Kernel Structure Discovery for Gaussian Process Classification

Nikola Mrkšić

Trinity College, University of Cambridge

June 9, 2014

The Automated Statistician

Big data is transforming the sciences and commerce. However, there are too few expert statisticians to make use of all the new data available

The Automated Statistician Project

Long term goal: remove the human expert from the process by making statistical inference automatic and abstracting it away from the end-user

Given an arbitrary data set, the Automated Statistician should:

- Apply many different models to identify structure in the data
- Suggest interpretable hypotheses and models to use for the data set
- Produce a natural language report about the data set, explaining the identified patterns to the user and representing them visually

Different types of data: regression, **classification**, etc.

Which models do we use?

Bayesian nonparametric models assume that the models have an infinite number of parameters: these models are as flexible as they need to be to describe the complex relationships encountered in real-world data
Gaussian Processes specify *distributions over functions*

- Infinite-dimensional generalisation of multinomial Normal distribution
- Priors for expressing beliefs about the functions we are modelling

GPs are fully specified by the mean function $\mu(x)$ and and the covariance (kernel) function $K(x, x')$. The *squared exponential* (SE) covariance function specifies a prior over smooth functions:

$$K(\mathbf{x}, \mathbf{x}') = \sigma_0^2 \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')^2}{2\lambda^2} \right)$$

The lengthscale λ and signal variance σ_0 are the kernel hyperparameters.

How do we choose/construct the kernel for the GP?

Given a data set, the Automated Statistician must identify the kernel to use and optimise its hyperparameters for the given data set.

Regression

(Duvenaud et al, 2013) *search over a grammar over kernel structures* defined by algebraic expressions over one-dimensional kernels:

- Base terms are 1D kernels defined on individual input dimensions
- Composite kernels are created by adding and multiplying base kernels
- Marginal likelihood is used to score different kernel structures
- Greedy breadth first search used to guide the search procedure

On time series data, models constructed by the search achieve competitive predictive performance and decompose into interpretable components. (Lloyd et al, 2014) use the results of the structure search to generate 10-15 page reports which describe trends, change-points, periodicity, etc.

Searching over the Kernel Grammar

Summing two kernels equates to an OR operator, while multiplication resembles an AND (and allows the kernels to capture non-local structure)

$$\mathrm{SE}_1 \quad \mathrm{SE}_2 \quad \mathrm{SE}_3 \quad \mathrm{SE}_4 \quad \mathrm{SE}_5 \quad \mathbf{SE_6}$$

$$\mathrm{SE}_6 + \mathrm{SE}_1 \quad \mathrm{SE}_6 + \mathrm{SE}_2 \quad \cdots \quad \mathbf{SE_6 + SE_5} \quad \cdots$$
$$\mathrm{SE}_6 \times \mathrm{SE}_1 \quad \mathrm{SE}_6 \times \mathrm{SE}_2 \quad \cdots \quad \mathrm{SE}_6 \times \mathrm{SE}_6$$

$$\mathrm{SE}_6 + \mathrm{SE}_5 + \mathrm{SE}_1 \quad \cdots \quad \mathrm{SE}_6 + \mathrm{SE}_5 + \mathrm{SE}_6$$
$$\mathrm{SE}_6 \times \mathrm{SE}_1 + \mathrm{SE}_5 \quad \cdots \quad \mathrm{SE}_6 \times \mathrm{SE}_6 + \mathrm{SE}_5$$
$$\mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_1 \quad \cdots \quad \mathbf{SE_6 + SE_5 \times SE_3} \quad \cdots$$

$$\mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_3 + \mathrm{SE}_1 \quad \cdots \quad \mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_3 + \mathrm{SE}_6$$
$$\mathrm{SE}_6 \times \mathrm{SE}_1 + \mathrm{SE}_5 \times \mathrm{SE}_3 \quad \cdots \quad \mathrm{SE}_6 \times \mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_3$$
$$\mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_3 \times \mathrm{SE}_1 \quad \cdots \quad \mathbf{SE_6 + SE_5 \times SE_3 \times SE_4} \quad \cdots$$

$$\mathbf{SE_6 + SE_5 \times SE_3 \times SE_4 + SE_1} \quad \cdots \quad \mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_3 \times \mathrm{SE}_4 + \mathrm{SE}_6$$
$$\mathrm{SE}_6 \times \mathrm{SE}_1 + \mathrm{SE}_5 \times \mathrm{SE}_3 \times \mathrm{SE}_4 \quad \cdots \quad \mathrm{SE}_6 \times \mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_3 \times \mathrm{SE}_4$$
$$\mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_3 \times \mathrm{SE}_4 \times \mathrm{SE}_1 \quad \cdots \quad \mathrm{SE}_6 + \mathrm{SE}_5 \times \mathrm{SE}_3 \times \mathrm{SE}_4 \times \mathrm{SE}_6$$

Extending Kernel Discovery to Classification

The Difficulties with Classification

1. Non-Gaussian likelihood functions (sigmoid, error function)
   Laplace's Method, Expectation Propagation, Variational Bayes
2. Model Selection
   What is the *number of effective hyperparameters*?
3. Interpretability: reduced information content
   Finding and plotting patterns becomes harder than for regression

Model Selection

- The marginal likelihood exhibits an Occam's razor effect for achieving an optimal trade-off between the *quality of fit* and *model complexity*
- Each model's hyperparameters are optimised with respect to the marginal likelihood (conjugate gradient methods, multiple restarts)
- Marginal likelihood values are still conditioned on the final hyperparameters; need to integrate these out:

$$\mathrm{BIC} = -2 \ln L + k \ln n \qquad \mathrm{AIC} = -2 \ln L + 2k$$

where k is the number of *effective hyperparameters* of the model

1. Removing variances lead to a third information criteria: *BIC light*
2. One can go one step further than AIC and just choose the models with the best *cross-validated training accuracy* (no stopping criterion)

Recovering the True Structure for Synthetic Data Sets

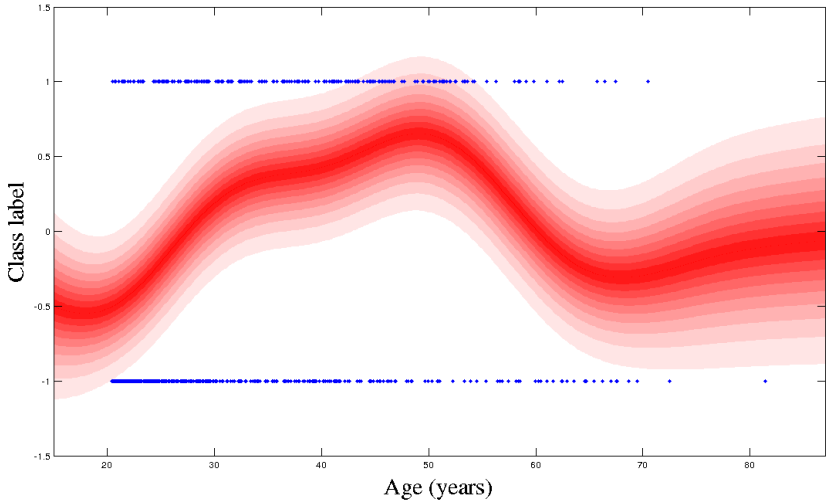
N	True Kernel	SNR = 1 Kernel recovered	SNR = 100 Kernel recovered
100		**1**	**1**
300	**1**	**1**	**1**
500	*3 dimensions*	**1**	**1**
100		**2×3**	**2×3**
300	**2×3**	**2×3**	**2×3**
500	*3 dimensions*	**2×3**	**2×3**
100		1 + 4	1 + 2 + 4
300	**1 + 2×3 + 4**	1×4 + 2×3	**1 + 2×3 + 4**
500	*4 dimensions*	1×4 + 2×3	**1 + 2×3 + 4**
100		4×7	**3×5×7**
300	**3×5×7**	2×3×5×7	**3×5×7**
500	*10 dimensions*	2×3×5×7	**3×5×7**
100		1×3 + 10	1×10
300	**1 + 3×5×7 + 10**	1 + 10	1 + 1×10 + 3×5×7
500	*10 dimensions*	1 + 10	**1 + 3×5×7 + 10**
100		3×7	1 + 7×9
300	**3×5×7×9**	7×9	**3×5×7×9**
500	*10 dimensions*	**3×5×7×9**	**3×5×7×9**
100		9 + 10	10
300	**1 + 3×5×7×9 + 10**	1 + 10	1×5 + 7 + 10
500	*10 dimensions*	1 + 3×5×9 + 10	**1 + 3×5×7×9 + 10**

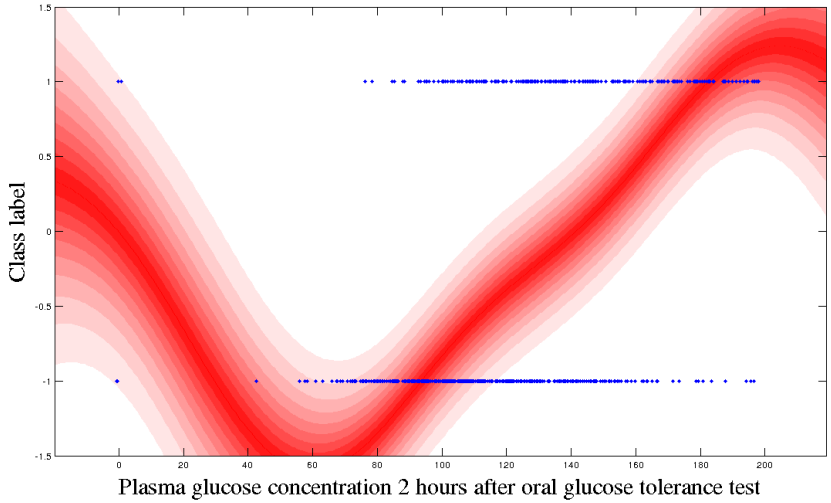
Real-world experiments

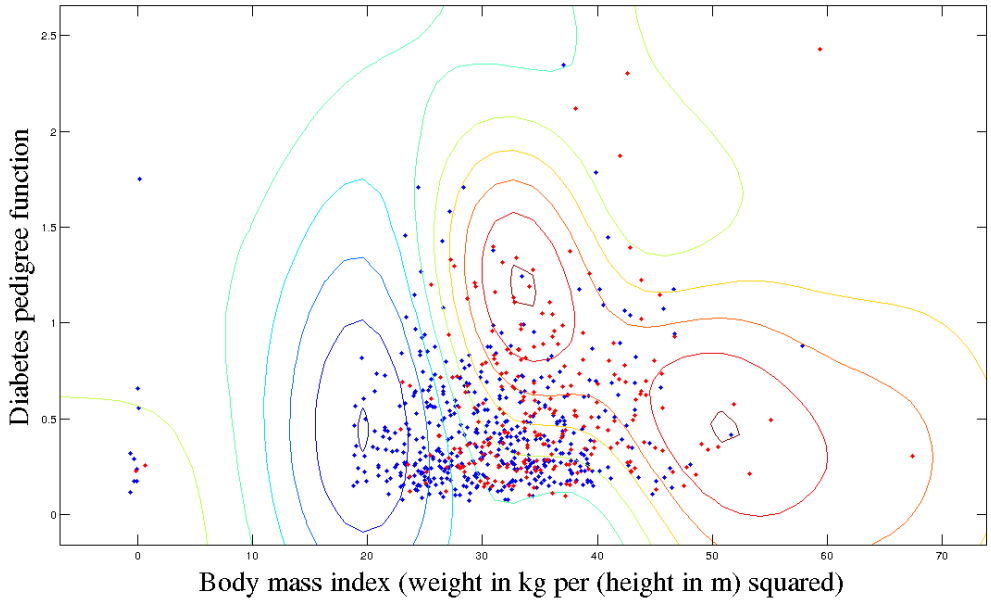
Table 1 : Mean classification errors of the 4 flavours of the GP Structure Search.

Method	Breast	Pima	Liver	Heart
Logistic Regression	*7.61*	*24.39*	*45.06*	**16.08**
GP GAM	**5.19**	**22.42**	**29.84**	**16.84**
HKL	**5.38**	*24.26*	**27.27**	**18.98**
GP Squared-exponential	**4.73**	**23.72**	**31.24**	**20.64**
GP Additive Model	**5.57**	**23.08**	**30.06**	**18.50**
GPSS (AIC)	6.43	**22.53**	28.92	**19.86**
GPSS (BIC)	**5.98**	**23.44**	37.01	**18.15**
GPSS (BIC Light)	6.43	**22.27**	**27.50**	**17.82**
GPSS (Cross Validation)	**5.09**	**23.70**	30.08	**17.16**
Random Forest	**4.22**	**23.44**	**24.03**	**17.13**

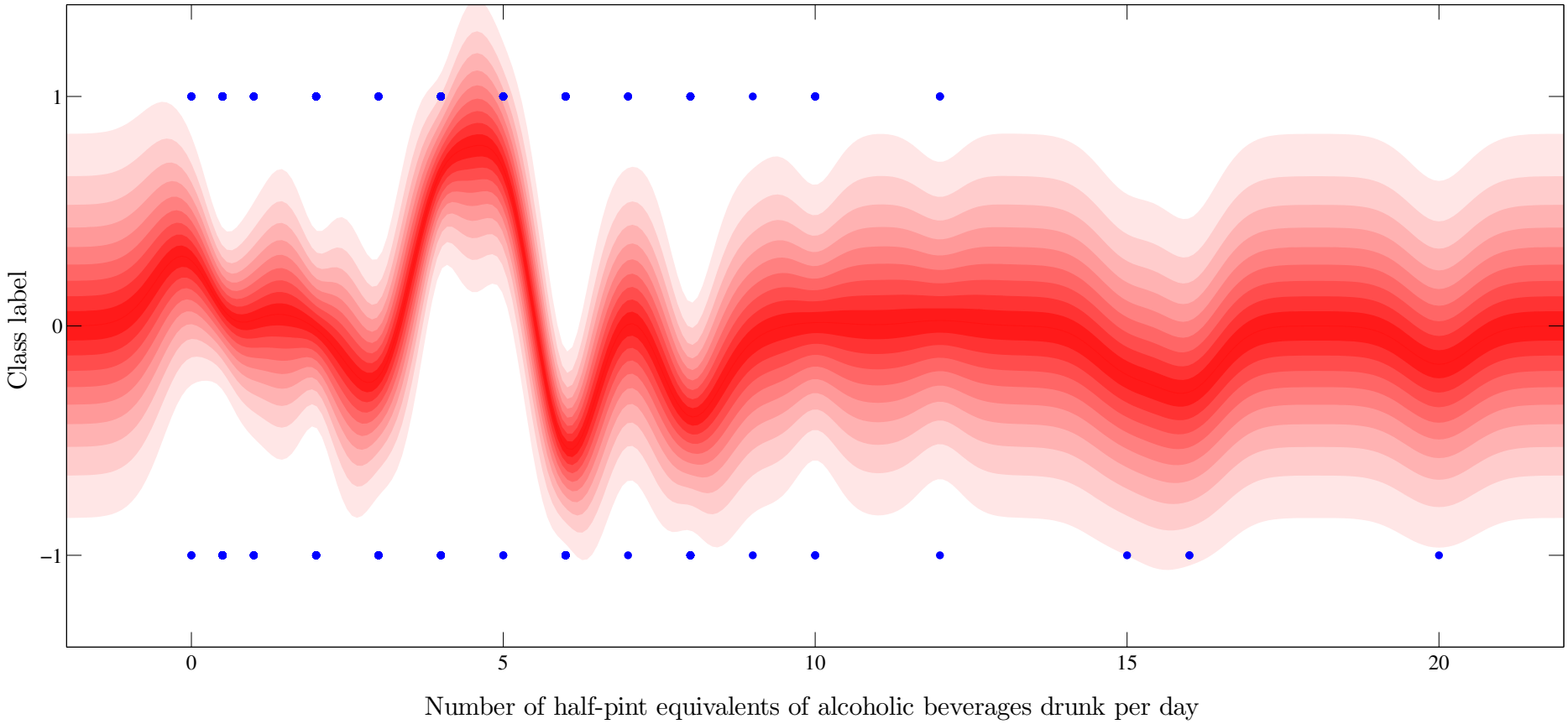
Interpreting the Structure of the Constructed Models







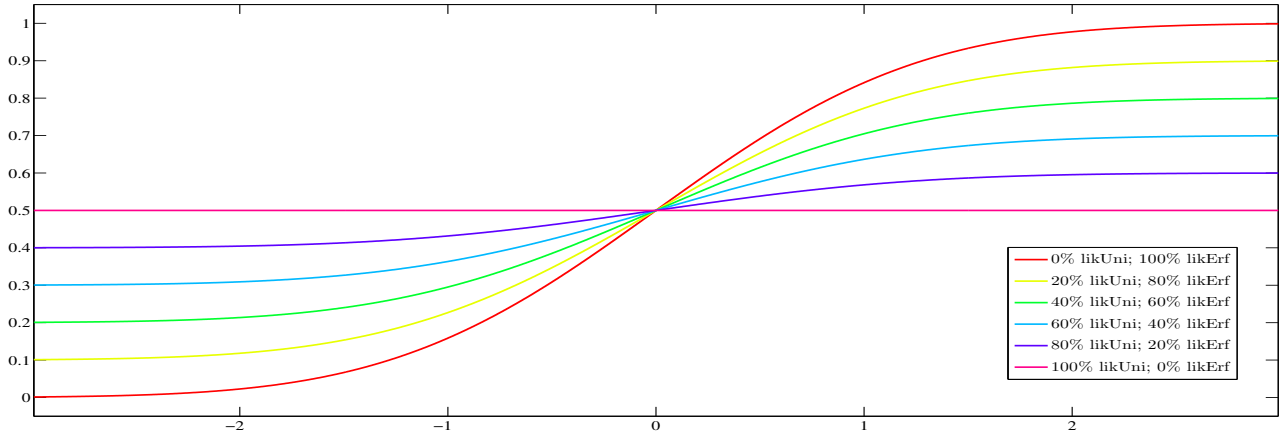
Overfitting with small lengthscales



Constructing Robust Classifiers

- Standard GP classifiers are very sensitive to outliers in the data.
- Classification-specific models such as soft-margin support vector machines are designed with the existence of outliers in mind.

We replaced the error function likelihood with a mixture of the error function and a uniform likelihood, allowing for a certain proportion of examples to have random class membership.



First experiments show that the idea has potential, but issues with the GPML implementation of the likelihood mixtures must be addressed.

Bayesian Model Averaging boosts predictive power

- A fully Bayesian approach would predict using a weighted average of all the models expressible in the kernel grammar.
- A prior over kernel structures should penalise overly complex ones.
- As an ad-hoc approximation, we used an average of the predictions made by the models evaluated during the structure search, weighted by their BIC values:

$$w_i = \frac{e^{-\alpha b_i}}{\sum_j e^{-\alpha b_j}}$$

- Model averaging boosted performance across most data sets, coming closer to that of the Random Forest Algorithm.

Further Work and Acknowledgements

Our Next Goal

Representing higher-order interactions in the data, textually and visually.

Prof Zoubin Ghahramani

James Lloyd

David Duvenaud

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B. and Ghahramani, Z. (2013), Structure Discovery in Nonparametric Regression through Compositional Kernel Search, ICML 2013.
Lloyd, J.R., Duvenaud, D., Grosse, R., Tenenbaum, J.B. and Ghahramani, Z. (2014) Automatic construction and natural-language description of nonparametric regression models, AAAI 2014.