Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.

Bernhard SCHÖLKOPF and Alexander SMOLA. Cambridge, MA: MIT Press, 2002. ISBN 0-262-19475-9. xviii + 626 pp. $60.00 (H).

Learning Kernel Classifiers.

Ralf HERBRICH. Cambridge, MA: MIT Press, 2002. ISBN 0-262-08306-X. xx + 364 pp. $40.00 (H).

Over the last 5 years or so, a major area of research activity in the machine-learning research community has been that of kernel machines. The best-known example of these is the support vector machine (SVM), derived from the work of Vladimir Vapnik and coworkers. Although some statisticians are active members of this research community, many are not, thus I start this review by outlining some of the basic ideas of kernel machines, and relating them to more familiar ideas from the statistical literature, before going on to consider the specific merits of the two books.

The basic idea behind kernel methods is that we start with an input pattern $\mathbf{x}$ and rerepresent it in a feature space as $\phi(\mathbf{x})$. A simple example of this would be a polynomial feature space in which a two-dimensional input pattern $\mathbf{x} = (x_1, x_2)^T$ would be represented under a quadratic polynomial transformation as $\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)^T$. The key idea behind most kernel algorithms is the so-called kernel trick: If algorithms making use of the feature space require only the dot product in feature space $\phi(\mathbf{x}) \cdot \phi(\mathbf{x}')$ of two input patterns $\mathbf{x}$ and $\mathbf{x}'$, and if this dot product can be computed by a kernel function $k(\mathbf{x}, \mathbf{x}')$, then there is no need to work explicitly in the feature space. For the foregoing quadratic example, the corresponding kernel is $(\mathbf{x} \cdot \mathbf{x}')^2$.

Of course, polynomial regression has a long history. Things can get more interesting with other kernels; for example, the widely used Gaussian radial basis function kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$ corresponds to a feature expansion in an infinite-dimensional feature space.

Kernel machines were initially used for supervised learning (i.e., prediction) problems, for both classification and regression. Thanks to the kernel trick, it turns out that the predictor at a test point $\mathbf{x}$ will be of the form $h(f(\mathbf{x}))$ with $f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i)$, where h is a given function, the α_i's are real coefficients, and $\{\mathbf{x}_i\}_{i=1}^n$ is the training set of patterns. For a two-class classification problem, h may be the sign function, whereas for a regression problem it may be the identity function. Of course, prediction methods of this form are not unknown in the statistical literature. Examples of such "kernel machines" include spline methods (e.g., Wahba 1990; Green and Silverman 1994) and Gaussian process models, and these can be related to notions of regularization in function space and to reproducing kernel Hilbert spaces. Gaussian process models have been used extensively in geostatistics (e.g., Cressie 1993) and also used in machine-learning contexts (e.g., Williams and Rasmussen 1996).

One important difference of the support vector machine to these methods is that the expansion $\sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i)$ is typically sparse; that is, many of the coefficients α_i are 0. Indeed, those training patterns that have non-0 α's are known as support patterns. From where does this sparsity derive? If we consider a classification problem (with classes labeled as $y = \{+1, -1\}$), then for spline prediction it would be natural to use $f(\mathbf{x})$ to model the log-odds ratio $\log P(y = +1|\mathbf{x})/P(y = -1|\mathbf{x})$. This corresponds to a negative log-likelihood term of $\log(1 + e^{-y_i f_i})$, where $f_i = f(\mathbf{x}_i)$. However, the SVM uses an alternative "data-fit" term $[1 - y_i f_i]_+$, which is broadly similar to the log-likelihood term but with a key difference that it is exactly 0 for $y_i f_i \geq 1$. The term $[1 - y_i f_i]_+$ is regarded as a soft constraint: if $y_i f_i \geq 1$, then the constraint is satisfied; otherwise, a penalty is paid for its violation. It is this feature that gives rise to the sparsity of the SVM; if a training point is safely classified (so that $y_i f_i \geq 1$), then the constraint is not active, and the point does not contribute to the expansion.

Under a Gaussian process prior on f, the optimization problem to find the value of $\mathbf{f} = (f_1, f_2, \ldots, f_n)^T$ that minimizes the negative log posterior $\frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} + \sum_{i=1}^n \log(1 + e^{-y_i f_i})$ is convex, where K is the $n \times n$ matrix with entries $k(\mathbf{x}_i, \mathbf{x}_j)$. The solution to the corresponding optimization for SVMs gives rise to a convex quadratic programming problem. The convexity of both of these optimization problems contrasts with earlier work on artificial neural networks, where local optima in the parameter space were a serious problem.

A sparse solution can also be obtained in regression problems by using the ϵ-insensitive loss function $\max(0, |y_i - f(\mathbf{x}_i)| - \epsilon)$ instead of the standard squared loss $(y_i - f(\mathbf{x}_i))^2$ that corresponds to a Gaussian noise model.

The algorithmic work on SVMs described here has been complemented by theoretical analysis of the generalizability of such machines, particularly under the probably-approximately-correct (PAC) framework. Of course, it is not only theoretical considerations that lead to interest in SVMs. Probably more important have been the reports of world-class prediction performance on a wide variety of tasks, including handwritten digit recognition, text classification, and some bioinformatics problems.

Kernel methods have been applied beyond prediction problems. Examples include lifting principal-components analysis (PCA) into feature space to obtain kernel PCA, and one can also carrying out one-class learning (similar to class-conditional density estimation) for problems such as novelty detection using kernel machines.

I now turn to the two books under review. *Learning With Kernels* is, as its length suggests, comprehensive. It is divided into three parts: I, Concepts and Tools; II, Support Vector Machines; and III, Kernel Methods, plus a tutorial introduction (Chap. 1) to kernel machines, which focuses mainly on the SVM but with some mention of kernel PCA.

Part I lays the groundwork for the book. There are chapters describing the statistical framework for prediction problems, defining and giving numerous examples of kernels (both positive definite and conditionally positive definite), regularization theory, statistical learning theory [uniform convergence theory and Vapnik–Chervonenkis (VC) bounds], and optimization theory (for both unconstrained and constrained problems).

Part II focuses is on SVMs. The particular form of the SVMs for classification, regression, and single-class problems and their corresponding basic algorithms are derived. There is also a chapter on implementation issues, which is very important because, naïvely, the time complexity of the quadratic-programming problems obtained would scale as $O(n^3)$, and careful approximations and exploitation of the sparsity of the SVM solution are needed to produce efficient solutions when faced with tens of thousands of training examples.

Part III considers a number of topics. One area covered is that of other kernelized algorithms, such as kernel PCA and the kernel Fisher discriminant. Also covered is the very important area of designing kernels; of particular interest here is the work (originated by C. Watkins and D. Haussler) on string kernels, where inputs $\mathbf{x}$ and $\mathbf{x}'$ are strings of symbols and similarity is measured in terms of the common substrings that the two input patterns contain. The book focuses mainly on the regularization theory point of view, but Chapter 16 also covers the Bayesian (Gaussian process) view of kernel machines.

Schölkopf and Smola have been two of the leading lights in kernel machines research for many years, and this book is likely to become one of the standard references—perhaps the standard reference—in the area. It gives a comprehensive and detailed treatment of most topics in the kernel-machines research area and extensive pointers to the literature. It also contains around 300 problems, ranging from simple to hard and on to open problems and questions for further research.

By way of comparison, the book by Cristianini and Shawe-Taylor (2000) provides an attractive, concise, and very readable introduction (in around 200 pages) to SVMs and related algorithms. It covers similar material to much of parts I and II of *Learning With Kernels*, although in rather less detail in some areas. The text of Vapnik (1998) provides a comprehensive treatment of statistical learning theory and includes a large amount of material on SVMs. However, it focuses mainly on the supervised-learning problem, and naturally is unable to cover important developments that have occurred since its publication.

Learning Kernel Classifiers developed from the author's doctoral thesis, written at the Technische Universität Berlin. Herbrich has been a leading researcher in the kernel community for a number of years, and his book concentrates on the supervised-learning problem, and indeed on the classification problem (as opposed to regression). It is divided into two main parts. Part I consists of two chapters. In the first chapter, kernels and support vector classification are developed. The second chapter takes a Bayesian view of the problem, looking at Gaussian process approaches to classification and other Bayesian techniques, such as the relevance vector machine and the Bayes point machine. Part II uses techniques from statistical learning theory to analyze the generalization ability of the algorithms presented in Part I. Chapter 4 provides PAC and VC analyses and the describes luckiness framework of Shawe-Taylor et al. (1998). This

allows PAC generalization error bounds to be applied in situations where the complexity bound depends on a data-dependent quantity (such as the margin in SVMs). In contrast with Chapter 4, Chapter 5 derives performance bounds for specific algorithms. These include PAC-Bayesian analysis of Bayesian algorithms and analysis of compression schemes (i.e., learning schemes where only a subset of the training points have an effect on the prediction, as in SVMs). In addition to the main part of the book (195 pages), two extensive appendixes (77 pages) give detailed derivations of theorems presented in the main text. The book does not provide exercises for the reader.

Clearly, Herbrich's book has a more limited scope than Schölkopf and Smola's. However, it does deal in more depth with Bayesian approaches to the classification problem (Chap. 3), and also gives thorough and detailed expositions of the learning theory material in Chapters 4 and 5.

In summary, these books have different niches. I would recommend the book by Cristianini and Shawe-Taylor (2000) for a reader looking to understand the basics of support vector machines, *Learning With Kernels* for someone looking for a comprehensive and in-depth treatment of the wide diversity of kernel algorithms, and *Learning Kernel Classifiers* for a reader looking to understand kernel classifiers and the learning theory that has been developed for them.

Looking to the future, I expect that the development of novel kernels, particularly those incorporating prior/domain knowledge, will be important. A second key issue for kernel methods is developing good approximation algorithms for the numerical optimizations encountered when dealing with large datasets. The website *http://www.kernel-machines.org/* acts as a central information source in the kernel world, allowing the interested reader to keep abreast of developments (as well as to obtain preprints, software, etc.).

Christopher K. I. WILLIAMS
University of Edinburgh

REFERENCES

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, New York: Wiley.

Cristianini, N., and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge, U.K.: Cambridge University Press.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman & Hall.

Shawe-Taylor, J. et al. (1998), "Structural Risk Minimization Over Data-dependent Hierarchies," *IEEE Transactions on Information Theory*, 44, 1926–1940.

Vapnik, V. (1998), *Statistical Learning Theory*, New York: Wiley.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics.

Williams, C. K. I., and Rasmussen, C. E. (1996), "Gaussian Processes for Regression," in *Advances in Neural Information Processing Systems* 8, eds. D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Cambridge, MA: MIT Press, pp. 514–520.