set of already chosen candidates $\mathcal{C}$. Then, to encourage exploitation, we add to $\mathcal{C}$ all "neighbors" of the best model seen thus far. To encourage exploration, we perform random walks to create diverse models, which we also add to $\mathcal{C}$. To constrain the number of candidates, we discard the models with the lowest EI values at the end of each round, keeping $|\mathcal{C}|$ no larger than 600.

3. Bayesian optimization for Gaussian process kernel search

We introduced above a general framework for searching over a space of probabilistic models $\mathbb{M}$ to explain a dataset $\mathcal{D}$ without making further assumptions about the nature of the models. In the following, we will provide specific suggestions in the case that all members of $\mathbb{M}$ are Gaussian process priors on a latent function.

We assume that our observations $\mathbf{y}$ were generated according to an unknown function $f\colon \mathcal{X} \to \mathbb{R}$ via a fixed probabilistic observation mechanism $p(\mathbf{y} \mid \mathbf{f})$, where $f_i = f(\mathbf{x}_i)$. In our experiments here, we will consider regression with additive Gaussian observation noise, but this is not integral to our approach. We further assume a GP prior distribution on f, $p(f) = \mathcal{GP}(f; \mu_f, K_f)$, where $\mu_f\colon \mathcal{X} \to \mathbb{R}$ is a mean function and $K_f\colon \mathcal{X}^2 \to \mathbb{R}$ is a positive-definite covariance function or kernel. For simplicity, we will assume that the prior on f is centered, $\mu_f(x) = 0$, which lets us fully define the prior on f by the kernel function K_f. The kernel function is parameterized by hyperparameters that we concatenate into a vector θ. In this restricted context, a model $\mathcal{M}$ is completely determined by the choice of kernel function and an associated hyperparameter prior $p(\theta \mid \mathcal{M})$. We use the kernel grammar proposed by Duvenaud et al. (2013) for constructing an infinite space of potential kernels to model the latent function f, and thus an infinite family of models $\mathbb{M}$.

Creating a "kernel kernel". The evidence function g is the objective function we are trying to optimize via Bayesian optimization. Our prior belief about g is given by a GP prior $p(g) = \mathcal{GP}(g; \mu_g, K_g)$, which is fully specified by the mean μ_g and covariance functions K_g. We define the former as a simple constant mean function $\mu_g(\mathcal{M}) = \theta_\mu$, where θ_μ is a hyperparameter to be learned through a regular GP training procedure given a set of observations. The latter we will construct as follows.

The basic idea in our construction is that is that we will consider the distribution of the observation locations in our dataset $\mathcal{D}$, $\mathbf{X}$ (the design matrix of the underlying problem). We note that selecting a model class $\mathcal{M}$ induces a prior distribution over the latent function values at $\mathbf{X}$, $p(\mathbf{f} \mid \mathbf{X}, \mathcal{M})$. This prior distribution is an infinite mixture of multivariate Gaussian prior distributions, each conditioned on specific hyperparameters θ. We consider these prior distributions as different explanations of the latent function f, restricted to the observed locations, offered by the model $\mathcal{M}$. We will compare two models in $\mathbb{M}$ according to how different the explanations they offer for $\mathbf{f}$ are, *a priori*.

The *Hellinger distance* is a probability metric that we adopt as a basic measure of similarity between two distributions. Although this quantity is defined between arbitrary probability distributions (and thus could be used with non-GP model spaces), we focus on the multivariate normal case. Suppose that $\mathcal{M}, \mathcal{M}' \in \mathbb{M}$ are two models that we wish to compare, in the context of explaining a fixed dataset $\mathcal{D}$. For now, suppose that we have conditioned each of these models on arbitrary hyperparameters (that is, we select a particular prior for f from each of these two families), giving $\mathcal{M}_\theta$ and $\mathcal{M}'_{\theta'}$, with $\theta \in \Theta_{\mathcal{M}}$ and $\theta' \in \Theta_{\mathcal{M}'}$. Now, we define the two distributions

$$P = p(\mathbf{f} \mid \mathbf{X}, \mathcal{M}, \theta) = \mathcal{N}(\mathbf{f}; \mu_P, \Sigma_P) \qquad Q = p(\mathbf{f} \mid \mathbf{X}, \mathcal{M}', \theta') = \mathcal{N}(\mathbf{f}; \mu_Q, \Sigma_Q).$$

The squared *Hellinger distance* between P and Q is

$$d^2_{\text{H}}(P, Q) = 1 - \frac{|\Sigma_P|^{1/4}|\Sigma_Q|^{1/4}}{\left|\frac{\Sigma_P + \Sigma_Q}{2}\right|^{1/2}} \exp\left\{-\frac{1}{8}(\mu_P - \mu_Q)^\top \left(\frac{\Sigma_P + \Sigma_Q}{2}\right)^{-1} (\mu_P - \mu_Q)\right\}. \tag{2}$$

The Hellinger distance will be small when P and Q are highly overlapping, and thus $\mathcal{M}_\theta$ and $\mathcal{M}'_{\theta'}$ provide similar explanations *for this dataset*. The distance will be larger, conversely, when $\mathcal{M}_\theta$ and $\mathcal{M}'_{\theta'}$ provide divergent explanations. Critically, we note that this distance depends on the dataset under consideration in addition to the GP priors.

Observe that the distance above is not sufficient to compare the similarity of two models $\mathcal{M}$, $\mathcal{M}'$ due to the fixing of hyperparameters above. To properly account for the different

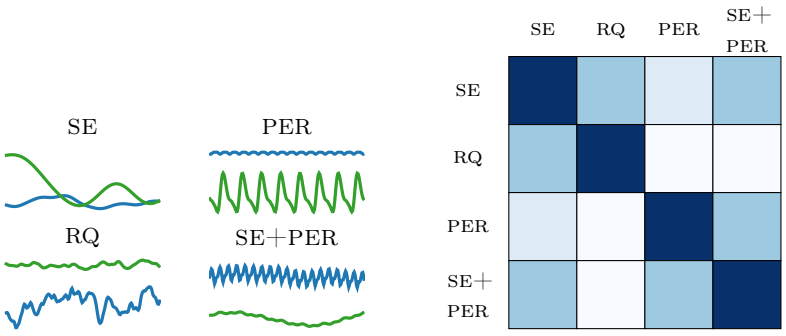

Figure 1: A demonstration of our model kernel K_g (4) based on expected Hellinger distance of induced latent priors. Left: four simple model classes on a $1d$ domain, showing samples from the prior $p(f \mid \mathcal{M}) \propto p(f \mid \theta, \mathcal{M})\, p(\theta \mid \mathcal{M})$. Right: our Hellinger squared exponential covariance evaluated for the grid domains on the left. Increasing intensity indicates stronger covariance. The sets $\{\text{SE}, \text{RQ}\}$ and $\{\text{SE}, \text{PER}, \text{SE+PER}\}$ show strong mutual correlation.

hyperparameters of different models, and the priors associated with them, we define the *expected squared Hellinger distance* of two models $\mathcal{M}, \mathcal{M}' \in \mathbb{M}$ as

$$\bar{d}^2_{\text{H}}(\mathcal{M}, \mathcal{M}'; \mathbf{X}) = \mathbb{E}\big[d^2_{\text{H}}(\mathcal{M}_\theta, \mathcal{M}'_{\theta'})\big] = \iint d^2_{\text{H}}(\mathcal{M}_\theta, \mathcal{M}'_{\theta'}; \mathbf{X})\, p(\theta \mid \mathcal{M})\, p(\theta' \mid \mathcal{M}')\, \mathrm{d}\theta\, \mathrm{d}\theta', \quad (3)$$

where the distance is understood to be evaluated between the priors provided on $\mathbf{f}$ induced at $\mathbf{X}$. Finally, we construct the *Hellinger squared exponential* covariance between models as

$$K_g(\mathcal{M}, \mathcal{M}'; \theta_g, \mathbf{X}) = \sigma^2 \exp\left(-\frac{1}{2}\frac{\bar{d}^2_{\text{H}}(\mathcal{M}, \mathcal{M}'; \mathbf{X})}{\ell^2}\right), \quad (4)$$

where $\theta_g = (\sigma, \ell)$ specifies output and length scale hyperparameters in this kernel/evidence space. This covariance is illustrated in Figure 1.

We make two notes before continuing. The first observation is that computing (2) scales cubically with $|\mathbf{X}|$, so it might appear that we might as well compute the evidence instead. This is misleading for two reasons. First, the (approximate) computation of a given model's evidence via either a Laplace approximation or the BIC requires optimizing its hyperparameters. Especially for complex models this can require hundreds-to-thousands of computations that each require cubic time. Further, as a result of our investigations, we have concluded that in practice we may approximate (2) and (3) by considering only a small *subset* of the observation locations $\mathbf{X}$ and that this usually sufficient to capture the similarity between models in terms of explaining a given dataset. In our experiments, we choose 20 points uniformly at random from those available in each dataset, fixed once for the entire procedure and for all kernels under consideration in the search. We then used these points to compute distances (2–4), significantly reducing the overall time to compute K_g.

Second, we note that the expectation in (3) is intractable. We approximate the expectation via quasi-Monte Carlo, using a low-discrepancy sequence (a Sobol sequence) of the appropriate dimension, and inverse transform sampling, to give consistent, representative samples from the hyperparameter space of each model. We used 100 (θ, θ') samples with good results.

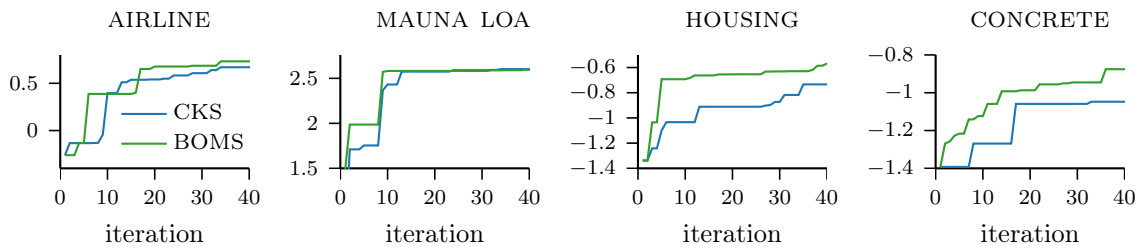

Figure 2: A plot of the best model evidence found (normalized by $|\mathcal{D}|$) as a function of the number of models evaluated, $g(\mathcal{M}^*; \mathcal{D})$, for six of the datasets considered (identical vertical axis labels omitted for greater horizontal resolution).

4. Experiments

Here, we evaluate our method's ability to quickly find a suitable model to explain a given dataset. The datasets we consider are publicly available[3] and were used in previous related work (Duvenaud et al., 2013; Bach, 2008). AIRLINE, MAUNA LOA are $1d$ time series, and CONCRETE and HOUSING have, respectively, 8 and 13 dimensions. To facilitate comparison of evidence across datasets, we report log evidence divided by dataset size, redefining $g(\mathcal{M}; \mathcal{D}) = \log(p(\mathbf{y} \mid \mathbf{X}, \mathcal{M}))/|\mathcal{D}|$. Our setup for the base kernels is the same as in (Duvenaud et al., 2013). We compare our approach with the greedy *compositional kernel search* (CKS) of Duvenaud et al. (2013). Both algorithms used the same kernel grammar, hyperparameter priors, and evidence approximation (Laplace approximation). We used L-BFGS to optimize model hyperparameters, using multiple restarts to avoid bad local maxima; each restart begins from a sample from $p(\theta \mid \mathcal{M})$. For BOMS, we always began our search evaluating SE first. We also initialized the set of active models with all models that are at most two edges distant from the base kernels. To avoid unnecessary re-training over g, we optimized the hyperparameters of μ_g and K_g every 10 iterations. This also allows us to perform rank-one updates for fast inference during the intervening iterations.

Results are depicted in Figure 2 for a budget of 40 evaluations of the model evidence. In three of the four datasets we substantially outperform CKS. Note the vertical axis is in the log domain. The overhead for computing the kernel K_g and performing the inference about g was approximately 10% of the total running time. On MAUNA LOA our method is competitive since we find a model with similar quality, but earlier.

5. Conclusion

We introduced a novel automated search for an appropriate kernel to explain a given dataset. Our mechanism explores a space of infinite candidate kernels and quickly and effectively selects a promising model. Focusing on the case where the models represent structural assumptions in GPs, we introduced a novel "kernel kernel" to capture the similarity in prior explanations that two models ascribe to a given dataset. We have empirically demonstrated that our choice of modeling the evidence (or marginal likelihood) with a GP in model space is capable of predicting the evidence value of unseen models with enough fidelity to effectively explore model space via Bayesian optimization.

3. https://archive.ics.uci.edu/ml/datasets.html

References

Francis R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Conference on Neural Information Processing Systems (NIPS)*, 2008.

Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In *International Conference on Machine Learning (ICML)*, 2013.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Conference on Neural Information Processing Systems*, 2012.