Kernel Structure Discovery in Gaussian Process Models

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

Abstract

Gaussian process (GP) models are used widely and successfully. However, their effictiveness depends critically on choosing an appropriate family of kernels. This aspect of GP modeling has been sorely underdeveloped. In this paper, we introduce a procedure for automatically and efficiently searching through a large space of GP models.

1 INTRODUCTION

Similar searches over large model classes have been succesfully used in machine vision [cite Cox + Pinto]. In general, learning the model class from data seems superior proposing the model beforehand. In high dimensional problems, it is also hard for a practitioner to propose an appropriate model even after examining a dataset closely. Choosing a kernel family is also a stumbling block for non-experts who wish to use Gaussian Process models.

2 GP STRUCTURE

[Standard GP intro Rasmussen & Williams (2006)]

3 MODEL-BASED SEARCH OVER MODELS

Bayesian optimization for hyper-parameter search: Snoek et al. (2012)

3.1 Bayesian Optimization

3.2 A Kernel between kernels

Hyperkernels Ong et al. (2002)

Preliminary work. Under review by AISTATS 2013. Do not distribute.

4 RELATED WORK

Compositional Model search for unsupervsied learning: Grosse et al. (2012)

Hyperkernels Ong et al. (2002)

4.1 ANOVA Kernels

Additive Gaussian Processes Duvenaud et al. (2011)

Support vector regression with ANOVA decomposition kernels Stitson et al. (1999)

4.1.1 Smoothing spline ANOVA models

A closely related procedure from the statistics literature is smoothing-splines ANOVA (SS-ANOVA)Wahba (1990); Gu (2002). An SS-ANOVA model is estimated as a weighted sum of splines along each dimension, plus a sum of splines over all pairs of dimensions, all triplets, etc, with each individual interaction term having a separate weighting parameter. Because the number of terms to consider grows exponentially in the order, in practice, only terms of first and second order are usually considered. Learning in SS-ANOVA is usually done via penalized-maximum likelihood with a fixed sparsity hyperparameter.

4.2 Hierarchical Kernel Learning

In "High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning", BachBach (2009) uses a regularized optimization framework to learn a weighted sum over an exponential number of kernels which can be computed in polynomial time. The subsets of kernels considered by this method are restricted to be a *hull* of kernels.[1] Given each dimension's kernel, and a pre-defined weighting over all terms, HKL performs model selection by searching over hulls of interaction terms. In Bach (2009), Bach

[1] In the setting we are considering in this paper, a hull can be defined as a subset of all terms such that if term $\prod_{j \in J} k_j(\mathbf{x}, \mathbf{x}')$ is included in the subset, then so are all terms $\prod_{j \in J/i} k_j(\mathbf{x}, \mathbf{x}')$, for all $i \in J$. For details, see Bach (2009).

also fixes the relative weighting between orders of interaction with a single term α, computing the sum over all orders by:

$$k_a(\mathbf{x}, \mathbf{x}') = v_D^2 \prod_{d=1}^{D} (1 + \alpha k_d(x_d, x'_d)) \qquad (1)$$

which has computational complexity $O(D)$. However, this formulation forces the weight of all nth order terms to be weighted by α^n.

The main difficulty with the approach of Bach (2009) is that hyperparameters are hard to set other than by cross-validation. In contrast, our method optimizes the hyperparameters of each dimension's base kernel, as well as the relative weighting of each order of interaction.

Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using Gaussian process models Plate (1999)

A related functional ANOVA GP modelKaufman & Sain (2010) decomposes the *mean* function into a weighted sum of GPs. However, the effect of a particular degree of interaction cannot be quantified by that approach. Also, computationally, the Gibbs sampling approach used in Kaufman & Sain (2010) is disadvantageous.

4.3 Genetic Searches

Evolving kernel functions for SVMs by genetic programming: Diosan et al. (2007)

A Genetic Programming based kernel construction and optimization method for Relevance Vector Machines: Bing et al. (2010)

4.4 Equation Learning

Equation discovery with ecological applications Dzeroski et al. (1999)

Discovering admissible model equations from observed data based on scale-types and identity constraints Washio et al. (1999)

4.5 Multiple Kernel Learning

Christoudias et al.Christoudias et al. (2009) previously showed how mixtures of kernels can be learnt by gradient descent in the Gaussian process framework. They call this *Bayesian localized multiple kernel learning*.

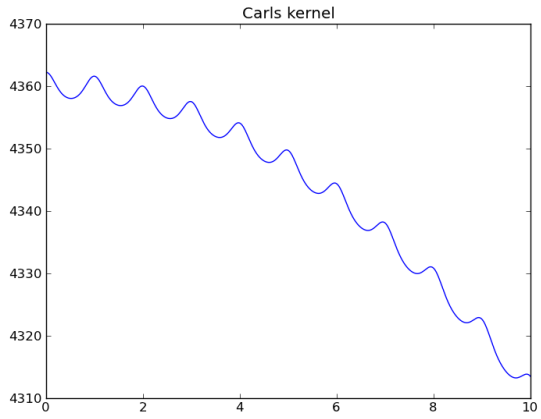

Figure 1: Carl's kernel function on the Mauna dataset.

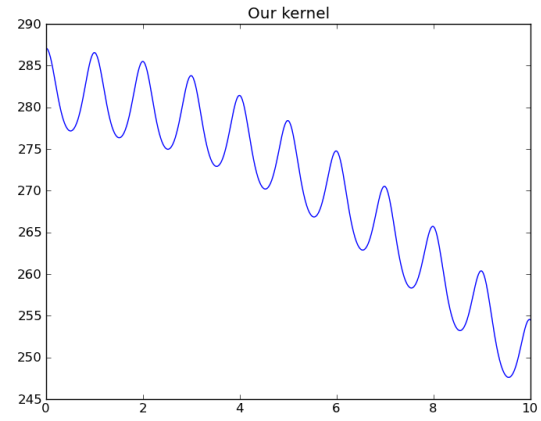

Figure 2: Our best kernel function on the Mauna dataset.

5 EXPERIMENTS

5.1 Real datasets

5.1.1 Maunal Loa Atmoshperic Carbon Dioxide

As an example of a GP modeling problem where choosing an appropriate structure is critical, we revisit a dataset explored in Rasmussen & Williams (2006), pages 120-126, where a kernel was hand-tailored to fit a GP model to the dataset.

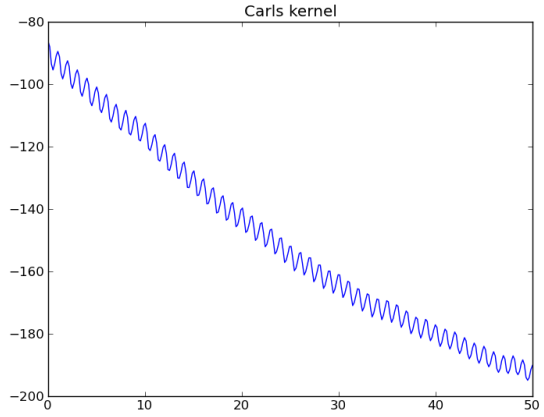

Figure 3: A draw from Carl's kernel.

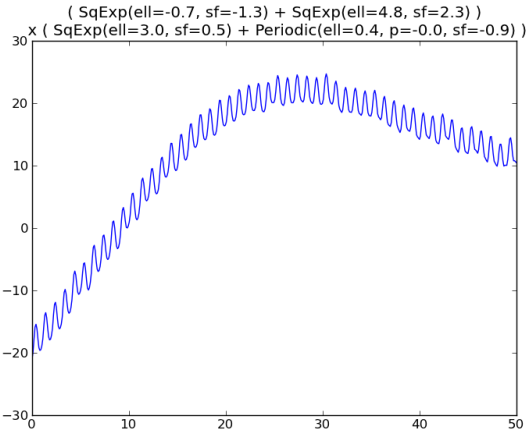

Figure 5: Another draw from our kernel.

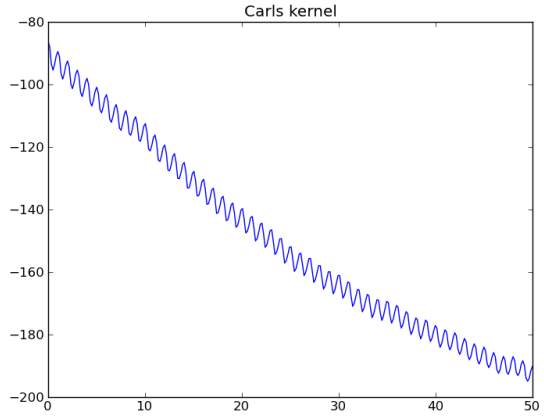

Figure 4: Another draw from Carl's kernel.

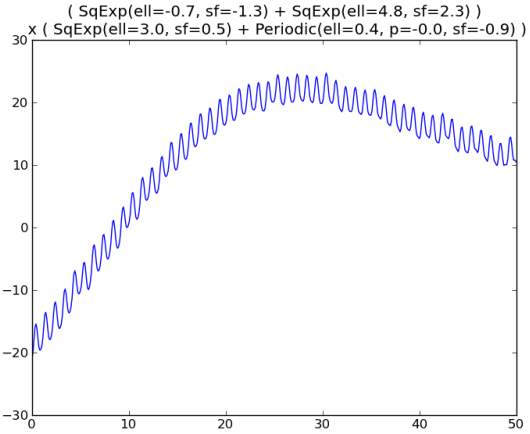

Figure 6: Another draw from our kernel.

5.1.2 Load forecasting

5.2 Synthetic experiments

5.2.1 Bach Synthetic Dataset

In addition to standard UCI repository datasets, we generated a synthetic dataset following the same recipe as Bach (2009): From a covariance matrix drawn from a Wishart distribution with 1024 degrees of freedom, we select 8 variables. We then construct the non-linear function $f(X) = \sum_{i=1}^{4} \sum_{j=1+1}^{4} X_i X_j + \epsilon$, which sums all 2-way products of the first 4 variables, and adds Gaussian noise ϵ. This dataset is one which can be predicted well by a kernel which is a sum of two-way interactions over the first 4 variables, ignoring the extra 4 noisy copies.

This dataset was designed by Bach (2009) to demonstrate the advantages of HKL over GP-ARD.

If the dataset is large enough, HKL can construct a hull around only those subsets of cross terms that are optimal for predicting the output. GP-ARD, in contrast, can only learn to ignore the noisy copy variables, but cannot learn to ignore the higher-term interactions between the predictive variables. However, a GP with an additive kernel can learn both to ignore irrelevant variables, and to ignore certain orders of interaction. In this example, the additive GP is able to recover the relevant structure.

5.3 Methods

5.3.1 Our method

All of the experiments in this paper were performed using the standard GPML toolbox[2]; code to perform all experiments is available at the authors' website.

[2] Available at `http://www.gaussianprocess.org/gpml/code/`

5.3.2 Hierarchical Kernel Learning

HKL[3] was run using the all-subsets kernel, which corresponds to the same set of kernels as considered by the additive GP with a squared-exp base kernel.

6 DISCUSSION

Machine learning can be more data-driven, analogous to the high-thoughput approaches being used in biology.

7 CONCLUSION

Acknowledgements

References

Bach, Francis. High-dimensional non-linear variable selection through hierarchical kernel learning. *CoRR*, abs/0909.0844, 2009.

Bing, W., Wen-qiong, Z., Ling, C., and Jia-hong, L. A gp-based kernel construction and optimization method for rvm. In *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, volume 4, pp. 419–423. IEEE, 2010.

Christoudias, M., Urtasun, R., and Darrell, T. Bayesian localized multiple kernel learning. *Technical report*, 2009.

Diosan, L., Rogozan, A., and Pecuchet, J.P. Evolving kernel functions for svms by genetic programming. In *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*, pp. 19–24. IEEE, 2007.

Duvenaud, David, Nickisch, Hannes, and Rasmussen, Carl Edward. Additive Gaussian processes. In *Advances in Neural Information Processing Systems 25*, pp. 1–8, Granada, Spain, December 2011.

Dzeroski, S., Todorovski, L., Bratko, I., Kompare, B., and Krizman, V. Equation discovery with ecological applications. *Machine learning methods for ecological applications*, pp. 185–207, 1999.

Grosse, R.B., Salakhutdinov, R., and Tenenbaum, J.B. Exploiting compositionality to explore a large space of model structures. In *Uncertainty in Artificial Intelligence*, 2012.

Gu, C. *Smoothing spline ANOVA models*. Springer Verlag, 2002. ISBN 0387953531.

Kaufman, C.G. and Sain, S.R. Bayesian functional anova modeling using Gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–150, 2010.

[3]Code for HKL available at `http://www.di.ens.fr/~fbach/hkl/`

Ong, C.S., Smola, A.J., and Williamson, R.C. Hyperkernels. *Advances in neural information processing systems*, 15:478–485, 2002.

Plate, T.A. Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using Gaussian process models. *Behaviormetrika*, 26:29–50, 1999. ISSN 0385-7417.

Rasmussen, C.E. and Williams, CKI. Gaussian Processes for Machine Learning. *The MIT Press, Cambridge, MA, USA*, 2006.

Snoek, J., Larochelle, H., and Adams, R.P. Practical bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944*, 2012.

Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., and Weston, J. Support vector regression with ANOVA decomposition kernels. *Advances in kernel methods: Support vector learning*, pp. 285–292, 1999.

Wahba, G. *Spline models for observational data*. Society for Industrial Mathematics, 1990. ISBN 0898712440.

Washio, T., Motoda, H., Niwa, Y., et al. Discovering admissible model equations from observed data based on scale-types and identity constraints. In *International Joint Conference On Artifical Intelligence*, volume 16, pp. 772–779, 1999.