Designing Kernel Functions Using the Karhunen-Loève Expansion

Masashi Sugiyama[1,2] and Hidemitsu Ogawa[2]

[1]Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489, Berlin, Germany

[2]Department of Computer Science, Tokyo Institute of Technology,
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

1 Introduction

In recent years, a number of kernel-based learning algorithms such as the regularization networks [1], the support vector machines [7, 4, 5], and the Gaussian process regression [8] have been investigated. These kernel machines are shown to work very well on real-world problems, given appropriate *kernel functions*. For general purposes, the Gaussian kernel is widely used and seems to work well [5]. On the other hand, a lot of attention have been paid recently to designing kernel functions using the problem-dependent prior knowledge. Various methods for constructing suitable kernels have been proposed [7, 3, 9, 6, 4]. In this contribution, we propose a framework for designing kernel functions for regression.

2 A Kernel Design Framework for Regression

Let us consider a one-dimensional regression problem. Kernel regression tries to approximate an unknown function $f(x)$ by the sum of kernel functions centered at training input points:

$$\hat{f}(x) = \sum_{i=1}^{n} K(x, x_i), \tag{1}$$

where $K(x, x')$ is a kernel function and $\{x_i\}_{i=1}^n$ are training input points. Roughly speaking, the target function $f(x)$ is locally approximated by the kernel function. For this reason, we consider the problem of approximating local functions by a single kernel function.

Let Ω be a set of all local functions $\{\Pi(x)\}$. Let $\mathcal{H}_\Pi$ be a functional Hilbert space that includes Ω. Then the well-known *Karhunen-Loève expansion* [2] asserts that the best approximation to the set Ω of all local functions $\{\Pi(x)\}$ is given by the eigenfunction $\phi_0(x)$ associated with the largest eigenvalue λ_0 of the correlation operator R of the local functions $\{\Pi(x)\}$. More specifically, it holds that

$$\phi_0 = \underset{\phi \in \mathcal{H}_\Pi,\ \|\phi\|=1}{\operatorname{argmin}} \ E\|\Pi - \langle \Pi, \phi \rangle \phi\|^2, \tag{2}$$

where E denotes the expectation over Π, $\|\cdot\|$ denotes the norm in $\mathcal{H}_\Pi$, and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathcal{H}_\Pi$. Note that $\langle \Pi, \phi \rangle \phi$ is the orthogonal projection of Π onto the subspace spanned by ϕ, provided $\|\phi\| = 1$. The leading eigenfunction $\phi_0(x)$ is referred to as the *principal component* of R. Based on this fact, we propose using the principal component $\phi_0(x)$ as the kernel function, i.e.,

$$K(x, x') = \phi_0 \left(\frac{|x - x'|}{c} \right), \tag{3}$$

where c is a positive scalar that controls the kernel width. We call the above kernel the *principal component (PC) kernels*. For multi-dimensional regression problems, PC kernels may be constructed by the componentwise product of the one-dimensional PC kernel or by replacing $|x - x'|$ with the Euclidean norm.

For actually obtaining the PC kernels, the probability distribution of the local functions and the topology in the functional Hilbert space $\mathcal{H}_\Pi$ (i.e., the inner product and norm) should be specified. If a prior knowledge of the probability distribution of the local functions is available, it can be effectively incorporated into constructing the kernel function. In the absence of such a prior knowledge, we may use a non-informative prior knowledge such as a uniform distribution.

3 Constructing Kernel Functions for Binary Regression

Using the above framework, we design a kernel function for binary regression problems, where the learning target function $f(x)$ is binary. In the binary regression case, Ω is a set of all rectangle functions $\{\Pi(x)\}$ with different widths.

In order to obtain the PC kernels, the probability distribution of the rectangle functions should be specified. Here let us treat the width of the rectangle functions as probabilistic. Since we do not have any prior knowledge of the width of the rectangle functions, we use the uniform distribution on a closed interval. If the standard L_2-norm is used in the functional Hilbert space $\mathcal{H}_\Pi$, we have the following PC kernel.

$$K(x, x') = \begin{cases} \cos\left(\dfrac{|x - x'|}{c}\right) & \text{if } \dfrac{|x - x'|}{c} \leq \dfrac{\pi}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

References

[1] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

[2] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, Inc., Boston, second edition, 1990.

[3] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, Cambridge, MA., 1999. MIT Press.

[4] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[5] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines.* World Scientific Pub. Co., Singapore, 2002.

[6] K. Tsuda, T. Kin, and K. Asai. Marginalized kernels for biological sequences. *Bioinformatics*, 18(1):268–275, 2002.

[7] V. N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc., New York, 1998.

[8] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 514–520. The MIT Press, 1996.

[9] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.