Kernel Methods and Machine Learning

Offering a fundamental basis in kernel-based learning theory, this book covers both statistical and algebraic principles. It provides over 30 major theorems for kernel-based supervised and unsupervised learning models. The first of the theorems establishes a condition, arguably necessary and sufficient, for the kernelization of learning models. In addition, several other theorems are devoted to proving mathematical equivalence between seemingly unrelated models.

With nearly 30 closed-form and iterative algorithms, the book provides a step-by-step guide to algorithmic procedures and analyzing which factors to consider in tackling a given problem, enabling readers to improve specifically designed learning algorithms and to build models for new application paradigms such as green IT and big data learning technologies.

Numerous real-world examples and over 200 problems, several of which are MATLAB-based simulation exercises, make this an essential resource for undergraduate and graduate students in computer science, and in electrical and biomedical engineering. It is also a useful reference for researchers and practitioners in the field of machine learning. Solutions to some problems and additional resources are provided online for instructors.

S. Y. Kung is a Professor in the Department of Electrical Engineering at Princeton University. His research areas include VLSI array/parallel processors, system modeling and identification, wireless communication, statistical signal processing, multimedia processing, sensor networks, bioinformatics, data mining, and machine learning. He is a Fellow of the IEEE.

Kernel Methods and Machine Learning

S. Y. KUNG
Princeton University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314-321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi - 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107024960

First published 2014

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Kung, S. Y. (Sun Yuan)
Kernel methods and machine learning / S.Y. Kung, Princeton University, New Jersey.
pages cm
ISBN 978-1-107-02496-0 (hardback)
1. Support vector machines. 2. Machine learning. 3. Kernel functions. I. Title.
Q325.5.K86 2014
006.3´10151252–dc23 2014002487

ISBN 978-1-107-02496-0 Hardback

Additional resources for this publication at www.cambridge.org/9781107024960

To Jaemin, Soomin, Timmy, and Katie, who have been our constant source of joy and inspiration

Cambridge University Press
978-1-107-02496-0 — Kernel Methods and Machine Learning
S. Y. Kung
Frontmatter
More Information

Contents

Preface

Machine learning is a research field involving the study of theories and technologies to adapt a system model using a training dataset, so that the learned model will be able to generalize and provide a correct classification or useful guidance even when the inputs to the system are previously unknown. Machine learning builds its foundation on linear algebra, statistical learning theory, pattern recognition, and artificial intelligence. The development of practical machine learning tools requires multi-disciplinary knowledge including matrix theory, signal processing, regression analysis, discrete mathematics, and optimization theory. It covers a broad spectrum of application domains in multimedia processing, network optimization, biomedical analysis, etc.

Since the publication of Vapnik's book entitled *The Nature of Statistical Learning Theory* (Springer-Verlag, 1995) and the introduction of the celebrated support vector machine (SVM), research on kernel-based machine learning has flourished steadily for nearly two decades. The enormous amount of research findings on unsupervised and supervised learning models, both theory and applications, should already warrant a new textbook, even without considering the fact that this fundamental field will undoubtedly continue to grow for a good while.

The book first establishes algebraic and statistical foundations for kernel-based learning methods. It then systematically develops kernel-based learning models both for unsupervised and for supervised scenarios.

- The secret of success of a machine learning system lies in finding an effective representation for the objects of interest. In a basic representation, an object is represented as a feature vector in a finite-dimensional vector space. However, in numerous machine learning applications, two different types of modified representations are often employed: one involving dimension reduction and another involving dimension expansion.

  Dimension reduction. Dimension reduction is vital for visualization because of humans' inability to see objects geometrically in high-dimensional space. Likewise, dimension reduction may become imperative because of a machine's inability to process computationally demanding data represented by an extremely huge dimensionality. Subspace projection is a main approach to dimension reduction. This book will study principal component analysis (PCA) and discriminant component analysis (DCA), two such projection methods for unsupervised and supervised learning scenarios, respectively.

Dimension expansion. In other application scenarios, the dimensionality of the original feature space may be too small, which in turn limits the design freedom of any linear methods, rendering them ineffective for classifying datasets with complex data distributions. In this case, dimension expansion offers a simple and effective solution. One of the most systematic approaches to dimension expansion is the kernel methods, which are based on polynomial or Gaussian kernels. The higher the order of the kernel functions the more expanded the new feature space.

As shown later, the kernel methods, when applied to PCA or DCA, will lead to kernel PCA and kernel DCA, respectively. Likewise, the same methods may be used to derive various kernelized learning models both for unsupervised and for supervised scenarios.

- **Unsupervised learning models.** The book presents conventional unsupervised learning models for clustering analysis. They include K-means, expectation-maximization (EM), self-organizing-map (SOM), and neighbor-joining (NJ) methods. All these unsupervised learning models can be formulated as ℓ_2-based optimizers, thus they satisfy a critical learning subspace property (LSP). This in turn assures the existence of their kernelized counterparts, i.e. kernelized learning models. The latter models are formulated in terms of pairwise similarities between two objects, as opposed to the representative feature vectors for individual objects. Hence kernelized learning models are naturally applicable to non-vectorial data analysis, such as network segmentation.
- **Supervised learning models.** The book also presents conventional supervised learning models for classfication. They include least-squares error (LSE), Fisher discriminant analysis (FDA), ridge regression (RR) and linear SVM. All these supervised learning models can be formulated as ℓ_2-based optimizers, thus they satisfy the LSP condition, which in turn leads to their respective kernelized formulations, such as kernel RR (KRR) and kernel SVM. The combination of KRR and SVM further yields a hybrid classifier, named Ridge-SVM. The Ridge-SVM is endowed with a sufficient set of design parameters to embrace existing classifiers as its special cases, including KDA, KRR, and SVM. With properly adjusted parameters, again, all these kernelized supervised learning models are naturally applicable to nonvectorial data analysis, such as subcellular protein-sequence prediction.

In the book, the presentation of these topics and their extensions will be subdivided into the following parts:

(i) Part I: Machine learning and kernel vector spaces
(ii) Part II: Dimension-reduction: PCA/KPCA and feature selection
(iii) Part III: Unsupervised learning models for cluster analysis
(iv) Part VI: Kernel ridge regressors and variants
(v) Part V: Support vector machines and variants
(vi) Part VI: Kernel methods for green machine learning technologies
(vii) Part VII: Kernel methods for statistical estimation theory
(viii) Part VIII: Appendices.

The table of contents provides a more detailed description of the scope of the book.

From the perspective of new feature representation

The study of kernel-based machine learning involves a natural extension of the linear methods into their nonlinear counterparts. This book starts by devoting much of the discussion to establishing formally the linear learning models so as to make sure that students are given an opportunity to acquire a solid grasp of the underlying linear algebra and statistical principles of the learning models. The mathematical principle of kernel methods, instead of linear methods, hinges upon replacing the conventional pairwise similarity metric by a nonlinear kernel function. This ultimately leads to the nonlinear (and more flexible) decision boundaries for pattern classification. In summary, this basic mapping approach is conceptually simple. It involves (1) mapping the original representative vectors to the (dimension-expanded) intrinsic space, resulting in a training-data-independent feature representation; and (2) applying the same linear methods to the new and higher-dimensional feature vectors to yield a kernel-based learning model, which is defined over the intrinsic space.

From the perspective of the kernel trick

If the LSP holds, the above two-step mapping procedure can ultimately lead to a kernelized learning model, defined over the "empirical space" with a training-data-dependent feature representation. In the literature, the tedious two-step re-mapping process has often been replaced by a shortcut, nicknamed the "kernel trick." Most authors present the kernel trick as an elegant and simple notion. However, as evidenced by the following two aspects, a deeper understanding will prove essential to fully appreciating the limitation/power of the kernel trick.

- **The pre-requisite of applying the kernel trick.** First of all, note that not all linear learning models are amenable to the kernel trick. Let us briefly explain the precondition for applying the kernel trick. Conceptually, machine learning methods are built upon the principle of learning from examples. Algebraically, the range of the training vectors forms a learning subspace prescribing the subspace on which the solution is most likely to fall. This leads to a formal condition named the learning subspace property (LSP). It can be shown that the kernel trick is applicable to a linear learning model if and only if the LSP holds for the model. In other words, the LSP is the pre-requisite for the kernelizability of a linear learning model.
- **The interplay between two kernel-induced representations.** Given the kernelizability, we have at our disposal two learning models, defined over two different kernel-induced vector spaces. Now let us shift our attention to the interplay between two kernel-induced representations. Even though the two models are theoretically equivalent, they could incur very different implementation costs for learning and prediction. For cost-effective system implementation, one should choose the lower-cost representation, irrespective of whether it is intrinsic or empirical. For example, if the dimensionality of the empirical space is small and manageable, an empirical-space learning model will be more appealing. However, this will not be so if the number of

training vectors is extremely large, which is the case for the "big-data" learning scenario. In this case, one must give serious consideration to the intrinsic model, whose cost can be controlled by properly adjusting the order of the kernel function.

Presentation style and coverage of the book

For an introductory textbook, it would be wise to keep the mathematics to a minimum and choose materials that are easily accessible to beginning students and practitioners. After all, one of the overriding reasons for my undertaking of this project is because the original book by Vapnik is mathematically so deep that it is accessible only to the most able researchers.

Moreover, an editor keenly reminded me of the famous cliché that "for every equation in the book the readership would be halved." To be fair, my original intention was indeed to write a mathematically much simpler textbook. The book can hardly be considered a success by this measure – having included nearly a thousand equations, thirty or so algorithms, and almost as many theorems.

From another viewpoint, however, such heavy use of equations does serve some very useful purposes.

- This book includes nearly sixty numerical examples, many with step-by-step descriptions of an algorithmic procedure. Concrete examples with numerical equations may go a long way towards clarifying the mathematical algorithm or theorem. They provide a tangible, and much less abstract, illustration of the actual procedure.
- This book contains equations specifying the bounds of computational complexities or estimates of prediction performance associated with a learning model, each of which could serve as a preliminary and quantitative guideline on the effectiveness of the learning model for specific applications.
- The book aims at demonstrating how machine learning models can be integrated into a recognition application system. Some theorems and equations in the book are devoted to establishing connections between equivalent learning models, paving a way to avoid redundant experiments on equivalent (and thus predictable) models. In short, the mathematical equivalence both improves the understanding of the models and prevents repetitive coding efforts.
- Compared with natural language or computer language (e.g. pseudocodes), the mathematics and equations provide a more concise descriptive language. With somewhat casual mathematical language, the semi-formal presentation style of this book should help beginning readers to more easily appreciate the power of the linear algebra and statistical theory behind the machine learning tools.

Comprehensiveness versus cohesiveness

Since machine learning covers a vast range of subjects, the selection of materials for this book inevitably involves a tradeoff between comprehensiveness and cohesiveness. Admittedly, the coverage of the book is far from being comprehensive. The constraint on space was certainly an important factor. On the other hand, there is already a large volume of publications on SVM and its variants. In order to save space, it was necessary to leave out many SVM-related subjects, knowing that several excellent presentations of SVM are already available in textbook form.

What sets the book apart from others is unlikely to be its scope of coverage; rather, it may very well be the cohesive presentation and novel results.

- **Cohesive presentation.** The book aims at offering a cohesive, organized, and yet balanced presentation with natural flow between sections. This streamlined approach facilitates the presentation of key ideas in a single flow, without digression into the analytical details. Moreover, the streamlined approach also reflects a personal (and subjective) viewpoint on how to relate the loosely connected subjects.
- **Novel results.** Some significant novel results have been introduced here for the first time in textbook form. For example, under the supervised scenario, DCA for optimal subspace projection will outperform PCA, which is meant for use in unsupervised scenarios. A hybrid learning model of KRR and SVM, named Ridge-SVM, covers many existing classifiers as special cases, including KDA, KRR, and SVM. With properly adjusted parameters, it has been shown to deliver improved generalization and prediction capability. The book also establishes the theoretical foundation linking kernel methods and the rich theory in estimation, prediction, and system identification. Curiously, the presentation of these novel ideas seemed to fall naturally into appropriate places in their respective chapters.

Finally, due to its emphasis being placed on a cohesive and streamlined presentation of key ideas, the book necessarily had to forgo some otherwise important research results. I would like to take this opportunity to express my most sincere apologies and profound regret to researchers whose contributions have inadvertently been omitted here.

Readership of the book

The book was designed for senior and graduate students with a diversity of educational experiences in computer science, electrical engineering, financial engineering, applied statistics, etc. The main focus of the book aims at taking a beginning student, with some prior exposure to linear algebra, statistical theory, and convex optimization, through an integrated understanding of the underlying principles and potential applications of kernel-based learning models. In addition, the book should provide enough material for it to be used either as a textbook for classroom instruction or as a reference book for self-study.

- **As a textbook for machine learning course.** The book may be adopted for one-semester senior or graduate courses in machine learning in, say, electrical engineering and computer science departments. For example, by carefully picking some fundamental materials from Chapters 1 through 13, it should be possible to find enough material to be organized into a one-semester course that covers feature representations, and unsupervised and supervised learning models, with balanced yet rigorous treatments in statistics and linear algebra.

Just like in other textbooks, exercises are included at the end of each chapter. They should be useful for self-study and for probing into some of the more intricate aspects of the subjects treated in the text.

- **As a recommended or supplementary reference for courses on artificial intelligence.** The scope of the materials covered here is sufficiently broad to allow it to be re-structured for many other educational purposes. For example, the book may be adopted as a recommended reference for artificial intelligence and machine learning. It may also be adopted as a textbook/reference for a two-semester course. In this case, the first semester can be devoted to fundamental concepts, with the second semester covering advanced research areas such as big-data learning and kernel-based statistical estimation. For the latter area, Chapters 14 and 15 present statistical estimation techniques with errors-in-variables methods, Gauss–Markov theorems, and kernel methods for time-series analysis.
- **As a reference book for research and development.** The book is also intended for professional engineers, scientists, and system integrators who want to learn systematic ways of implementing machine learning systems. Throughout the book, application examples are provided to motivate the learning model developed. The book provides practitioners with basic mathematical knowledge so that they know how to apply off-the-shelf machine learning codes to solve new problems. In addition, efforts have been made to make the book relatively self-contained. For example, some basic matrix algebra and statistical theory are included in the book, making the book more accessible to newcomers from other fields and to those who have become rusty with some aspects of their undergraduate curriculum.

Acknowledgements

I found this writing project to be expectedly difficult at the beginning, but surprisingly enjoyable towards the end. It was truly rewarding seeing so many old and new results fall so nicely into place together. I also came to the realization that I had been so very fortunate to be surrounded by many fine people, professors, colleagues, students, and friends. The emotional parting with a seven-year-long project is somewhat offset by the pleasure of being able to finally acknowledge this unique group of people who made it possible.

I am pleased to acknowledge the generous support of a gift grant from Mitsubishi (MERL), a research grant from Motorola, multiple research grants from the Hong Kong Research Grants Council, and the DARPA Research Program on active authentication. The project was also indirectly supported by various fellowships, received by some of my collaborators, from Princeton University, the Canadian Government, and Microsoft Inc.

I was fortunate to benefit from the outstanding professional support of many fine people at Cambridge University Press (CUP), including Phil Meyler, Sarah Marsh, Elizabeth Horne, Kirsten Bot, Jessica Murphy, Dr. Steven Holt, and numerous others. I wish to thank the anonymous CUP reviewer who kindly suggested the current title of the book.

During the period of the book project, I was a Distinguished Visiting Professor at the EEE Department of the University of Hong Kong for several summers. I am grateful for the kind hospitality, warm friendship, and stimulating exchange with C. Q. Chang, Fei Mai, Y. S. Hung, and many others.

This book was an outgrowth of many years of teaching and research on neural networks, biometric authentication, and machine learning. I am grateful to the Department of Electrical Engineering of Princeton University and my fellow colleagues for having created such a scholarly environment for teaching and research. In particular, I would like to acknowledge Sarah McGovern, Stacey Weber, and Lori Baily for their cheerful spirit and generous assistance.

I am much indebted to my Ph.D. students, former and current, for their participation in building my understanding of machine learning. They include Xinying Zhang, Yunnan Wu, C. L. Myers, Ilias Tagkopoulos, Yuhui Luo, Peiyuan Wu, and Yinan Yu (Princeton University), as well as Jian Guo, Shibiao Wan, and F. Tobar (outside Princeton University). Their research studies have provided an important foundation for this book. Moreover, they have helped develop this book in various ways.

I would like to acknowledge the invaluable contributions of all of the students in my class during the past six years, undergraduate and graduate, for their invaluable contribution to examples and exercises. In particular, I would like to mention Tiffany Tong, Chun-Yi Lee, Chia-Chun Lin, Dan Li, K. H. Lee, Si Chen, Yang Yang, Clement Canonne, Pei-yuan Wu, Zhang Zhuo, Xu Chen, Pingmei Xu, Shang Shang, Rasmus Rothe, Vincent Pham, and Jintao Zhang.

I express my sincere gratitude to my visiting professors for stimulating discussions and for their proofreading of numerous versions of the previous drafts when they visited Princeton University. They are Young-Shik Moon, Shang-Hung Lai, Shaikh Fattah, Jie Lin, Wei-Kuang Lai, Xiao-Dong Gu, Yu Liu, and K. Diamantaras. I also benefited greatly from the enlightening exchanges with many external collaborators, in particular, Professors J. Morris Chang, Y. K. Chen, Y. B. Kan, T. S. Lin, Mahesan Niranjan, D. Mandic, T. McKelvery, Jin-Shiuh Taur, Yue Wang, and Juan Zhou.

There is little doubt that I must have missed some important names of people whom I would like to thank, and to whom I wish to offer my most sincere apologies and profound regret in that regard.

It is always fun and brings back fond memories recalling my Stanford years, so I must express my special appreciation of Professor Thomas Kailath, my advisor and life-time mentor, for his constant inspiration and friendship. I am proud to be closely associated with a group of outstanding scholars including Professors Patrick DeWilde, Lenart Ljung, Bernard Levy, George Verghese, and Erik Verriest, among many others. Moreover, my first exposure to machine learning was a course taught by none other than Professor R. O. Duda, which was based on his now classical book *Pattern Classification and Scene Analysis* (John Wiley, 1973).

Their mention so far does not fully acknowledge the measure of the contributions by Professor Man-Wai Mak, Mr. Peiyuan Wu, and Miss Yinan Yu. For their invaluable and indispensable roles, they could conceivably have been named as co-authors of the book.

Finally, a book project of such scale would not have been possible without strong support from my parents, my wife, and all our children. In recent years, the center of attention of my (much extended) family seems to have been focused upon our four grandchildren. It is only fitting that the book is dedicated to them.

S. Y. KUNG
Princeton