Information Science and Statistics

Information Science and Statistics

Akaike and Kitagawa: The Practice of Time Series Analysis.
Bishop: Pattern Recognition and Machine Learning.
Cowell, Dawid, Lauritzen, and Spiegelhalter: Probabilistic Networks and Expert Systems.
Doucet, de Freitas, and Gordon: Sequential Monte Carlo Methods in Practice.
Fine: Feedforward Neural Network Methodology.
Hawkins and Olwell: Cumulative Sum Charts and Charting for Quality Improvement.
Jensen and Nielsen: Bayesian Networks and Decision Graphs, Second Edition.
Lee and Verleysen: Nonlinear Dimensionality Reduction.
Marchette: Computer Intrusion Detection and Network Monitoring: A Statistical Viewpoint.
Rissanen: Information and Complexity in Statistical Modeling.
Rubinstein and Kroese: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation, and Machine Learning.
Studený: Probabilistic Conditional Independence Structures.
Vapnik: The Nature of Statistical Learning Theory, Second Edition.
Wallace: Statistical and Inductive Inference by Minimum Massage Length.

John A. Lee Michel Verleysen

Nonlinear Dimensionality Reduction

Springer

John Lee
Molecular Imaging and Experimental Radiotherapy
Université catholique de Louvain
Avenue Hippocrate 54/69
B-1200 Bruxelles
Belgium
john.lee@uclouvain.be

Michel Verleysen
Machine Learning Group – DICE
Université catholique de Louvain
Place du Levant 3
B-1348 Louvain-la-Neuve
Belgium
michel.verleysen@uclouvain.be

Series Editors:

Michael Jordan
Division of Computer Science and Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
USA

Jon Kleinberg
Department of Computer Science
Cornell University
Ithaca, NY 14853
USA

Bernhard Schölkopf
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen
Germany

Library of Congress Control Number: 2006939149

ISBN-13: 978-0-387-39350-6 e-ISBN-13: 978-0-387-39351-3

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To our families

Preface

Methods of dimensionality reduction are innovative and important tools in the fields of data analysis, data mining, and machine learning. They provide a way to understand and visualize the structure of complex data sets. Traditional methods like principal component analysis and classical metric multidimensional scaling suffer from being based on linear models. Until recently, very few methods were able to reduce the data dimensionality in a nonlinear way. However, since the late 1990s, many new methods have been developed and nonlinear dimensionality reduction, also called manifold learning, has become a hot topic. New advances that account for this rapid growth are, for example, the use of graphs to represent the manifold topology, and the use of new metrics like the geodesic distance. In addition, new optimization schemes, based on kernel techniques and spectral decomposition, have led to spectral embedding, which encompasses many of the recently developed methods.

This book describes existing and advanced methods to reduce the dimensionality of numerical databases. For each method, the description starts from intuitive ideas, develops the necessary mathematical details, and ends by outlining the algorithmic implementation. Methods are compared with each other with the help of different illustrative examples.

The purpose of the book is to summarize clear facts and ideas about well-known methods as well as recent developments in the topic of nonlinear dimensionality reduction. With this goal in mind, methods are all described from a unifying point of view, in order to highlight their respective strengths and shortcomings.

The book is primarily intended for statisticians, computer scientists, and data analysts. It is also accessible to other practitioners having a basic background in statistics and/or computational learning, such as psychologists (in psychometry) and economists.

Louvain-la-Neuve, Belgium *John A. Lee*
October 2006 *Michel Verleysen*

Contents

Notations

$\mathbb{N}$	The set of positive natural numbers: $\{0, 1, 2, 3, \ldots\}$
$\mathbb{R}$	The set of real numbers
y, x	Known or unknown random variables taking their values in $\mathbb{R}$
$\mathbf{A}$	A matrix
$a_{i,j}$	An entry of the matrix $\mathbf{A}$ (located at the crossing of the ith row and the jth column)
N	Number of points in the data set
M	Number of prototypes in the codebook $\mathbf{C}$
D	Dimensionality of the data space (which is usually $\mathbb{R}^D$)
P	Dimensionality of the latent space (which is usually $\mathbb{R}^P$) (or its estimation as the intrinsic dimension of the data)
$\mathbf{I}_D$	D-dimensional identity matrix
$\mathbf{I}_{P\times D}$	Rectangular matrix containing the first P rows of $\mathbf{I}_D$
$\mathbf{1}_N$	N-dimensional column vector containing ones everywhere
$\mathbf{y}$	Random vector in the known data space: $\mathbf{y} = [y_1, \ldots, y_d, \ldots, y_D]^T$
$\mathbf{x}$	Random vector in the unknown latent space: $\mathbf{x} = [x_1, \ldots, x_p, \ldots, x_P]^T$
$\mathbf{y}(i)$	The ith vector of the data set
$\mathbf{x}(i)$	(Unknown) latent vector that generated $\mathbf{y}(i)$
$\hat{\mathbf{x}}(i)$	The estimate of $\mathbf{x}(i)$
$\mathcal{Y}$	The data set $\mathcal{Y} = \{\ldots, \mathbf{y}(i), \ldots\}_{1\leq i\leq N}$
$\mathcal{X}$	The (unknown) set of latent vectors that generated $\mathcal{Y}$
$\hat{\mathcal{X}}$	Estimation of $\mathcal{X}$
$\mathbf{Y}$	The data set in matrix notation: $\mathcal{Y} = [\ldots, \mathbf{y}(i), \ldots]_{1\leq i\leq N}$
$\mathbf{X}$	The (unknown) ordered set of latent vectors that generated $\mathbf{Y}$
$\hat{\mathbf{X}}$	Estimation of $\mathbf{X}$

$\mathcal{M}$	A manifold (noted as a set)
$\mathbf{m}$	The functional notation of $\mathcal{M}$: $\mathbf{y} = \mathbf{m}(\mathbf{x})$
$E_x\{x\}$	The expectation of the random variable x
$\mu_x(x)$	The mean value of the random variable x (computed with its known values $x(i)$, $i = 1, \ldots, N$)
μ_i	The ith-order centered moment
μ'_i	The ith-order raw moment
$\mathbf{C}_{\mathbf{xy}}$	The covariance matrix between the random vectors $\mathbf{x}$ and $\mathbf{y}$
$\hat{\mathbf{C}}_{\mathbf{xy}}$	The estimate of the covariance matrix
$f(\mathbf{x})$, $\mathbf{f}(\mathbf{x})$	Uni- or multivariate function of the random vector $\mathbf{x}$
$\frac{\partial f(\mathbf{x})}{\partial x_p}$	Partial derivative of f with respect to x_p
$\nabla_{\mathbf{x}} f(\mathbf{x})$	Gradient vector of f with respect to $\mathbf{x}$
$\mathbf{H}_{\mathbf{x}} f(\mathbf{x})$	Hessian matrix of f with respect to $\mathbf{x}$
$\mathbf{J}_{\mathbf{x}} \mathbf{f}(\mathbf{x})$	Jacobian matrix of $\mathbf{f}$ with respect to $\mathbf{x}$
$\langle \mathbf{y}(i) \cdot \mathbf{y}(j) \rangle$	Scalar product between the two vectors $\mathbf{y}(i)$ and $\mathbf{y}(j)$
$d(\mathbf{y}(i), \mathbf{y}(j))$	Distance function between the two vectors $\mathbf{y}(i)$ and $\mathbf{y}(j)$ (often a spatial distance, like the Euclidean one) shortened as $d_{\mathbf{y}}(i, j)$ or $d_{\mathbf{y}}$ when the context is clear
$\delta(\mathbf{y}(i), \mathbf{y}(j))$	Geodesic or graph distance between $\mathbf{y}(i)$ and $\mathbf{y}(j)$
$\mathcal{C}$, $\mathcal{G}$	Codebook (noted as a set) in the data and latent spaces
$\mathbf{C}$, $\mathbf{G}$	Codebook (noted as a matrix) in the data and latent spaces
$\mathbf{c}(r)$, $\mathbf{g}(r)$	Coordinates of the rth prototypes in the codebook (respectively, in the data and latent spaces)

Acronyms

DR	Dimensionality reduction
LDR	Linear dimensionality reduction
NLDR	Nonlinear dimensionality reduction

ANN	Artificial neural networks
EVD	Eigenvalue decomposition
SVD	Singular value decomposition
SVM	Support vector machines
VQ	Vector quantization

CCA	Curvilinear component analysis	*NLDR method*
CDA	Curvilinear distance analysis	*NLDR method*
EM	Expectation-maximization	*optimization technique*
GTM	Generative topographic mapping	*NLDR method*
HLLE	Hessian LLE (see LLE)	*NLDR method*
KPCA	Kernel PCA (see PCA)	*NLDR method*
LE	Laplacian eigenmaps	*NLDR method*
LLE	Locally linear embedding	*NLDR method*
MDS	Multidimensional scaling	*LDR/NLDR method*
MLP	Multilayer perceptron	*ANN for function approx.*
MVU	Maximum variance unfolding (see SDE)	*NLDR method*
NLM	(Sammon's) nonlinear mapping	*NLDR method*
PCA	Principal component analysis	*LDR method*
RBFN	Radial basis function network	*ANN for function approx.*
SDE	Semidefinite embedding	*NLDR method*
SDP	Semidefinite programming	*optimization technique*
SNE	Stochastic neighbor embedding	*NLDR method*
SOM	(Kohonen's) self-organizing map	*NLDR method*
TRN	Topology-representing network	*ANN*