

# Machine Learning Methods in Environmental Sciences

*Neural Networks and Kernels*

**William W. Hsieh**

University of British Columbia  
Vancouver, B.C., Canada

Cambridge University Press

August 31, 2008

# Preface

Machine learning is a major subfield 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 bibliography of this book. They are heavily used in satellite data processing, in general circulation models (GCM) for emulating physics, in the 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 the 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 his/her 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 the 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 web sites listed in Appendix A. Codes for many machine learning methods are also available from sites listed in Appendix A.

On a personal note, writing this book has been both exhilarating and grueling. 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 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 pass 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  $\mathbf{v}^T$  is a row vector, i.e.  $\mathbf{v}^T = (v_1, v_2, \dots, v_n)$  and  $\mathbf{v} = (v_1, v_2, \dots, v_n)^T$ , and the inner or dot product of two vectors  $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b} = \mathbf{b}^T \mathbf{a}$ . The elements of a matrix  $\mathbf{A}$  is written as  $A_{ij}$  or  $(\mathbf{A})_{ij}$ . The probability for discrete variables is denoted by the upper case  $P$ , whereas the probability density for continuous variables is denoted by the lower case  $p$ . The expectation is denoted by  $E[\dots]$  or  $\langle \dots \rangle$ . The natural logarithm is denoted by  $\ln$  or  $\log$ .

### Acronyms:

AO = Arctic Oscillation

BNN = Bayesian neural network

CART = classification and regression tree  
CCA = canonical correlation analysis  
CDN = conditional density network  
EC = evolutionary computation  
ENSO = El Niño-Southern Oscillation  
EOF = empirical orthogonal function  
EEOF = extended empirical orthogonal function  
GCM = general circulation model (or global climate model)  
GA = genetic algorithm  
GP = Gaussian process model  
IC = information criterion  
LP = linear projection  
MAE = mean absolute error  
MJO = Madden-Julian Oscillation  
MLP = multi-layer perceptron neural network  
MLR = multiple linear regression  
MOS = model output statistics  
MSE = mean square error  
MSSA = multichannel singular spectrum analysis  
NAO = North Atlantic Oscillation  
NN = neural network  
NLCCA = nonlinear canonical correlation analysis  
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 square error  
SSA = singular spectrum analysis  
SLP = sea level pressure  
SOM = self-organizing map  
SST = sea surface temperature (sum of squares in Chapter 1)  
SVD = singular value decomposition  
SVM = support vector machine  
SVR = support vector regression

# Contents

<b>1</b>	<b>Basic notions in classical data analysis</b>	<b>1</b>
1.1	Expectation and mean . . . . .	1
1.2	Variance and covariance . . . . .	2
1.3	Correlation . . . . .	3
1.3.1	Rank correlation . . . . .	4
1.3.2	Autocorrelation . . . . .	5
1.3.3	Correlation matrix . . . . .	6
1.4	Regression . . . . .	7
1.4.1	Linear regression . . . . .	7
1.4.2	Relating regression to correlation . . . . .	8
1.4.3	Partitioning the variance . . . . .	9
1.4.4	Multiple linear regression . . . . .	10
1.4.5	Perfect Prog and MOS . . . . .	11
1.5	Bayes theorem . . . . .	12
1.6	Discriminant functions and classification . . . . .	13
1.7	Clustering . . . . .	16
<b>2</b>	<b>Linear Multivariate Statistical Analysis</b>	<b>19</b>
2.1	Principal component analysis (PCA) . . . . .	19
2.1.1	Geometric approach to PCA . . . . .	19
2.1.2	Eigenvector approach to PCA . . . . .	20
2.1.3	Real and complex data . . . . .	23
2.1.4	Orthogonality relations . . . . .	24
2.1.5	PCA of the tropical Pacific climate variability . . . . .	25
2.1.6	Scaling the PCs and eigenvectors . . . . .	30
2.1.7	Degeneracy of eigenvalues . . . . .	33
2.1.8	A smaller covariance matrix . . . . .	34
2.1.9	Temporal and spatial mean removal . . . . .	35
2.1.10	Singular value decomposition . . . . .	35
2.1.11	Missing data . . . . .	37
2.1.12	Significance tests . . . . .	38
2.2	Rotated PCA . . . . .	39
2.3	PCA for vectors . . . . .	44
2.4	Canonical correlation analysis (CCA) . . . . .	49

2.4.1	CCA theory . . . . .	49
2.4.2	Pre-filter with PCA . . . . .	52
2.4.3	Singular value decomposition and maximum covariance analysis . . . . .	56
<b>3</b>	<b>Basic time series analysis</b>	<b>58</b>
3.1	Spectrum . . . . .	58
3.1.1	Autospectrum . . . . .	59
3.1.2	Cross-spectrum . . . . .	63
3.2	Windows . . . . .	65
3.3	Filters . . . . .	66
3.4	Singular spectrum analysis . . . . .	69
3.5	Multichannel singular spectrum analysis . . . . .	73
3.6	Principal oscillation patterns . . . . .	76
3.7	Spectral principal component analysis . . . . .	85
<b>4</b>	<b>Feed-forward neural network models</b>	<b>88</b>
4.1	McCulloch and Pitts model . . . . .	89
4.2	Perceptrons . . . . .	89
4.3	Multi-layer perceptrons (MLP) . . . . .	94
4.4	Back-propagation . . . . .	98
4.5	Hidden neurons . . . . .	105
4.6	Radial basis functions (RBF) . . . . .	107
4.7	Conditional probability distributions . . . . .	110
4.7.1	Mixture models . . . . .	112
<b>5</b>	<b>Nonlinear optimization</b>	<b>115</b>
5.1	Gradient descent method . . . . .	117
5.2	Conjugate gradient method . . . . .	118
5.3	Quasi-Newton methods . . . . .	121
5.4	Nonlinear least squares methods . . . . .	124
5.5	Evolutionary computation & genetic algorithms . . . . .	126
<b>6</b>	<b>Learning and generalization</b>	<b>130</b>
6.1	Mean square error and maximum likelihood . . . . .	131
6.2	Objective functions and robustness . . . . .	132
6.3	Variance and bias errors . . . . .	135
6.4	Reserving data for validation . . . . .	137
6.5	Regularization . . . . .	138
6.6	Cross-validation . . . . .	139
6.7	Bayesian neural networks (BNN) . . . . .	141
6.7.1	Estimating the hyperparameters . . . . .	143
6.7.2	Estimate of predictive uncertainty . . . . .	147
6.8	Ensemble of models . . . . .	149
6.9	Approaches to predictive uncertainty . . . . .	154
6.10	Linearization from time-averaging . . . . .	155

<b>7</b>	<b>Kernel methods</b>	<b>160</b>
7.1	From neural networks to kernel methods . . . . .	160
7.2	Primal and dual solutions for linear regression . . . . .	162
7.3	Kernels . . . . .	164
7.4	Kernel ridge regression . . . . .	167
7.5	Advantages and disadvantages . . . . .	168
7.6	The pre-image problem . . . . .	170
<b>8</b>	<b>Nonlinear classification</b>	<b>174</b>
8.1	Multi-layer perceptron classifier . . . . .	175
8.1.1	Cross entropy error function . . . . .	179
8.2	Multi-class classification . . . . .	179
8.3	Bayesian neural network (BNN) classifier . . . . .	181
8.4	Support vector machine (SVM) classifier . . . . .	182
8.4.1	Linearly separable case . . . . .	182
8.4.2	Linearly non-separable case . . . . .	186
8.4.3	Nonlinear classification by SVM . . . . .	188
8.4.4	Multi-class classification by SVM . . . . .	190
8.5	Forecast verification . . . . .	191
8.5.1	Skill scores . . . . .	193
8.5.2	Multiple classes . . . . .	196
8.5.3	Probabilistic forecasts . . . . .	196
8.6	Unsupervised competitive learning . . . . .	197
<b>9</b>	<b>Nonlinear regression</b>	<b>201</b>
9.1	Support vector regression (SVR) . . . . .	201
9.2	Classification and regression trees (CART) . . . . .	207
9.3	Gaussian processes (GP) . . . . .	211
9.3.1	Learning the hyperparameters . . . . .	214
9.3.2	Other common kernels . . . . .	216
9.4	Probabilistic forecast scores . . . . .	216
<b>10</b>	<b>Nonlinear principal component analysis</b>	<b>218</b>
10.1	Auto-associative NN for nonlinear PCA . . . . .	220
10.1.1	Open curves . . . . .	220
10.1.2	Application . . . . .	224
10.1.3	Overfitting . . . . .	228
10.1.4	Closed curves . . . . .	233
10.2	Principal curves . . . . .	238
10.3	Self-organizing maps (SOM) . . . . .	239
10.4	Kernel principal component analysis . . . . .	244
10.5	Nonlinear complex PCA . . . . .	247
10.6	Nonlinear singular spectrum analysis . . . . .	251

<b>11 Nonlin. canonical correlation analysis</b>	<b>260</b>
11.1 MLP-based NLCCA model . . . . .	260
11.1.1 Tropical Pacific climate variability . . . . .	268
11.1.2 Atmospheric teleconnection . . . . .	269
11.2 Robust NLCCA . . . . .	272
11.2.1 Biweight midcorrelation . . . . .	274
11.2.2 Inverse mapping . . . . .	277
11.2.3 Prediction . . . . .	279
<b>12 Applications in environmental sciences</b>	<b>284</b>
12.1 Remote sensing . . . . .	285
12.1.1 Visible light sensing . . . . .	286
12.1.2 Infrared sensing . . . . .	290
12.1.3 Passive microwave sensing . . . . .	293
12.1.4 Active microwave sensing . . . . .	294
12.2 Oceanography . . . . .	297
12.2.1 Sea level . . . . .	297
12.2.2 Equation of state of sea water . . . . .	297
12.2.3 Wind wave modelling . . . . .	299
12.2.4 Ocean temperature and heat content . . . . .	300
12.3 Atmospheric science . . . . .	302
12.3.1 Hybrid coupled modelling of the tropical Pacific . . . . .	302
12.3.2 Climate variability and climate change . . . . .	303
12.3.3 Radiation in atmospheric models . . . . .	307
12.3.4 Post-processing and downscaling of numerical model out- put . . . . .	309
12.3.5 Severe weather forecasting . . . . .	316
12.3.6 Air quality . . . . .	317
12.4 Hydrology . . . . .	321
12.5 Ecology . . . . .	324
<b>A Sources for data and codes</b>	<b>327</b>
<b>B Lagrange multipliers</b>	<b>329</b>
<b>Bibliography</b>	<b>332</b>
<b>Index</b>	<b>365</b>