

# **Introduction to Machine Learning Lecture Notes**

**CSC C11**

Department of Computer and Mathematical Sciences  
University of Toronto Scarborough

Version: December 4, 2023

Copyright © 2023 Aaron Hertzmann and David Fleet

<b>Conventions and Notation</b>	<b>v</b>
<b>1 Introduction to Machine Learning</b>	<b>1</b>
1.1 A Simple Problem . . . . .	3
<b>2 Linear Regression</b>	<b>5</b>
2.1 The 1D Case . . . . .	5
2.2 Multi-Dimensional Inputs . . . . .	6
2.3 Multi-Dimensional Outputs . . . . .	8
<b>3 Nonlinear Regression</b>	<b>10</b>
3.1 Basis Function Regression . . . . .	10
3.2 Overfitting and Regularization . . . . .	13
3.3 Artificial Neural Networks . . . . .	15
3.4 $k$ -Nearest Neighbors . . . . .	16
<b>4 Quadratics</b>	<b>18</b>
4.1 Optimizing a Quadratic . . . . .	19
<b>5 Basic Probability Theory</b>	<b>22</b>
5.1 Classical Logic . . . . .	22
5.2 Basic Definitions and Rules . . . . .	23
5.3 Discrete Random Variables . . . . .	25
5.4 Binomial and Multinomial Distributions . . . . .	26
5.5 Mathematical Expectation . . . . .	27
<b>6 Probability Density Functions (PDFs)</b>	<b>28</b>
6.1 Mathematical Expectation, Mean, and Variance . . . . .	29
6.2 Uniform Distributions . . . . .	30
6.3 Gaussian Distributions . . . . .	30
6.3.1 Diagonalization . . . . .	31
6.3.2 Marginal and Conditional Distributions . . . . .	34
<b>7 Estimation</b>	<b>36</b>
7.1 Learning a Binomial Distribution . . . . .	36
7.2 Bayes' Rule . . . . .	37
7.3 Parameter Estimation . . . . .	38
7.3.1 MAP, ML and Bayes' Estimates . . . . .	39
7.4 Learning Gaussians . . . . .	40
7.5 MAP Nonlinear Regression . . . . .	40
<b>8 Information Theory</b>	<b>43</b>
8.1 Entropy . . . . .	43
8.2 Conditional and Relative Entropy . . . . .	44

8.3	Mutual Information . . . . .	45
8.4	Cross Entropy . . . . .	46
<b>9</b>	<b>Classification</b>	<b>47</b>
9.1	Classification by Regression . . . . .	47
9.2	k-Nearest Neighbors Classification . . . . .	49
9.3	Decision Trees . . . . .	50
9.3.1	Learning . . . . .	51
9.3.2	Decision Forests . . . . .	52
9.4	Class Conditionals . . . . .	53
9.5	Naïve Bayes . . . . .	54
9.5.1	Discrete Input Features . . . . .	55
9.5.2	Learning . . . . .	57
9.6	Logistic Regression . . . . .	58
9.6.1	Learning . . . . .	60
9.7	Generative vs. Discriminative Models . . . . .	62
<b>10</b>	<b>Gradient Descent</b>	<b>64</b>
10.1	Finite Differences . . . . .	66
<b>11</b>	<b>Cross Validation</b>	<b>67</b>
11.1	Hold-Out Validation . . . . .	67
11.2	<i>K</i> -Fold Cross Validation . . . . .	68
11.3	Issues with Cross Validation . . . . .	69
<b>12</b>	<b>Bayesian Methods</b>	<b>70</b>
12.1	Bayesian Regression . . . . .	71
12.2	Hyper-Parameters . . . . .	74
12.3	Bayesian Model Selection . . . . .	76
<b>13</b>	<b>Monte Carlo Methods</b>	<b>79</b>
13.1	Sampling Gaussians . . . . .	80
13.2	Sampling Categorical Distributions . . . . .	80
13.3	Importance Sampling . . . . .	81
13.4	Markov Chain Monte Carlo (MCMC) . . . . .	82
<b>14</b>	<b>Principal Component Analysis</b>	<b>84</b>
14.1	Modelling and Learning . . . . .	85
14.2	Representation and Reconstruction of New Data . . . . .	86
14.3	Properties of PCA . . . . .	87
14.4	Whitening . . . . .	89
14.5	Modeling . . . . .	89
14.6	Probabilistic PCA . . . . .	90

<b>15 Lagrange Multipliers</b>	<b>94</b>
15.1 Examples	95
15.2 Least-Squares PCA in 1D	98
15.3 Multiple Constraints	101
15.4 Inequality Constraints	101
<b>16 Clustering</b>	<b>103</b>
16.1 $K$ -Means	103
16.2 Hierarchical $K$ -Means	105
16.3 Product Quantization	106
16.4 $K$ -Medoids	106
16.5 Gaussian Mixture Models	107
16.5.1 Learning	108
16.5.2 Numerical Issues	110
16.5.3 Free Energy	110
16.5.4 Proofs	112
16.5.5 Relation to $k$ -Means	113
16.5.6 Degeneracy	114
16.6 Determining the number of clusters	114
<b>17 AdaBoost</b>	<b>115</b>
17.1 Decision Stumps	116
17.2 Why Does It Work?	117
17.3 Early Stopping	119
<b>18 Support Vector Machines</b>	<b>121</b>
18.1 Maximizing the Margin	121
18.2 Slack Variables for Non-Separable Datasets	123
18.3 Loss Functions	123
18.4 The Lagrangian and the Kernel Trick	125
18.5 Choosing Parameters	127
18.6 Software	128
<b>19 Hidden Markov Models</b>	<b>129</b>
19.1 Markov Models	129
19.2 Hidden Markov Models	129
19.3 Viterbi Algorithm	132
19.4 The Forward-Backward Algorithm	133
19.5 EM: The Baum-Welch Algorithm	136
19.5.1 Numerical Issues: Renormalization	136
19.5.2 Free Energy	138
19.6 Most Likely State Sequences	139

## Conventions and Notation

Scalars are typically written with lower-case italics, e.g.,  $x$ . Column-vectors are written in bold, lower-case, e.g.,  $\mathbf{x}$ . Matrices are usually written in bold uppercase, e.g.,  $\mathbf{B}$ .

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

*Aside:*

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

## Acknowledgements

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