An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods

A Review

Tong Zhang

This book is an introduction to support vector machines and related kernel methods in supervised learning, whose task is to estimate an input-output functional relationship from a training set of examples. A learning problem is referred to as *classification* if its output take discrete values in a set of possible categories and *regression* if it has continuous real-valued output.

A simple and useful model of an input-output functional relationship is to assume that the output variable can be expressed approximately as a linear combination of its input vector components. These linear models include the linear least squares method for regression and the logistic regression method for classification. Because a linear model has limited prediction power by itself, there has been extensive research in nonlinear models such as neural networks. However, there are two major problems with the use of nonlinear models: First, they are theoretically difficult to analyze, and second, they are computationally difficult to solve. Linear methods have recently regained their popularity because of their simplicity both theoretically and computationally. It has also been realized that with appropriate features, the prediction power of linear models can be as good as nonlinear models. For example, one can linearize a neural network model by using weighted averaging over all possible neural networks in the model.

To use a linear model to represent a nonlinear functional relationship, we need to include nonlinear feature components. A useful technique for constructing nonlinear features is *kernel methods*, where each feature is a function of the current input and one of the example input (such as their distance). This idea has recently received much attention because of the introduction of support vector machines (SVMs) and the renewed interest in Gaussian processes. SVMs, introduced by Vapnik and his collaborators, were originally formulated for binary-classification problems. The resulting method is similar to penalized maximum-likelihood estimators using logistic models (also known as conditional maximum entropy classification models). Later, SVMs were extended to regression problems, and the resulting formulations are similar to ridge regression.

An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Nello Cristianini and John Shawe-Taylor, Cambridge University Press, Cambridge, U.K., 2000, 189 pp., ISBN 0-521-78019-5.

The general framework of support vector learning includes the following components: (1) regularized linear learning models (such as classification and regression), (2) theoretical bounds, (3) convex duality and the associated dual-kernel representation, and (4) sparseness of the dual-kernel representation. Although most of these concepts are not new, the combination is unique, which makes SVMs and related kernel-based learning methods special and interesting. As a result, hundreds of research papers have been published in recent years on different aspects of this new learning methodology. SVMs have also successfully been applied in practice, especially for classification problems. Although many problems that have been successfully solved by SVMs could also have been solved successfully by standard statistical methods such as penalized logistic regression, in practice, SVMs can still have significant computational advantages.

Certain aspects of SVMs have already been described in Vapnik's monograph entitled *Statistical Learning Theory* (Wiley, 1998). However, this book is highly mathematical, which makes it intimidating to an average reader. Thus, it is not an ideal introductory textbook. However, because of the popularity of SVMs in recent years, there is a need for a gentle introduction to the subject that is suitable for the average reader.

The book by Cristianini and Shawe-Taylor successfully fulfills the demand for such a gentle introduction. It is structured as a self-contained textbook for machine learning students and practitioners. However, the book is also a valuable reference for active researchers working in this field. A useful aspect of the book is that it contains a literature survey and discussion section at the end of every chapter. Pointers to many of the references can also be found on a dedicated web site.[1] The discussion broadens the reader's view of relevant topics, and the references point the reader to some important recent research works. Because the authors of the book are active researchers in the field who have made tremendous contributions themselves, the information chosen in

the book reflects their unique, often insightful perspectives.

The book contains eight chapters, which cover topics including the mathematical foundation of kernels, Vapnik-Chervonenkis–style learning theory, duality in mathematical programming, SVM formulations, implementation issues, and successful SVM applications. Although these topics are not comprehensive, they are carefully chosen. Some more advanced materials are left out or barely mentioned in the book. Often, these advanced topics are still under active research and, thus, are not mature enough to be included in an introductory text. Among the missing topics I consider important are (1) approximation properties of kernel representation, (2) non–Vapnik-Chervonenkis–style and non-margin–style theoretical analysis, (3) the relationship between SVM and regularization methods in statistics and numeric mathematics, (4) the impact of different regularization conditions and the impact of different loss terms, (5) direct-kernel formulations in the primal form (these formulations have some advantages over the dual-kernel formulations considered in the book), and (6) sparseness properties of the dual-kernel representation in an SVM (this aspect of SVM is quite important and relates to many ideas that existed before its introduction).

The book starts with a brief introduction to different learning models and then focuses on linear learning machines using perceptrons and least squares regression. Many other linear learning machines are not mentioned in the book, probably because they are less relevant to dual-kernel methods. The concept of kernel-induced feature space is developed in a subsequent chapter. The authors include some novel and very interesting kernel examples, such as a string kernel that induces a feature space consisting of all possible subsequences in a string. The main advantage of kernel-induced feature spaces is the ability to represent nonlinear functions linearly in a space spanned by the kernel functions. Although a kernel representation can naturally be obtained from the dual form of SVMs as presented in the book, it can also be formulated directly in a primal-form learning machine.

The book has a nicely written chapter on the learning theory aspect of SVMs, mainly from the Vapnik-Chervonenkis analysis and the maximal margin point of view. The authors have made substantial contributions in the theoretical development of the field, and the material selection in this chapter reflects their perspective on the problem. These theoretical results are later used to motivate some specific SVM formulations considered in the book. As acknowledged by the authors, this approach only leads to learning methods that reflect what are given in the theoretical bounds, which might not be optimal. In addition, for computational reasons, in the later proposed formulations, there are some heuristics not directly reflected in the theoretical results. For these reasons, the theory itself does not justify SVMs as a method superior to other standard statistical methods, such as logistic regression for classification or least squares for regression. In fact, one might start with non-SVM formulations and develop similar theoretical results.

Following the theoretical analysis chapter, the book gives an introduction of optimization and Lagrangian duality theory, which is an essential component of support vector learning. Motivated by the theoretical bounds given earlier, a number of specific SVM formulations are presented. By using the Lagrangian duality theory, these SVMs are transformed into dual forms using kernel representation. Some implementation issues that are useful for practitioners are carefully discussed. Although this book mainly considers the standard approach to obtaining dual-kernel representation using Lagrangian duality, we can also use kernel representation directly in primal-form linear learning machines. The latter approach does not require the Lagrangian duality theory and can be made either equivalent or non-equivalent to the dual-kernel formulation.

Although this book does not include all aspects of learning methods related to SVMs, it achieves the right balance of topic selection, clarity, and mathematical rigor as an introductory text. The topics covered in the book are the most fundamental and important ones in support vector learning and related kernel methods. This book is essential both for practitioners who want to quickly learn some implementation tricks to apply SVMs to real problems and for theoretically oriented readers who want to understand the mathematical foundations of the underlying theory. It is also a handy reference book for researchers working in the field.

Note

1. www.support-vector.net.

Tony Zhang received a B.A. in mathematics and computer science from Cornell University in 1994 and a Ph.D. in computer science from Stanford University in 1998. Since 1998, he has been with IBM Research, T. J. Watson Research Center, Yorktown Heights, New York, where he is now a research staff member in the Mathematical Sciences Department. His research interests include machine learning, numeric algorithms, and their applications. His e-mail address is tzhang@watson.ibm.com.