

Machine Learning and Data Mining Lecture Notes

CSC C11/D11

Department of Computer and Mathematical Sciences
University of Toronto Scarborough

Version: September 28, 2015

Copyright © 2015 Aaron Hertzmann, David Fleet and Marcus Brubaker

Conventions and Notation	iv
1 Introduction to Machine Learning	1
1.1 Types of Machine Learning	2
1.2 A simple problem	2
2 Linear Regression	5
2.1 The 1D case	5
2.2 Multidimensional inputs	6
2.3 Multidimensional outputs	8
3 Nonlinear Regression	9
3.1 Basis function regression	9
3.2 Overfitting and Regularization	11
3.3 Artificial Neural Networks	13
3.4 K -Nearest Neighbors	15
4 Quadratics	17
4.1 Optimizing a quadratic	18
5 Basic Probability Theory	21
5.1 Classical logic	21
5.2 Basic definitions and rules	22
5.3 Discrete random variables	24
5.4 Binomial and Multinomial distributions	25
5.5 Mathematical expectation	26
6 Probability Density Functions (PDFs)	27
6.1 Mathematical expectation, mean, and variance	28
6.2 Uniform distributions	29
6.3 Gaussian distributions	29
6.3.1 Diagonalization	31
6.3.2 Conditional Gaussian distribution	33
7 Estimation	35
7.1 Learning a binomial distribution	35
7.2 Bayes' Rule	36
7.3 Parameter estimation	37
7.3.1 MAP, ML, and Bayes' Estimates	38
7.4 Learning Gaussians	39
7.5 MAP nonlinear regression	40

8	Classification	42
8.1	Class Conditionals	42
8.2	Logistic Regression	44
8.3	Artificial Neural Networks	46
8.4	K -Nearest Neighbors Classification	46
8.5	Generative vs. Discriminative models	47
8.6	Classification by LS Regression	48
8.7	Naïve Bayes	49
8.7.1	Discrete Input Features	49
8.7.2	Learning	51
9	Gradient Descent	53
9.1	Finite differences	55
10	Cross Validation	56
10.1	Cross-Validation	56
11	Bayesian Methods	59
11.1	Bayesian Regression	60
11.2	Hyperparameters	63
11.3	Bayesian Model Selection	65
12	Monte Carlo Methods	69
12.1	Sampling Gaussians	70
12.2	Importance Sampling	70
12.3	Markov Chain Monte Carlo (MCMC)	73
13	Principal Components Analysis	75
13.1	The model and learning	75
13.2	Reconstruction	76
13.3	Properties of PCA	77
13.4	Whitening	78
13.5	Modeling	79
13.6	Probabilistic PCA	79
14	Lagrange Multipliers	83
14.1	Examples	84
14.2	Least-Squares PCA in one-dimension	87
14.3	Multiple constraints	90
14.4	Inequality constraints	90

15 Clustering	92
15.1 K -means Clustering	92
15.2 K -medoids Clustering	94
15.3 Mixtures of Gaussians	94
15.3.1 Learning	96
15.3.2 Numerical issues	98
15.3.3 The Free Energy	99
15.3.4 Proofs	100
15.3.5 Relation to K -means	101
15.3.6 Degeneracy	101
15.4 Determining the number of clusters	102
16 Hidden Markov Models	103
16.1 Markov Models	103
16.2 Hidden Markov Models	104
16.3 Viterbi Algorithm	106
16.4 The Forward-Backward Algorithm	107
16.5 EM: The Baum-Welch Algorithm	110
16.5.1 Numerical issues: renormalization	110
16.5.2 Free Energy	112
16.6 Most likely state sequences	114
17 Support Vector Machines	115
17.1 Maximizing the margin	115
17.2 Slack Variables for Non-Separable Datasets	117
17.3 Loss Functions	118
17.4 The Lagrangian and the Kernel Trick	120
17.5 Choosing parameters	121
17.6 Software	122
18 AdaBoost	123
18.1 Decision stumps	126
18.2 Why does it work?	126
18.3 Early stopping	128

Conventions and Notation

Scalars are written with lower-case italics, e.g., x . Column-vectors are written in bold, lower-case: \mathbf{x} , and matrices are written in bold uppercase: \mathbf{B} .

The set of real numbers is represented by \mathbb{R} ; N -dimensional Euclidean space is written \mathbb{R}^N .

Aside:

Text in “aside” boxes provide extra background or information that you are not required to know for this course.

Acknowledgements

Graham Taylor and James Martens assisted with preparation of these notes.