Cambridge University Press
978-0-521-79192-2 - Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels
William W. Hsieh
Frontmatter
More information

MACHINE LEARNING METHODS IN THE ENVIRONMENTAL SCIENCES

Neural Networks and Kernels

William W. Hsieh

Machine learning methods, having originated from computational intelligence (i.e. artificial intelligence), are now ubiquitous in the environmental sciences. This is the first single-authored textbook to give a unified treatment of machine learning methods and their applications in the environmental sciences.

Machine learning methods began to infiltrate the environmental sciences in the 1990s. Today, thanks to their powerful nonlinear modelling capability, they are no longer an exotic fringe species, as they are heavily used in satellite data processing, in general circulation models (GCM), in weather and climate prediction, air quality forecasting, analysis and modelling of environmental data, oceanographic and hydrological forecasting, ecological modelling, and in the monitoring of snow, ice and forests, etc. End-of-chapter review questions are included, allowing readers to develop their problem-solving skills and monitor their understanding of the material presented. An appendix lists websites available for downloading computer code and data sources. A resources website is available containing datasets for exercises, and additional material to keep the book completely up-to-date.

This book presents machine learning methods and their applications in the environmental sciences (including satellite remote sensing, atmospheric science, climate science, oceanography, hydrology and ecology), written at a level suitable for beginning graduate students and advanced undergraduates. It is also valuable for researchers and practitioners in environmental sciences interested in applying these new methods to their own work.

William W. Hsieh is a Professor in the Department of Earth and Ocean Sciences and in the Department of Physics and Astronomy, as well as Chair of the Atmospheric Science Programme, at the University of British Columbia. He is internationally known for his pioneering work in developing and applying machine learning methods in the environmental sciences. He has published over 80 peer-reviewed journal publications covering areas of climate variability, machine learning, oceanography, atmospheric science and hydrology.

MACHINE LEARNING METHODS IN THE ENVIRONMENTAL SCIENCES

Neural Networks and Kernels

WILLIAM W. HSIEH

University of British Columbia
Vancouver, BC, Canada

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521791922

First published 2009

Printed in the United Kingdom at the University Press, Cambridge

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

ISBN 978-0-521-79192-2 hardback

Contents

Preface

Machine learning is a major sub-field in computational intelligence (also called artificial intelligence). Its main objective is to use computational methods to extract information from data. Machine learning has a wide spectrum of applications including handwriting and speech recognition, object recognition in computer vision, robotics and computer games, natural language processing, brain–machine interfaces, medical diagnosis, DNA classification, search engines, spam and fraud detection, and stock market analysis. Neural network methods, generally regarded as forming the first wave of breakthrough in machine learning, became popular in the late 1980s, while kernel methods arrived in a second wave in the second half of the 1990s.

In the 1990s, machine learning methods began to infiltrate the environmental sciences. Today, they are no longer an exotic fringe species, since their presence is ubiquitous in the environmental sciences, as illustrated by the lengthy References section of this book. They are heavily used in satellite data processing, in general circulation models (GCM) for emulating physics, in post-processing of GCM model outputs, in weather and climate prediction, air quality forecasting, analysis and modelling of environmental data, oceanographic and hydrological forecasting, ecological modelling, and in monitoring of snow, ice and forests, etc.

This book presents machine learning methods (mainly neural network and kernel methods) and their applications in the environmental sciences, written at a level suitable for beginning graduate students and advanced undergraduates. It is also aimed at researchers and practitioners in environmental sciences, who having been intrigued by exotic terms like neural networks, support vector machines, self-organizing maps, evolutionary computation, etc., are motivated to learn more about these new methods and to use them in their own work. The reader is assumed to know multivariate calculus, linear algebra and basic probability.

Chapters 1–3, intended mainly as background material for students, cover the standard statistical methods used in environmental sciences. The machine learning methods of later chapters provide powerful nonlinear generalizations for many of these standard linear statistical methods. The reader already familiar with the background material of Chapters 1–3 can start directly with Chapter 4, which introduces neural network methods. While Chapter 5 is a relatively technical chapter on nonlinear optimization algorithms, Chapter 6 on learning and generalization is essential to the proper use of machine learning methods – in particular, Section 6.10 explains why a nonlinear machine learning method often outperforms a linear method in weather applications but fails to do so in climate applications. Kernel methods are introduced in Chapter 7. Chapter 8 covers nonlinear classification, Chapter 9, nonlinear regression, Chapter 10, nonlinear principal component analysis, and Chapter 11, nonlinear canonical correlation analysis. Chapter 12 broadly surveys applications of machine learning methods in the environmental sciences (remote sensing, atmospheric science, oceanography, hydrology, ecology, etc.). For exercises, the student could test the methods on data from their own area or from some of the websites listed in Appendix A. Codes for many machine learning methods are also available from sites listed in Appendix A. The book website `www.cambridge.org/hsieh` also provides datasets for some of the exercises given at the ends of the chapters.

On a personal note, writing this book has been both exhilarating and gruelling. When I first became intrigued by neural networks through discussions with Dr Benyang Tang in 1992, I recognized that the new machine learning methods would have a major impact on the environmental sciences. However, I also realized that I had a steep learning curve ahead of me, as my background training was in physics, mathematics and environmental sciences, but not in statistics nor computer science. By the late 1990s I became convinced that the best way for me to learn more about machine learning was to write a book. What I thought would take a couple of years turned into a marathon of over eight years, as I desperately tried to keep pace with a rapidly expanding research field. I managed to limp past the finish line in pain, as repetitive strain injury from overusage of keyboard and mouse struck in the final months of intensive writing!

I have been fortunate in having supervised numerous talented graduate students, post-doctoral fellows and research associates, many of whom taught me far more than I taught them. I received helpful editorial assistance from the staff at the Cambridge University Press and from Max Ng. I am grateful for the support from my two university departments (Earth and Ocean Sciences, and Physics and Astronomy), the Peter Wall Institute of Advanced Studies, the Natural Sciences and Engineering Research Council of Canada and the Canadian Foundation for Climate and Atmospheric Sciences.

Without the loving support from my family (my wife Jean and my daughters, Teresa and Serena), and the strong educational roots planted decades ago by my parents and my teachers, I could not have written this book.

Notation used in this book

In general, vectors are denoted by lower case bold letters (e.g. $\mathbf{v}$), matrices by upper case bold letters (e.g. $\mathbf{A}$) and scalar variables by italics (e.g. x or J). A column vector is denoted by $\mathbf{v}$, while its transpose $\mathbf{v}^{\mathrm{T}}$ is a row vector, i.e. $\mathbf{v}^{\mathrm{T}} = (v_1, v_2, \ldots, v_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)^{\mathrm{T}}$, and the inner or dot product of two vectors $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^{\mathrm{T}}\mathbf{b} = \mathbf{b}^{\mathrm{T}}\mathbf{a}$. The elements of a matrix $\mathbf{A}$ are written as A_{ij} or $(\mathbf{A})_{ij}$. The probability for discrete variables is denoted by upper case P, whereas the probability density for continuous variables is denoted by lower case p. The expectation is denoted by $\mathrm{E}[\ldots]$ or $\langle \ldots \rangle$. The natural logarithm is denoted by ln or log.

Abbreviations

AO = Arctic Oscillation
BNN = Bayesian neural network
CART = classification and regression tree
CCA = canonical correlation analysis
CDN = conditional density network
EC = evolutionary computation
EEOF = extended empirical orthogonal function
ENSO = El Niño-Southern Oscillation
EOF = empirical orthogonal function
GA = genetic algorithm
GCM = general circulation model (or global climate model)
GP = Gaussian process model
IC = information criterion
LP = linear projection
MAE = mean absolute error
MCA = maximum covariance analysis
MJO = Madden–Julian Oscillation
MLP = multi-layer perceptron neural network
MLR = multiple linear regression
MOS = model output statistics
MSE = mean squared error
MSSA = multichannel singular spectrum analysis
NAO = North Atlantic Oscillation
NLCCA = nonlinear canonical correlation analysis
NLCPCA = nonlinear complex PCA
NN = neural network
NLPC = nonlinear principal component
NLPCA = nonlinear principal component analysis

NLSSA = nonlinear singular spectrum analysis
PC = principal component
PCA = principal component analysis
PNA = Pacific-North American teleconnection
POP = principal oscillation pattern
QBO = Quasi-Biennial Oscillation
RBF = radial basis function
RMSE = root mean squared error
SLP = sea level pressure
SOM = self-organizing map
SSA = singular spectrum analysis
SST = sea surface temperature (sum of squares in Chapter 1)
SVD = singular value decomposition
SVM = support vector machine
SVR = support vector regression