Local Kernel Shaping for Function Approximation

Jo-Anne Ting*, Stefan Schaal
University of Southern California
Los Angeles, CA 90034
joanneti@usc.edu, sschaal@usc.edu

Kernel-based methods have been highly popular in statistical learning, starting with Parzen windows, kernel regression, locally weighted regression, radial basis function networks and leading to newer formulations such as Reproducing Kernel Hilbert Spaces, Support Vector Machines and Gaussian process regression. Most algorithms start out with kernel parameterizations that are the same for all kernels, independent of where in data space the kernel is used, but later recognize the advantage of locally adaptive kernels [1][2][3]. Such locally adaptive kernels are advantageous in scenarios where the data characteristics vary greatly in different parts of the workspace (e.g., in terms of data density and level of nonlinearity). For instance, in Gaussian process regression, using a nonstationary covariance function [4] allows for such a treatment.

In multidimensional systems with d input dimensions and N data points, performing optimizations individually for every kernel becomes infeasible because there are d times more open parameters than constraints (data points). Hence, for effective local kernel adaptation, there should be at least d data points contributing to every kernel. Previous work has suggested gradient descent techniques with cross-validation methods or involved statistical hypothesis testing for optimizing the shape and size of a kernel in a learning system, e.g., [5][6][7]. We propose a complete Bayesian formulation that, with the help of variational approximations, results in an EM-like algorithm for simultaneous estimation of regression parameters and kernel parameters. We consider local kernel shaping by averaging over data points with the help of a locally linear model, and we envision that this approach can be augmented to generalized linear models, e.g., Gaussian Processes, in the near future. A Bayesian approach offers error bounds on the distance metrics and incorporates this uncertainty in the predictive distributions. Evaluations demonstrate that the resulting learning system has excellent local kernel adaptation abilities and that it is able to automatically determine the size of the bandwidth of each local kernel.

*Presenting author

Topic: learning algorithms, control
Preference: oral/poster

References:

[1] **Friedman, J.H.**, "A variable span smoother", Technical Report, Stanford University, (1984).

[2] **Poggio, T. and Girosi F.**, "Regularization algorithms for learning that are equivalent to multilayer networks", Science 247:213-225 (1990).
[3] **Fan, J. and Gijbels, I.**, "Local polynomial modeling and its applications", London: Chapman & Hall (1996).
[4] **Paclorek, C.J. and Schervish, M.J.**, "Nonstationary covariance functions for Gaussian process regression", Advances in Neural Information Processing Systems 16, MIT Press (2004).
[5] **Fan, J. and Gijbels, I.**, "Variable bandwidth and local linear regression smoothers", The Annals of Statistics 20:2008-2036 (1992).
[6] **Fan, J. and Gijbels, I.**, "Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation", Journal of the Royal Statistical Society B 57:371-395 (1995).
[7] **Schaal, S. and Atkeson, C.G.**, "Assessing the quality of learned local models", Advances in Neural Information Processing Systems 6, Morgan Kaufmann (1998).