An Efficient Convolutional Framework for Multitask Learning

Michalis K. Titsias [1], Mauricio Alvarez [1], David Luengo [1,2], Neil D. Lawrence [1]
School of Computer Science, University of Manchester, UK [1]
Dpto. Teoría de la Señal y Comunicaciones, Universidad Carlos III, Spain [2]
mtitsias@cs.man.ac.uk , alvarezm@cs.man.ac.uk, luengod@ieee.org, neill@cs.man.ac.uk

Introduction

Structured prediction of multiple outputs (commonly referred to as multi-task learning) presents a problem for kernel methods: how do we best compute the kernel between the different outputs? Several solutions have been suggested (e.g. [1, 2, 3, 4]), but many of them can be seen as corresponding to an affine transformation of the target data followed by independent modelling of outputs. Such linear transformations are clearly limiting. A potentially more powerful class of approximations involves convolutions [5, 6]. Temporal or spatial convolutions contain affine transformations of outputs as a (trivial) special case, providing a much broader class of models. The major problem with these models is their computational complexity, prohibitive for systems involving thousands of data points or outputs.

We approach the problem from a Gaussian process (GP) perspective. A way to model multiple correlated outputs with GPs is to assume that they are generated from a small set of R latent GP functions analogously to the linear latent variable models commonly used in machine learning [3, 4]. A significant generalization of this approach is to combine it with a convolution-based framework as proposed in [5, 6, 7]. Thus, the value $y_q(\mathbf{x})$ for the q-th output at input $\mathbf{x}$ is generated according to

$$y_q(\mathbf{x}) = \sum_{r=1}^{R} \int G_{qr}(\mathbf{x}-\mathbf{z}) f_r(\mathbf{z}) \mathrm{d}\mathbf{z} + \epsilon, \tag{1}$$

where $\epsilon \sim N(0, v_q^2)$, each $G_{qr}(\mathbf{x}-\mathbf{z})$ is a smoothing kernel and the latent functions $f_r(\mathbf{z})$ can be both smooth GP functions or continuous but nowhere differentiable Gaussian white noise processes.

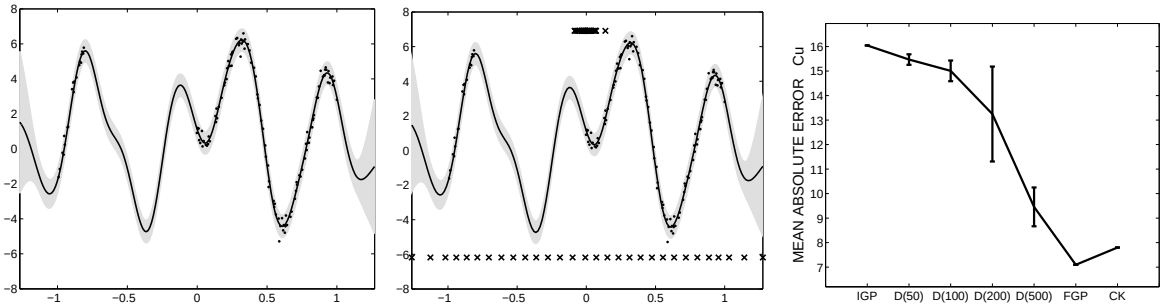
The convolution-based framework is appealing because it is connected with models of linear stochastic differential equations. For example, a special case of (1) corresponds to a Bayesian hierarchical model of linear stochastic differential equations driven by continuous white noise. Thus, one can build models by incorporating prior knowledge about the physical system or phenomenon which generated the output data (as already done in [7] using simple mechanistic assumptions). This is analogous to the idea of "learning the kernel" which has been recently investigated in machine learning (see e.g. [8]). Furthermore, the convolution operation in (1) guarantees that the newly generated process has a positive definite kernel function across the outputs $\{y_q(\mathbf{x})\}_{q=1}^{Q}$. However, even though the convolution-based framework is an elegant way for constructing dependent output processes, the fact that the full covariance function of the joint multi-output model must be considered results in significant computational demand and memory requirements when performing inference, since the size of the covariance matrix now scales as $\mathcal{O}(N^3Q^3)$, where N is the number of observation points per output and Q is the number of outputs.

Proposed method

In this work we propose a variational sparse approximation to the exact full GP model which scales as $\mathcal{O}(NM^2Q)$, where M is the number of support or inducing variables (usually much smaller than N). This method does not modify or discretize the exact GP model, but instead applies a principled variational approximation. In particular, the approximation is derived by minimizing the Kullback-Leibler (KL) divergence between the exact posterior GP and a variational distribution. The support or inducing variables are treated as variational parameters, which are rigorously selected by the minimization of the KL divergence. Additionally, model parameters are inferred by minimizing the difference between a variational lower bound and the exact GP marginal likelihood. This framework builds on previous work described in [9], which deals only with single-output regression. Our method also generalizes the concept of support or inducing variables, which are the basic tools used to construct sparse kernel machines [10]. This generalization is required in order to work with non-smooth latent functions, such as continuous but non-differentiable white Gaussian noise processes, for which current sparse techniques, such as [11], are not applicable. The new inducing variables are now computed by filtering these non-smooth processes through convolution operations using an appropriate smoothing kernel. The parameters of this new kernel are also variational quantities which are rigorously selected again by the minimization of the KL divergence.

Results

For all experiments, the variational and the model smoothing kernel follow a Gaussian form and the necessary integrals are tractable. We first setup a toy problem which consists of $Q = 4$ outputs and $N = 200$ observation points for each output. For the sparse approximations we used $M = 30$ inducing points. We sought the kernel parameters and the positions of the inducing inputs maximizing the variational bound for the marginal likelihood using a scaled conjugate gradient algorithm. For test data we removed a portion of one output (points in the interval $[-0.8, 0]$) as shown in Figure 1. The variational approximation captures the signal even in the missing observation range.



(a) (b) (c)

Figure 1: Predictive mean and variance using the full multi-output GP (in (a)) and the the sparse variational approximation (in (b)) for output 4 in the toy example. Black dots indicate data observations. There is a missing range of observations in the interval $[-0.8, 0]$. For the sparse approximation, the crosses in the upper part of the figure indicate the initial position of the inducing variables. The crosses in the lower part of the figure indicate the position of the inducing variables after the maximization of the variational bound. In (c) we show the mean absolute error and standard deviation for five repetitions of the experiment for the Jura dataset. In the bottom of the figure, IGP stands for independent GP, D(M) indicates sparse approximation with M inducing values, FGP stands for full GP and CK stands for ordinary co-kriging (see [2] for detailed description).

As another example we employ the Jura dataset, which consists of measurements of concentrations of several heavy metal pollutants collected in the topsoil of a 14.5 km^2 region of the Swiss Jura. The data is divided into a prediction set (259 locations) and a validation set (100 locations).[1] We follow the experiment described in [2, pp. 248,249] in which a *primary variable* (copper), at prediction locations, in conjunction with some *secondary variables* (lead, nickel and zinc), at prediction and validation locations, are employed to predict the concentration of the primary variable at validation locations. In figure 1(c), the performance of the approximation in terms of the mean absolute error is compared against independent GPs, the full GP constructed through the convolution approach and ordinary cokriging [2]. The performance of the sparse method approximates the performance of the full GP better as more inducing variables are included. However, as the values of the ouputs are uniformly spread over the input space, it is necessary to include a large amount of inducing variables to reach the performance of the full GP.

Conclusions

We have presented a variational sparse approximation for efficient learning of the kernels within a convolutional framework. We have shown its feasibility in a toy problem and a geostatistical application. Current and future work concentrates on applying it to problems in finance, systems biology and human motion modelling.

References

[1] Evgeniou, T., Micchelli, C. A. & Pontil, M. *Journal of Machine Learning Research* **6**, 615–637 (2005).

[2] Goovaerts, P. *Geostatistics For Natural Resources Evaluation* (Oxford University Press, 1997).

[3] Teh, Y. W., Seeger, M. & Jordan, M. I. In Cowell, R. G. & Ghahramani, Z. (eds.) *AISTATS 10*, 333–340 (2005).

[4] Bonilla, E. V., Chai, K. M. & Williams, C. K. I. In Platt, J. C., Koller, D., Singer, Y. & Roweis, S. (eds.) *NIPS*, vol. 20 (2008).

[5] Higdon, D. M. In Anderson, C., Barnett, V., Chatwin, P. & El-Shaarawi, A. (eds.) *Quantitative methods for current environmental issues*, 37–56 (Springer-Verlag, 2002).

[6] Boyle, P. & Frean, M. In Saul, L., Weiss, Y. & Bouttou, L. (eds.) *NIPS*, vol. 17, 217–224 (2005).

[7] Álvarez, M., Luengo, D. & Lawrence, N. D. (2009). To appear at AISTATS 12.

[8] Bach, F., Lanckriet, G. & Jordan, M. I. In *ICML* (2004).

[9] Titsias, M. K. (2009). To appear at AISTATS 12.

[10] Snelson, E. & Ghahramani, Z. In Weiss, Y., Schölkopf, B. & Platt, J. C. (eds.) *NIPS*, vol. 18 (2006).

[11] Alvarez, M. & Lawrence, N. D. In Koller, D., Schuurmans, D., Bengio, Y. & Bottou, L. (eds.) *NIPS* (2009).

[1] This data is available at `http://www.ai-geostats.org/`