

Contents

1 Preliminaries	1
1.1 Introduction	1
1.1.1 What is Machine Learning?	1
1.1.2 Wellsprings of Machine Learning	3
1.1.3 Varieties of Machine Learning	5
1.2 Learning Input-Output Functions	6
1.2.1 Types of Learning	6
1.2.2 Input Vectors	8
1.2.3 Outputs	9
1.2.4 Training Regimes	9
1.2.5 Noise	10
1.2.6 Performance Evaluation	10
1.3 Learning Requires Bias	10
1.4 Sample Applications	13
1.5 Sources	14
1.6 Bibliographical and Historical Remarks	15
2 Boolean Functions	17
2.1 Representation	17
2.1.1 Boolean Algebra	17
2.1.2 Diagrammatic Representations	18
2.2 Classes of Boolean Functions	19
2.2.1 Terms and Clauses	19
2.2.2 DNF Functions	20

2.2.3	CNF Functions	24
2.2.4	Decision Lists	25
2.2.5	Symmetric and Voting Functions	26
2.2.6	Linearly Separable Functions	26
2.3	Summary	27
2.4	Bibliographical and Historical Remarks	28
3	Using Version Spaces for Learning	29
3.1	Version Spaces and Mistake Bounds	29
3.2	Version Graphs	31
3.3	Learning as Search of a Version Space	34
3.4	The Candidate Elimination Method	35
3.5	Bibliographical and Historical Remarks	37
4	Neural Networks	39
4.1	Threshold Logic Units	39
4.1.1	Definitions and Geometry	39
4.1.2	Special Cases of Linearly Separable Functions	41
4.1.3	Error-Correction Training of a TLU	42
4.1.4	Weight Space	45
4.1.5	The Widrow-Hoff Procedure	46
4.1.6	Training a TLU on Non-Linearly-Separable Training Sets	49
4.2	Linear Machines	50
4.3	Networks of TLUs	51
4.3.1	Motivation and Examples	51
4.3.2	Madalines	54
4.3.3	Piecewise Linear Machines	56
4.3.4	Cascade Networks	57
4.4	Training Feedforward Networks by Backpropagation	58
4.4.1	Notation	58
4.4.2	The Backpropagation Method	60
4.4.3	Computing Weight Changes in the Final Layer	62
4.4.4	Computing Changes to the Weights in Intermediate Layers	64

4.4.5	Variations on Backprop	66
4.4.6	An Application: Steering a Van	66
4.5	Synergies Between Neural Network and Knowledge-Based Methods	68
4.6	Bibliographical and Historical Remarks	68
5	Statistical Learning	69
5.1	Using Statistical Decision Theory	69
5.1.1	Background and General Method	69
5.1.2	Gaussian (or Normal) Distributions	71
5.1.3	Conditionally Independent Binary Components	75
5.2	Learning Belief Networks	77
5.3	Nearest-Neighbor Methods	77
5.4	Bibliographical and Historical Remarks	79
6	Decision Trees	81
6.1	Definitions	81
6.2	Supervised Learning of Univariate Decision Trees	83
6.2.1	Selecting the Type of Test	83
6.2.2	Using Uncertainty Reduction to Select Tests	84
6.2.3	Non-Binary Attributes	88
6.3	Networks Equivalent to Decision Trees	88
6.4	Overfitting and Evaluation	89
6.4.1	Overfitting	89
6.4.2	Validation Methods	90
6.4.3	Avoiding Overfitting in Decision Trees	91
6.4.4	Minimum-Description Length Methods	92
6.4.5	Noise in Data	93
6.5	The Problem of Replicated Subtrees	94
6.6	The Problem of Missing Attributes	96
6.7	Comparisons	96
6.8	Bibliographical and Historical Remarks	96

7	Inductive Logic Programming	97
7.1	Notation and Definitions	99
7.2	A Generic ILP Algorithm	100
7.3	An Example	103
7.4	Inducing Recursive Programs	107
7.5	Choosing Literals to Add	110
7.6	Relationships Between ILP and Decision Tree Induction . .	111
7.7	Bibliographical and Historical Remarks	114
8	Computational Learning Theory	117
8.1	Notation and Assumptions for PAC Learning Theory	117
8.2	PAC Learning	119
8.2.1	The Fundamental Theorem	119
8.2.2	Examples	121
8.2.3	Some Properly PAC-Learnable Classes	122
8.3	The Vapnik-Chervonenkis Dimension	124
8.3.1	Linear Dichotomies	124
8.3.2	Capacity	126
8.3.3	A More General Capacity Result	127
8.3.4	Some Facts and Speculations About the VC Dimension	129
8.4	VC Dimension and PAC Learning	129
8.5	Bibliographical and Historical Remarks	130
9	Unsupervised Learning	131
9.1	What is Unsupervised Learning?	131
9.2	Clustering Methods	133
9.2.1	A Method Based on Euclidean Distance	133
9.2.2	A Method Based on Probabilities	136
9.3	Hierarchical Clustering Methods	138
9.3.1	A Method Based on Euclidean Distance	138
9.3.2	A Method Based on Probabilities	138
9.4	Bibliographical and Historical Remarks	143

10 Temporal-Difference Learning	145
10.1 Temporal Patterns and Prediction Problems	145
10.2 Supervised and Temporal-Difference Methods	146
10.3 Incremental Computation of the $(\Delta \mathbf{W})_i$	148
10.4 An Experiment with TD Methods	150
10.5 Theoretical Results	152
10.6 Intra-Sequence Weight Updating	153
10.7 An Example Application: TD-gammon	155
10.8 Bibliographical and Historical Remarks	156
11 Delayed-Reinforcement Learning	159
11.1 The General Problem	159
11.2 An Example	160
11.3 Temporal Discounting and Optimal Policies	161
11.4 Q -Learning	164
11.5 Discussion, Limitations, and Extensions of Q -Learning . . .	167
11.5.1 An Illustrative Example	167
11.5.2 Using Random Actions	169
11.5.3 Generalizing Over Inputs	170
11.5.4 Partially Observable States	171
11.5.5 Scaling Problems	172
11.6 Bibliographical and Historical Remarks	173
12 Explanation-Based Learning	175
12.1 Deductive Learning	175
12.2 Domain Theories	176
12.3 An Example	178
12.4 Evaluable Predicates	182
12.5 More General Proofs	183
12.6 Utility of EBL	183
12.7 Applications	183
12.7.1 Macro-Operators in Planning	184
12.7.2 Learning Search Control Knowledge	186
12.8 Bibliographical and Historical Remarks	187

Preface

These notes are in the process of becoming a textbook. The process is quite unfinished, and the author solicits corrections, criticisms, and suggestions from students and other readers. Although I have tried to eliminate errors, some undoubtedly remain—*caveat lector*. Many typographical infelicities will no doubt persist until the final version. More material has yet to be added. Please let me have your suggestions about topics that are too important to be left out. I hope that future versions will cover Hopfield nets, Elman nets and other recurrent nets, radial basis functions, grammar and automata learning, genetic algorithms, and Bayes networks . . . I am also collecting exercises and project suggestions which will appear in future versions. Yes, the final version will have a good index.

Some of my plans for additions and other reminders are mentioned in marginal notes.

My intention is to pursue a middle ground between a theoretical textbook and one that focusses on applications. The book concentrates on the important *ideas* in machine learning. I do not give proofs of many of the theorems that I state, but I do give plausibility arguments and citations to formal proofs. And, I do not treat many matters that would be of practical importance in applications; the book is not a handbook of machine learning practice. Instead, my goal is to give the reader sufficient preparation to make the extensive literature on machine learning accessible.

Students in my Stanford courses on machine learning have already made several useful suggestions, as have my colleague, Pat Langley, and my teaching assistants, Ron Kohavi, Karl Pfleger, Robert Allen, and Lise Getoor.