An Introduction to Support Vector Machines

A Review

Yiling Chen and Isaac G. Councill

In the preface of the book, Cristianini and Shawe-Taylor state that their intention is to present an organic, integrated introduction to support vector machines (SVMs). The authors believe that SVMs are a topic now sufficiently mature that it should be viewed as its own subfield of machine learning. SVMs, first introduced by Vladimir Vapnik, are a type of linear learning machines much like the famous perceptron algorithm and, thus, function to classify input patterns by first being trained on labeled data sets (supervised learning). However, SVMs represent a significant enhancement in function over perceptrons. The power of SVMs lies in their use of nonlinear kernel functions that implicitly map input into high-dimensional feature spaces. In the *high-dimensional feature spaces,* linear classifications are possible; they become nonlinear in the transformation back to the original input space. Thus, although SVMs are linear learning machines with respect to the high-dimensional feature spaces, they are in effect nonlinear classifiers.

The authors review and synthesize a wide range of materials, including the dual representation characteristic of linear learning machines, feature spaces, learning theory, generalization theory, and optimization theory, that are necessary for a comprehensive introduction to SVMs. The topics are introduced in an iterative and problem-triggered manner: Problems are presented, followed by concepts

■ *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Nello Cristianini and John Shawe-Taylor, New York, Cambridge University Press, 2000, 189 pp., $45, ISBN 0-521-78019-5.

and methods that can overcome them but can cause new problems, followed by new theories to solve new problems, and so on. In this way, readers are naturally attracted to follow the logical path leading to the discovery of SVMs and are exposed to the advantages and beauty of SVMs along the way.

The book is divided into eight chapters. The first, "The Learning Methodology," introduces basic machine learning concepts and provides a road map for the rest of the book. In this chapter, the generalization criterion is presented to the reader along with the classic problem of overfitting.

In the second chapter, "Linear Learning Machines," linear classification and regression are presented. As the best understood and simplest learning machines, linear learning machines provide the framework within which more complex systems will be constructed. However, because of their simplicity, basic linear learning machines are only suitable for linearly separable problems. Real-world applications often require a more expressive hypothesis space than linear functions can provide. In

the last section of chapter 2, the dual representation of linear learning machines is introduced. Dual representation is one of the crucial concepts in developing SVMs.

The limited computational power of linear learning machines leads to the topic of the third chapter, "Kernel-Induced Feature Spaces." To increase the computational power of the linear learning machines, nonlinear mappings can be used to transform the data into a high-dimensional feature space in which a linear learning methodology is then applied. Kernel functions can implicitly combine these two steps (nonlinear mapping and linear learning) into one step in constructing a nonlinear learning machine. A linearly inseparable problem can become linearly separable in a higher-dimensional feature space. As a consequence of the dual representation of linear learning machines, the dimension of the feature space need not affect the computation because only the inner product is computed by evaluating the kernel function. The use of kernel functions is an attractive computational shortcut.

The use of kernel functions to construct nonlinear learning machines greatly increases the expressive power of learning machines and retains the underlying linearity that ensures the tractability of learning. However, the increased flexibility increases the risk of overfitting, which can lead to bad generalization performance. Chapter 4, "Generalization Theory," introduces the theory of Vapnik and Chervonenkis (VC) to control the increased flexibility of kernel-induced feature space and lead to good generalization. Loosely speaking, the most important result of VC theory is that the upper bound of the generalization risk for a learning machine is controlled by the empirical risk and the VC dimension, which is fixed for a hypothesis space of the learning machine.

The theory presented in chapter 4 shows that the learning machine with the lowest upper bound of generalization risk is achieved by selecting a machine that minimizes the empirical risk. This discussion sets

the stage for the topic of chapter 5, "Optimization Theory." *Optimization theory* deals with problems of finding a vector of parameters that minimizes or maximizes a certain cost function, typically subject to some constraints. Chapter 5 focuses on some of the results of optimization theory that apply to cases where the cost function is a convex quadratic function, but the constraints are linear. This class of optimization problems is what is needed for training SVMs.

The materials discussed in the previous five chapters form the theoretical foundation of SVMs. Chapter 6 brings these topics together to introduce the SVM learning system. The chapter discusses both support vector classification and support vector regression. In chapter 7, "Implementation Techniques," the authors introduce some specific techniques that have been developed to exploit SVM training. Chapter 8, "Applications of Support Vector Machines," illustrates the successful applications of SVMs to text categorization, image recognition, handwritten digit recognition, and bioinformatics.

The book is an excellent introduction to SVM learning systems. Despite the fact that it covers a wide range of material, the book presents the concept gradually in accessible and self-contained stages and consistently steers away from the deeper theoretical side of learning machines without sacrificing too much mathematical rigor. The book contains plenty of pseudocode examples and exercise questions and an excellent reference of related subjects. Most (but not all) of the background mathematics required to understand SVM theory (including vector spaces, inner product spaces, Hilbert spaces, and eigenvalues) are presented in an appendix. The reader needs a fair background of linear algebra and matrix theory to experience the excitement of the book. However, we feel that a book such as this belongs in the personal library of everyone with a serious interest in machine learning.

Yiling Chen is a Ph.D. candidate at the School of Information Sciences and Technology at the Pennsylvania State University. She received her B.S. in economics from Renmin University of China and her M.S. in finance from Tsinghua University, China. Her research interests include electronic commerce, machine learning, and information integration. Her e-mail address is yxc194@psu.edu.

Isaac G. Councill is a graduate student at the Pennsylvania State University School of Information Sciences and Technology. His research interests include developing high-level ontologies for structuring knowledge representations and modeling affective states in cognitive models. His e-mail address is igc2@psu.edu.

Natural Language Processing and Knowledge Representation

Language for Knowledge and Knowledge for Language

edited by

Łucja M. Iwańska
and **Stuart C. Shapiro**

Published by AAAI Press
www.aaaipress.org
650-328-3123