Structure Discovery in Nonparametric Classification through Compositional Kernel Search

Abstract

Our work builds on previous work on structure discovery for GP regression by extending the approach to facilitate automatic kernel structure discovery for GP classification.

1. Introduction

.....

2. Expressing structure through kernels

.....

Composing Kernels Plots of summation and multiplication of SE kernels.

3. Searching over structures

The base kernel family consists of one dimensional SE kernels. Our search procedure begins by evaluating base SE kernels applied to all input dimensions. Two search operators are defined over our set of expressions:

- Add SE: We can add an arbitrary base kernel $\mathcal{B}$ to the entire expression to obtain $\mathcal{S} + \mathcal{B}$, where $\mathcal{B}$ is an arbitrary base SE kernel.
- Multiply with SE: Any subexpression $\mathcal{S}$ representing a product of base kernels can be multiplied with another base SE kernel to obtain $\mathcal{S} \times \mathcal{B}$,

This context-free grammar allows us to reach any arithmetic expression where multiplication is fully distributed across addition. The algorithm searches over this space using a greedy search: at each stage, we choose the highest scoring kernel (in terms of BIC) and expand it using all applicable operators.

Scoring kernel families Marginal likelihood balances the fit and complexity of a model (Rasmussen & Ghahramani, 2001). For regression, the marginal likelihood of a GP can be computed analytically. This is not true for classification, as the likelihood function is no longer Gaussian. In our work, we use the Laplace approximation to compute the marginal likelihood. To evaluate a kernel family we must integrate out the hyperparameters. We first optimize to find the maximum-likelihood hyperparameters and then approximate the intractable integral using the Bayesian information criterion (Schwarz, 1978), in order to minimise the effect of the prior on our estimate.

Finding the optimal hyperparameters is not a convex optimization problem, as the space can have many local optima. This issue was especially pronounced in (Duvenaud et al., 2013), as periodic kernels were part of the search grammar. This issue is not as serious with the kernel family restricted to smooth SE functions, but the search procedure still relies on random initializations for the newly introduced hyperparameters in advance of the maximum-likelihood optimisation.

Table 1. Classification Percent Error

Method	breast	pima	liver	heart
Logistic Regression	7.611	24.392	45.060	**16.082**
GP GAM	**5.189**	**22.419**	**29.842**	**16.839**
HKL	**5.377**	24.261	**27.270**	**18.975**
GP Squared-exp	**4.734**	**23.722**	**31.237**	**20.642**
GP Additive	**5.566**	**23.076**	**30.060**	**18.496**
GPSS (AIC)	6.430	22.529	28.924	–
GPSS (BIC)	5.980	23.440	37.010	18.150
GPSS (BIC light)	6.430	22.270	27.5	17.820

4. Related Work

Nonparametric classification .....

Kernel learning .....

Structure discovery .....

*Equal contribution. †U. Cambridge. ‡MIT.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

5. Structure discovery for classification

In this section, we compare the performance of models found in our search with related methods and show that the performance of our structurally simpler models is on par with more complicated models such as additive GPs (Duvenaud et al., 2011) and Hierarchical Kernel Learning.

Heart

Liver

Pima

Breast

6. Validation on synthetic data

We validated our method's ability to recover known structure on a set of synthetic datasets. For several composite kernel expressions, we constructed synthetic data by first sampling 100, 300 and 500 points uniformly at random, then sampling function values at those points from a GP prior. We then added i.i.d. Gaussian noise to the functions, at various signal-to-noise ratios (SNR), as well as different amounts of salt and pepper noise (random outliers in the data set).

Table 2 lists the true kernels we used to generate the data. Subscripts indicate which dimension each kernel was applied to. Subsequent columns show the dimensionality D of the input space, and the kernels chosen by our search for different SNRs and different amounts of added salt and pepper noise. We also show the kernel optimal rates (the accuracy the kernel used to generate the data achieves on the noisy test set) and the function optimal rates (the rate a classifier which knew the *exact* funtion used to generate the data achieves on the noisy test data set).

Table 2. True kernel: $\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$, with log-lengthscales -1, -1, -1. The synthetic data is three dimensional. Kernels chosen by our method on synthetic data generated using known kernel structures. SNR indicates the signal-to-noise ratio, sp_noise the proportion of random outliers inserted.

Data size	SNR	sp_noise	Kernel chosen	Test accuracy	Kernel rate	Bayes optimal rate
100	1	0%	SE_3	72.0%	62.0%	77.8%
300	1	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	73.7%	70.7%	77.8%
500	1	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	74.4%	73.8%	77.8%
100	1	5%	$\mathrm{SE}_1 + \mathrm{SE}_3$	66.0%	68.0%	75.8%
300	1	5%	$\mathrm{SE}_1 + \mathrm{SE}_3$	71.7%	75.0%	75.8%
500	1	5%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	73.4%	73.0%	75.8%
100	1	20%	$\mathrm{SE}_1 + \mathrm{SE}_2$	64.0%	62.0%	69.4%
300	1	20%	$\mathrm{SE}_1 + \mathrm{SE}_2$	61.3%	61.7%	69.4%
500	1	20%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	66.2%	64.4%	69.4%
100	100	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	92.0%	90.0%	96.2%
300	100	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	91.0%	92.0%	96.2%
500	100	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	94.2%	94.4%	96.2%
100	100	5%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	86.0%	83.0%	94.8%
300	100	5%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	90.3%	89.7%	94.8%
500	100	5%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	90.2%	90.0%	94.8%
100	100	20%	SE_1	71.0%	69.0%	85.2%
300	100	20%	$\mathrm{SE}_1 \times \mathrm{SE}_2 + \mathrm{SE}_2$	75.0%	73.3%	85.2%
500	100	20%	$\mathrm{SE}_1 + \mathrm{SE}_2 + \mathrm{SE}_3$	82.0%	80.0%	85.2%

Table 3. True kernel: $\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3 + \mathrm{SE}_4$, with log-lengthscales all equal to 0 (all three terms have equal lengthscales). The synthetic data is three dimensional. Kernels chosen by our method on synthetic data generated using known kernel structures. SNR indicates the signal-to-noise ratio, sp_noise the proportion of random outliers inserted.

Data size	SNR	sp_noise	Kernel chosen	Test accuracy	Kernel rate	Bayes optimal rate
100	1	0%	SE_1	58.0%	74.0%	76.2%
300	1	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3$	70.0%	70.3%	76.2%
500	1	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3$	72.4%	74.4%	76.2%
100	1	5%	$\mathrm{SE}_2 \times \mathrm{SE}_3$	62.0%	66.0%	75.6%
300	1	5%	$\mathrm{SE}_2 \times \mathrm{SE}_3$	71.0%	71.3%	75.6%
500	1	5%	$\mathrm{SE}_1 \times \mathrm{SE}_4 + \mathrm{SE}_2 \times \mathrm{SE}_3$	71.8%	72.8%	75.6%
100	1	20%	SE_2	58.0%	64.0%	72.4%
300	1	20%	$\mathrm{SE}_2 \times \mathrm{SE}_3 \times \mathrm{SE}_4$	67.7%	69.3%	72.4%
500	1	20%	$\mathrm{SE}_2 \times \mathrm{SE}_3 \times \mathrm{SE}_4$	70.8%	69.0%	72.4%
100	100	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3 + \mathrm{SE}_4$	85.0%	84.0%	96.6%
300	100	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3 + \mathrm{SE}_4$	92.7%	93.3%	96.6%
500	100	0%	$\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3 + \mathrm{SE}_4$	94.2%	94.8%	96.6%
100	100	5%	$\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3 + \mathrm{SE}_4$	84.0%	82.0%	95.4%
300	100	5%	$\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3 \times \mathrm{SE}_4$	91.0%	88.3%	95.4%
500	100	5%	$\mathrm{SE}_1 + \mathrm{SE}_2 \times \mathrm{SE}_3 \times \mathrm{SE}_4 + \mathrm{SE}_4$	90.2%	90.6%	95.4%
100	100	20%	SE_1	61.0%	66.0%	88.4%
300	100	20%	$\mathrm{SE}_1 \times \mathrm{SE}_3$	71.3%	78.0%	88.4%
500	100	20%	$\mathrm{SE}_1 \times \mathrm{SE}_2 \times \mathrm{SE}_3$	81.4%	80.8%	88.4%

Table 4. True kernel: $SE_1 + SE_3 \times SE_7 + SE_{10}$, with log-lengthscales all equal to 0 (all three terms have equal lengthscales). The synthetic data is ten dimensional. Kernels chosen by our method on synthetic data generated using known kernel structures. SNR indicates the signal-to-noise ratio, sp_noise the proportion of random outliers inserted.

Data size	SNR	sp_noise	Kernel chosen	Test accuracy	Kernel rate	Bayes optimal rate
100	100	0%	$SE_1 + SE_3 \times SE_7$	96%	96%	98.2%
300	100	0%	$SE_1 + SE_3 \times SE_7 + SE_{10}$	97.33%	97.33%	98.2%
500	100	0%	$SE_1 + SE_3 \times SE_7$	96.6%	98%	98.2%
100	100	5%	$SE_1 + SE_3 \times SE_7 + SE_7$	90%	88%	98.2%
300	100	5%	$SE_3 \times SE_7$	94%	94.33%	98.2%
500	100	5%	$SE_3 \times SE_7$	94%	94.8%	98.2%
100	100	20%	SE_2	57%	68%	98.2%
300	100	20%	$SE_3 \times SE_7$	85%	75.67%	98.2%
500	100	20%	SE_8	45.8%	79.4%	98.2%
100	1	0%	$SE_3 \times SE_7$	75%	72%	79.2%
300	1	0%	$SE_3 \times SE_7$	77.33%	79.33%	79.2%
500	1	0%	$SE_3 \times SE_7$	76.8%	77%	79.2%
100	1	5%	SE_4	47%	66%	79.2%
300	1	5%	$SE_3 \times SE_7$	76%	75.67%	79.2%
500	1	5%	$SE_3 \times SE_7$	75.8%	73.2%	79.2%
100	1	20%	SE_7	63%	67%	79.2%
300	1	20%	SE_7	62.67%	63.67%	79.2%
500	1	20%	SE_7	66%	68%	79.2%

Table 5. True kernel: $SE_1 + SE_3 \times SE_5 \times SE_7 + SE_9$, with log-lengthscales all equal to -1 (all three terms have equal lengthscales). The synthetic data is ten dimensional. Kernels chosen by our method on synthetic data generated using known kernel structures. SNR indicates the signal-to-noise ratio, sp_noise the proportion of random outliers inserted.

Data size	SNR	sp_noise	Kernel chosen	Test accuracy	Kernel rate	Bayes optimal rate
100	100	0%	SE_7	42%	64%	95.8%
300	100	0%	SE_7	48%	74.33%	95.8%
500	100	0%	$SE_3 \times SE_5 \times SE_7$	75.6%	77.2%	95.8%
100	100	5%	SE_1	50%	62%	95.8%
300	100	5%	SE_7	47.67%	67.33%	95.8%
500	100	5%	$SE_3 \times SE_5 \times SE_7$	73.4%	71%	95.8%
100	100	20%	SE_1	52%	61%	95.8%
300	100	20%	$SE_9 \times SE_{10}$	56.33%	58.67%	95.8%
500	100	20%	SE_9	52.8%	57.2%	95.8%
100	1	0%	SE_4	60%	58%	80.4%
300	1	0%	SE_1	55.33%	65.67%	80.4%
500	1	0%	$SE_3 \times SE_5 \times SE_7 + SE_{10}$	64.8%	67.4%	80.4%
100	1	5%	SE_3	60%	59%	80.4%
300	1	5%	SE_3	52.33%	63.67%	80.4%
500	1	5%	SE_3	44.2%	65.6%	80.4%
100	1	20%	SE_4	58%	52%	80.4%
300	1	20%	SE_8	49.33%	53.33%	80.4%
500	1	20%	$SE_5 + SE_8$	46.8%	53%	80.4%

7. Quantitative evaluation

8. Discussion

Acknowledgements

References

Duvenaud, D., Nickisch, H., and Rasmussen, C.E. Additive Gaussian processes. In *Advances in Neural Information Processing Systems*, 2011.

Duvenaud, David, Lloyd, James Robert, Grosse, Roger, Tenenbaum, Joshua B., and Ghahramani, Zoubin. Structure discovery in nonparametric regression through compositional kernel search. pp. 1166–1174, June 2013.

Rasmussen, C.E. and Ghahramani, Z. Occam's razor. In *Advances in Neural Information Processing Systems*, 2001.

Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.