

Introduction  
to  
Machine  
Learning

Third  
Edition

## **Adaptive Computation and Machine Learning**

Thomas Dietterich, Editor

Christopher Bishop, David Heckerman, Michael Jordan, and Michael  
Kearns, Associate Editors

A complete list of books published in The Adaptive Computation and  
Machine Learning series appears at the back of this book.

# Introduction to Machine Learning

Third  
Edition

Ethem Alpaydın

The MIT Press  
Cambridge, Massachusetts  
London, England

© 2014 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email  
special\_sales@mitpress.mit.edu.

Typeset in 10/13 Lucida Bright by the author using  $\text{\LaTeX}$  2 $\epsilon$ .  
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Information

Alpaydin, Ethem.

Introduction to machine learning / Ethem Alpaydin—3rd ed.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-262-02818-9 (hardcover : alk. paper)

1. Machine learning. I. Title

Q325.5.A46 2014

006.3'1—dc23

2014007214

CIP

10 9 8 7 6 5 4 3 2 1

## ***Brief Contents***

<i>1</i>	<i>Introduction</i>	<i>1</i>
<i>2</i>	<i>Supervised Learning</i>	<i>21</i>
<i>3</i>	<i>Bayesian Decision Theory</i>	<i>49</i>
<i>4</i>	<i>Parametric Methods</i>	<i>65</i>
<i>5</i>	<i>Multivariate Methods</i>	<i>93</i>
<i>6</i>	<i>Dimensionality Reduction</i>	<i>115</i>
<i>7</i>	<i>Clustering</i>	<i>161</i>
<i>8</i>	<i>Nonparametric Methods</i>	<i>185</i>
<i>9</i>	<i>Decision Trees</i>	<i>213</i>
<i>10</i>	<i>Linear Discrimination</i>	<i>239</i>
<i>11</i>	<i>Multilayer Perceptrons</i>	<i>267</i>
<i>12</i>	<i>Local Models</i>	<i>317</i>
<i>13</i>	<i>Kernel Machines</i>	<i>349</i>
<i>14</i>	<i>Graphical Models</i>	<i>387</i>
<i>15</i>	<i>Hidden Markov Models</i>	<i>417</i>
<i>16</i>	<i>Bayesian Estimation</i>	<i>445</i>
<i>17</i>	<i>Combining Multiple Learners</i>	<i>487</i>
<i>18</i>	<i>Reinforcement Learning</i>	<i>517</i>
<i>19</i>	<i>Design and Analysis of Machine Learning Experiments</i>	<i>547</i>
<i>A</i>	<i>Probability</i>	<i>593</i>

# *Contents*

*Preface*      xvii

*Notations*      xxi

## **1 *Introduction*      1**

- 1.1 What Is Machine Learning?      1
- 1.2 Examples of Machine Learning Applications      4
  - 1.2.1 Learning Associations      4
  - 1.2.2 Classification      5
  - 1.2.3 Regression      9
  - 1.2.4 Unsupervised Learning      11
  - 1.2.5 Reinforcement Learning      13
- 1.3 Notes      14
- 1.4 Relevant Resources      17
- 1.5 Exercises      18
- 1.6 References      20

## **2 *Supervised Learning*      21**

- 2.1 Learning a Class from Examples      21
- 2.2 Vapnik-Chervonenkis Dimension      27
- 2.3 Probably Approximately Correct Learning      29
- 2.4 Noise      30
- 2.5 Learning Multiple Classes      32
- 2.6 Regression      34
- 2.7 Model Selection and Generalization      37
- 2.8 Dimensions of a Supervised Machine Learning Algorithm      41
- 2.9 Notes      42

2.10	Exercises	43
2.11	References	47
<b>3</b>	<b><i>Bayesian Decision Theory</i></b>	<b>49</b>
3.1	Introduction	49
3.2	Classification	51
3.3	Losses and Risks	53
3.4	Discriminant Functions	55
3.5	Association Rules	56
3.6	Notes	59
3.7	Exercises	60
3.8	References	64
<b>4</b>	<b><i>Parametric Methods</i></b>	<b>65</b>
4.1	Introduction	65
4.2	Maximum Likelihood Estimation	66
4.2.1	Bernoulli Density	67
4.2.2	Multinomial Density	68
4.2.3	Gaussian (Normal) Density	68
4.3	Evaluating an Estimator: Bias and Variance	69
4.4	The Bayes' Estimator	70
4.5	Parametric Classification	73
4.6	Regression	77
4.7	Tuning Model Complexity: Bias/Variance Dilemma	80
4.8	Model Selection Procedures	83
4.9	Notes	87
4.10	Exercises	88
4.11	References	90
<b>5</b>	<b><i>Multivariate Methods</i></b>	<b>93</b>
5.1	Multivariate Data	93
5.2	Parameter Estimation	94
5.3	Estimation of Missing Values	95
5.4	Multivariate Normal Distribution	96
5.5	Multivariate Classification	100
5.6	Tuning Complexity	106
5.7	Discrete Features	108
5.8	Multivariate Regression	109
5.9	Notes	111
5.10	Exercises	112

5.11	References	113
<b>6</b>	<b><i>Dimensionality Reduction</i></b>	<b>115</b>
6.1	Introduction	115
6.2	Subset Selection	116
6.3	Principal Component Analysis	120
6.4	Feature Embedding	127
6.5	Factor Analysis	130
6.6	Singular Value Decomposition and Matrix Factorization	135
6.7	Multidimensional Scaling	136
6.8	Linear Discriminant Analysis	140
6.9	Canonical Correlation Analysis	145
6.10	Isomap	148
6.11	Locally Linear Embedding	150
6.12	Laplacian Eigenmaps	153
6.13	Notes	155
6.14	Exercises	157
6.15	References	158
<b>7</b>	<b><i>Clustering</i></b>	<b>161</b>
7.1	Introduction	161
7.2	Mixture Densities	162
7.3	k-Means Clustering	163
7.4	Expectation-Maximization Algorithm	167
7.5	Mixtures of Latent Variable Models	172
7.6	Supervised Learning after Clustering	173
7.7	Spectral Clustering	175
7.8	Hierarchical Clustering	176
7.9	Choosing the Number of Clusters	178
7.10	Notes	179
7.11	Exercises	180
7.12	References	182
<b>8</b>	<b><i>Nonparametric Methods</i></b>	<b>185</b>
8.1	Introduction	185
8.2	Nonparametric Density Estimation	186
8.2.1	Histogram Estimator	187
8.2.2	Kernel Estimator	188
8.2.3	k-Nearest Neighbor Estimator	190
8.3	Generalization to Multivariate Data	192



8.4	Nonparametric Classification	193
8.5	Condensed Nearest Neighbor	194
8.6	Distance-Based Classification	196
8.7	Outlier Detection	199
8.8	Nonparametric Regression: Smoothing Models	201
8.8.1	Running Mean Smoother	201
8.8.2	Kernel Smoother	203
8.8.3	Running Line Smoother	204
8.9	How to Choose the Smoothing Parameter	204
8.10	Notes	205
8.11	Exercises	208
8.12	References	210
<b>9</b>	<b><i>Decision Trees</i></b>	<b>213</b>
9.1	Introduction	213
9.2	Univariate Trees	215
9.2.1	Classification Trees	216
9.2.2	Regression Trees	220
9.3	Pruning	222
9.4	Rule Extraction from Trees	225
9.5	Learning Rules from Data	226
9.6	Multivariate Trees	230
9.7	Notes	232
9.8	Exercises	235
9.9	References	237
<b>10</b>	<b><i>Linear Discrimination</i></b>	<b>239</b>
10.1	Introduction	239
10.2	Generalizing the Linear Model	241
10.3	Geometry of the Linear Discriminant	242
10.3.1	Two Classes	242
10.3.2	Multiple Classes	244
10.4	Pairwise Separation	246
10.5	Parametric Discrimination Revisited	247
10.6	Gradient Descent	248
10.7	Logistic Discrimination	250
10.7.1	Two Classes	250
10.7.2	Multiple Classes	254
10.8	Discrimination by Regression	257

10.9	Learning to Rank	260
10.10	Notes	263
10.11	Exercises	263
10.12	References	266
<b>11</b>	<b><i>Multilayer Perceptrons</i></b>	<b>267</b>
11.1	Introduction	267
11.1.1	Understanding the Brain	268
11.1.2	Neural Networks as a Paradigm for Parallel Processing	269
11.2	The Perceptron	271
11.3	Training a Perceptron	274
11.4	Learning Boolean Functions	277
11.5	Multilayer Perceptrons	279
11.6	MLP as a Universal Approximator	281
11.7	Backpropagation Algorithm	283
11.7.1	Nonlinear Regression	284
11.7.2	Two-Class Discrimination	286
11.7.3	Multiclass Discrimination	288
11.7.4	Multiple Hidden Layers	290
11.8	Training Procedures	290
11.8.1	Improving Convergence	290
11.8.2	Overtraining	291
11.8.3	Structuring the Network	292
11.8.4	Hints	295
11.9	Tuning the Network Size	297
11.10	Bayesian View of Learning	300
11.11	Dimensionality Reduction	301
11.12	Learning Time	304
11.12.1	Time Delay Neural Networks	304
11.12.2	Recurrent Networks	305
11.13	Deep Learning	306
11.14	Notes	309
11.15	Exercises	311
11.16	References	313
<b>12</b>	<b><i>Local Models</i></b>	<b>317</b>
12.1	Introduction	317
12.2	Competitive Learning	318

12.2.1	Online $k$ -Means	318
12.2.2	Adaptive Resonance Theory	323
12.2.3	Self-Organizing Maps	324
12.3	Radial Basis Functions	326
12.4	Incorporating Rule-Based Knowledge	332
12.5	Normalized Basis Functions	333
12.6	Competitive Basis Functions	335
12.7	Learning Vector Quantization	338
12.8	The Mixture of Experts	338
12.8.1	Cooperative Experts	341
12.8.2	Competitive Experts	342
12.9	Hierarchical Mixture of Experts	342
12.10	Notes	343
12.11	Exercises	344
12.12	References	347
<b>13</b>	<b><i>Kernel Machines</i></b>	<b>349</b>
13.1	Introduction	349
13.2	Optimal Separating Hyperplane	351
13.3	The Nonseparable Case: Soft Margin Hyperplane	355
13.4	$\nu$ -SVM	358
13.5	Kernel Trick	359
13.6	Vectorial Kernels	361
13.7	Defining Kernels	364
13.8	Multiple Kernel Learning	365
13.9	Multiclass Kernel Machines	367
13.10	Kernel Machines for Regression	368
13.11	Kernel Machines for Ranking	373
13.12	One-Class Kernel Machines	374
13.13	Large Margin Nearest Neighbor Classifier	377
13.14	Kernel Dimensionality Reduction	379
13.15	Notes	380
13.16	Exercises	382
13.17	References	383
<b>14</b>	<b><i>Graphical Models</i></b>	<b>387</b>
14.1	Introduction	387
14.2	Canonical Cases for Conditional Independence	389
14.3	Generative Models	396

14.4	d-Separation	399
14.5	Belief Propagation	399
14.5.1	Chains	400
14.5.2	Trees	402
14.5.3	Polytrees	404
14.5.4	Junction Trees	406
14.6	Undirected Graphs: Markov Random Fields	407
14.7	Learning the Structure of a Graphical Model	410
14.8	Influence Diagrams	411
14.9	Notes	412
14.10	Exercises	413
14.11	References	415
<b>15</b>	<b><i>Hidden Markov Models</i></b>	<b>417</b>
15.1	Introduction	417
15.2	Discrete Markov Processes	418
15.3	Hidden Markov Models	421
15.4	Three Basic Problems of HMMs	423
15.5	Evaluation Problem	423
15.6	Finding the State Sequence	427
15.7	Learning Model Parameters	429
15.8	Continuous Observations	432
15.9	The HMM as a Graphical Model	433
15.10	Model Selection in HMMs	436
15.11	Notes	438
15.12	Exercises	440
15.13	References	443
<b>16</b>	<b><i>Bayesian Estimation</i></b>	<b>445</b>
16.1	Introduction	445
16.2	Bayesian Estimation of the Parameters of a Discrete Distribution	449
16.2.1	$K > 2$ States: Dirichlet Distribution	449
16.2.2	$K = 2$ States: Beta Distribution	450
16.3	Bayesian Estimation of the Parameters of a Gaussian Distribution	451
16.3.1	Univariate Case: Unknown Mean, Known Variance	451

16.3.2	Univariate Case: Unknown Mean, Unknown Variance	453
16.3.3	Multivariate Case: Unknown Mean, Unknown Covariance	455
16.4	Bayesian Estimation of the Parameters of a Function	456
16.4.1	Regression	456
16.4.2	Regression with Prior on Noise Precision	460
16.4.3	The Use of Basis/Kernel Functions	461
16.4.4	Bayesian Classification	463
16.5	Choosing a Prior	466
16.6	Bayesian Model Comparison	467
16.7	Bayesian Estimation of a Mixture Model	470
16.8	Nonparametric Bayesian Modeling	473
16.9	Gaussian Processes	474
16.10	Dirichlet Processes and Chinese Restaurants	478
16.11	Latent Dirichlet Allocation	480
16.12	Beta Processes and Indian Buffets	482
16.13	Notes	483
16.14	Exercises	484
16.15	References	485
<b>17</b>	<b>Combining Multiple Learners</b>	<b>487</b>
17.1	Rationale	487
17.2	Generating Diverse Learners	488
17.3	Model Combination Schemes	491
17.4	Voting	492
17.5	Error-Correcting Output Codes	496
17.6	Bagging	498
17.7	Boosting	499
17.8	The Mixture of Experts Revisited	502
17.9	Stacked Generalization	504
17.10	Fine-Tuning an Ensemble	505
17.10.1	Choosing a Subset of the Ensemble	506
17.10.2	Constructing Metalearners	506
17.11	Cascading	507
17.12	Notes	509
17.13	Exercises	511
17.14	References	513

<b>18 Reinforcement Learning</b>	<b>517</b>
18.1 Introduction	517
18.2 Single State Case: $K$ -Armed Bandit	519
18.3 Elements of Reinforcement Learning	520
18.4 Model-Based Learning	523
18.4.1 Value Iteration	523
18.4.2 Policy Iteration	524
18.5 Temporal Difference Learning	525
18.5.1 Exploration Strategies	525
18.5.2 Deterministic Rewards and Actions	526
18.5.3 Nondeterministic Rewards and Actions	527
18.5.4 Eligibility Traces	530
18.6 Generalization	531
18.7 Partially Observable States	534
18.7.1 The Setting	534
18.7.2 Example: The Tiger Problem	536
18.8 Notes	541
18.9 Exercises	542
18.10 References	544
<b>19 Design and Analysis of Machine Learning Experiments</b>	<b>547</b>
19.1 Introduction	547
19.2 Factors, Response, and Strategy of Experimentation	550
19.3 Response Surface Design	553
19.4 Randomization, Replication, and Blocking	554
19.5 Guidelines for Machine Learning Experiments	555
19.6 Cross-Validation and Resampling Methods	558
19.6.1 $K$ -Fold Cross-Validation	559
19.6.2 $5 \times 2$ Cross-Validation	560
19.6.3 Bootstrapping	561
19.7 Measuring Classifier Performance	561
19.8 Interval Estimation	564
19.9 Hypothesis Testing	568
19.10 Assessing a Classification Algorithm's Performance	570
19.10.1 Binomial Test	571
19.10.2 Approximate Normal Test	572
19.10.3 $t$ Test	572
19.11 Comparing Two Classification Algorithms	573
19.11.1 McNemar's Test	573

19.11.2	$K$ -Fold Cross-Validated Paired $t$ Test	573
19.11.3	$5 \times 2$ cv Paired $t$ Test	574
19.11.4	$5 \times 2$ cv Paired $F$ Test	575
19.12	Comparing Multiple Algorithms: Analysis of Variance	576
19.13	Comparison over Multiple Datasets	580
19.13.1	Comparing Two Algorithms	581
19.13.2	Multiple Algorithms	583
19.14	Multivariate Tests	584
19.14.1	Comparing Two Algorithms	585
19.14.2	Comparing Multiple Algorithms	586
19.15	Notes	587
19.16	Exercises	588
19.17	References	590
<b>A</b>	<b>Probability</b>	<b>593</b>
A.1	Elements of Probability	593
A.1.1	Axioms of Probability	594
A.1.2	Conditional Probability	594
A.2	Random Variables	595
A.2.1	Probability Distribution and Density Functions	595
A.2.2	Joint Distribution and Density Functions	596
A.2.3	Conditional Distributions	596
A.2.4	Bayes' Rule	597
A.2.5	Expectation	597
A.2.6	Variance	598
A.2.7	Weak Law of Large Numbers	599
A.3	Special Random Variables	599
A.3.1	Bernoulli Distribution	599
A.3.2	Binomial Distribution	600
A.3.3	Multinomial Distribution	600
A.3.4	Uniform Distribution	600
A.3.5	Normal (Gaussian) Distribution	601
A.3.6	Chi-Square Distribution	602
A.3.7	$t$ Distribution	603
A.3.8	$F$ Distribution	603
A.4	References	603
<b>Index</b>		<b>605</b>

## *Preface*

Machine learning must be one of the fastest growing fields in computer science. It is not only that the data is continuously getting “bigger,” but also the theory to process it and turn it into knowledge. In various fields of science, from astronomy to biology, but also in everyday life, as digital technology increasingly infiltrates our daily existence, as our digital footprint deepens, more data is continuously generated and collected. Whether scientific or personal, data that just lies dormant passively is not of any use, and smart people have been finding ever new ways to make use of that data and turn it into a useful product or service. In this transformation, machine learning plays a larger and larger role.

This data evolution has been continuing even stronger since the second edition appeared in 2010. Every year, datasets are getting larger. Not only has the number of observations grown, but the number of observed attributes has also increased significantly. There is more structure to the data: It is not just numbers and character strings any more but images, video, audio, documents, web pages, click logs, graphs, and so on. More and more, the data moves away from the parametric assumptions we used to make—for example, normality. Frequently, the data is dynamic and so there is a time dimension. Sometimes, our observations are multi-view—for the same object or event, we have multiple sources of information from different sensors and modalities.

Our belief is that behind all this seemingly complex and voluminous data, there lies a simple explanation. That although the data is big, it can be explained in terms of a relatively simple model with a small number of hidden factors and their interaction. Think about millions of customers who each day buy thousands of products online or from their local supermarket. This implies a very large database of transactions, but there is a



pattern to this data. People do not shop at random. A person throwing a party buys a certain subset of products, and a person who has a baby at home buys a different subset; there are hidden factors that explain customer behavior.

This is one of the areas where significant research has been done in recent years—namely, to infer this hidden model from observed data. Most of the revisions in this new edition are related to these advances. Chapter 6 contains new sections on feature embedding, singular value decomposition and matrix factorization, canonical correlation analysis, and Laplacian eigenmaps.

There are new sections on distance estimation in chapter 8 and on kernel machines in chapter 13: Dimensionality reduction, feature extraction, and distance estimation are three names for the same devil—the ideal distance measure is defined in the space of the ideal hidden features, and they are fewer in number than the values we observe.

Chapter 16 is rewritten and significantly extended to cover such generative models. We discuss the Bayesian approach for all major machine learning models, namely, classification, regression, mixture models, and dimensionality reduction. Nonparametric Bayesian modeling, which has become increasingly popular during these last few years, is especially interesting because it allows us to adjust the complexity of the model to the complexity of data.

New sections have been added here and there, mostly to highlight different recent applications of the same or very similar methods. There is a new section on outlier detection in chapter 8. Two new sections in chapters 10 and 13 discuss ranking for linear models and kernel machines, respectively. Having added Laplacian eigenmaps to chapter 6, I also include a new section on spectral clustering in chapter 7. Given the recent resurgence of deep neural networks, it became necessary to include a new section on deep learning in chapter 11. Chapter 19 contains a new section on multivariate tests for comparison of methods.

Since the first edition, I have received many requests for the solutions to exercises from readers who use the book for self-study. In this new edition, I have included the solutions to some of the more didactic exercises. Sometimes they are complete solutions, and sometimes they give just a hint or offer only one of several possible solutions.

I would like to thank all the instructors and students who have used the previous two editions, as well as their translations into German, Chinese, and Turkish, and their reprints in India. I am always grateful to those

who send me words of appreciation, criticism, or errata, or who provide feedback in any other way. Please keep them coming. My email address is `alpaydin@boun.edu.tr`. The book's web site is `http://www.cmpe.boun.edu.tr/~ethem/i2ml3e`.

It has been a pleasure to work with the MIT Press again on this third edition, and I thank Marie Lufkin Lee, Marc Lowenthal, and Kathleen Caruso for all their help and support.

## ***Notations***

$x$	Scalar value
$\mathbf{x}$	Vector
$\mathbf{X}$	Matrix
$\mathbf{x}^T$	Transpose
$\mathbf{X}^{-1}$	Inverse
$X$	Random variable
$P(X)$	Probability mass function when $X$ is discrete
$p(X)$	Probability density function when $X$ is continuous
$P(X Y)$	Conditional probability of $X$ given $Y$
$E[X]$	Expected value of the random variable $X$
$\text{Var}(X)$	Variance of $X$
$\text{Cov}(X, Y)$	Covariance of $X$ and $Y$
$\text{Corr}(X, Y)$	Correlation of $X$ and $Y$
$\mu$	Mean
$\sigma^2$	Variance
$\Sigma$	Covariance matrix
$m$	Estimator to the mean
$s^2$	Estimator to the variance
$\mathbf{S}$	Estimator to the covariance matrix

$\mathcal{N}(\mu, \sigma^2)$	Univariate normal distribution with mean $\mu$ and variance $\sigma^2$
$\mathcal{Z}$	Unit normal distribution: $\mathcal{N}(0, 1)$
$\mathcal{N}_d(\mu, \Sigma)$	$d$ -variate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$
$x$	Input
$d$	Number of inputs (input dimensionality)
$y$	Output
$r$	Required output
$K$	Number of outputs (classes)
$N$	Number of training instances
$z$	Hidden value, intrinsic dimension, latent factor
$k$	Number of hidden dimensions, latent factors
$C_i$	Class $i$
$\mathcal{X}$	Training sample
$\{x^t\}_{t=1}^N$	Set of $x$ with index $t$ ranging from 1 to $N$
$\{x^t, r^t\}_t$	Set of ordered pairs of input and desired output with index $t$
$g(x \theta)$	Function of $x$ defined up to a set of parameters $\theta$
$\arg \max_{\theta} g(x \theta)$	The argument $\theta$ for which $g$ has its maximum value
$\arg \min_{\theta} g(x \theta)$	The argument $\theta$ for which $g$ has its minimum value
$E(\theta \mathcal{X})$	Error function with parameters $\theta$ on the sample $\mathcal{X}$
$l(\theta \mathcal{X})$	Likelihood of parameters $\theta$ on the sample $\mathcal{X}$
$\mathcal{L}(\theta \mathcal{X})$	Log likelihood of parameters $\theta$ on the sample $\mathcal{X}$
$1(c)$	1 if $c$ is true, 0 otherwise
$\#\{c\}$	Number of elements for which $c$ is true
$\delta_{ij}$	Kronecker delta: 1 if $i = j$ , 0 otherwise