An Introduction to Support Vector Machines
A Review

Yiling Chen (yxc194@psu.edu)
Isaac G. Councill (igc2@psu.edu)

School of Information Sciences and Technology
The Pennsylvania State University
004 Joab L. Thomas Building
University Park, PA 16802

2/24/03

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) which,
the authors believe, is 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 machine 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 functionality over perceptrons.
The power of SVMs lies in their use of non-linear kernel functions that implicitly map
input into high dimensional "feature spaces." In feature spaces linear classifications may
be possible, which become non-linear in the transformation back to the original input
space. Thus, although SVMs are linear learning machines with respect to feature spaces,
they are in effect non-linear 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 which 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 and methods that can overcome
them but may 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 roadmap 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, due to 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 two, 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."  In order 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 may 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 as only the inner product is computed by evaluating the kernel function.  The use of kernel functions is an attractive computational short-cut.

The use of kernel functions to construct nonlinear learning machines greatly increases the expressive power of learning machines while retaining the underlying linearity that ensures the tractability of learning.  However, the increased flexibility increases the risk of overfitting which may 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 that has the lowest upper bound of generalization risk is achieved by selecting a machine that minimizes the empirical risk.  This 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 that the cost function is a convex quadratic function, while 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 SVMs training. Chapter 8, "Applications of Support Vector Machines", illustrates the successful applications of SVMs to text categorization, image recognition, hand-written 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 rigour.  The book contains plenty of pseudo-code examples, exercise questions, and an excellent reference of related subjects.  Most of the background mathematics required to understand SVM theory

(including vector spaces, inner product spaces, Hilbert spaces, and eigenvalues) are presented in an appendix, though not all. 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.