

Contents

1	Machine Learning Definition and Basics	1
1.1	Introduction	1
1.1.1	Resurgence of ML	2
1.1.2	Relation with Artificial Intelligence (AI)	3
1.1.3	Machine Learning Problems	4
1.2	Matrices	4
1.2.1	Vector and Tensors	5
1.2.2	Matrix Addition (or Subtraction)	5
1.2.3	Matrix Transpose	5
1.2.4	Matrix Multiplication	6
1.2.5	Identity Matrix	7
1.2.6	Matrix Inversion	7
1.2.7	Solving Equations Using Matrices	8
1.3	Numerical Methods	9
1.4	Probability and Statistics	10
1.4.1	Sampling the Distribution	11
1.4.2	Random Variables	11
1.4.3	Expectation	11
1.4.4	Conditional Probability and Distribution	12
1.4.5	Maximum Likelihood	12
1.5	Linear Algebra	13
1.6	Differential Calculus	14
1.6.1	Functions	14
1.6.2	Slope	14
1.7	Computer Architecture	15
1.8	Next Steps	16
2	Learning Models	19
2.1	Supervised Learning	19
2.1.1	Classification Problem	20

2.1.2	Regression Problem	20
2.2	Unsupervised Learning	21
2.3	Semi-supervised Learning	22
2.4	Reinforcement Learning	22
3	Regressions	25
3.1	Introduction	25
3.2	The Model	25
3.3	Problem Formulation	26
3.4	Linear Regression	27
3.4.1	Normal Method	28
3.4.2	Gradient Descent Method	30
3.4.3	Normal Equation Method vs Gradient Descent Method	36
3.5	Logistic Regression	36
3.5.1	Sigmoid Function	37
3.5.2	Cost Function	38
3.5.3	Gradient Descent	38
3.6	Next Steps	39
3.7	Key Takeaways	40
4	Improving Further	41
4.1	Nonlinear Contribution	41
4.2	Feature Scaling	42
4.3	Gradient Descent Algorithm Variations	43
4.3.1	Cost Contour	43
4.3.2	Stochastic Gradient Descent	44
4.3.3	Mini Batch Gradient Descent	47
4.3.4	Map Reduce and Parallelism	48
4.3.5	Basic Theme of Algorithm Variations	49
4.4	Regularization	49
4.4.1	Regularization for Normal Equation	52
4.4.2	Regularization for Logistic Regression	52
4.4.3	Determining Appropriate λ	52
4.4.4	Comparing Hypothesis	53
4.5	Multi-class Classifications	54
4.5.1	One-vs-All Classification	54
4.5.2	SoftMax	55
4.6	Key Takeaways and Next Steps	56
5	Classification	57
5.1	Decision Boundary	57
5.1.1	Nonlinear Decision Boundary	58
5.2	Skewed Class	60
5.2.1	Optimizing Precision vs Recall	61
5.2.2	Single Metric	61
5.3	Naïve Bayes' Algorithm	62

5.4	Support Vector Machines	63
5.4.1	Kernel Selection	66
6	Clustering	67
6.1	<i>K</i> -Means	67
6.1.1	Basic Algorithm	68
6.1.2	Distance Calculation	68
6.1.3	Algorithm Pseudo Code	69
6.1.4	Cost Function	69
6.1.5	Choice of Initial Random Centers	70
6.1.6	Number of Clusters	71
6.2	<i>K</i> -Nearest Neighbor (KNN)	72
6.2.1	Weight Consideration	73
6.2.2	Feature Scaling	73
6.2.3	Limitations	73
6.2.4	Finding the Nearest Neighbors	74
6.3	Next Steps	76
7	Random Forests	77
7.1	Decision Tree	77
7.2	Information Gain	79
7.3	Gini Impurity Criterion	86
7.4	Disadvantages of Decision Trees	89
7.5	Random Forests	89
7.5.1	Data Bagging	90
7.5.2	Feature Bagging	90
7.5.3	Cross Validation in Random Forests	90
7.5.4	Prediction	91
7.6	Variable Importance	91
7.7	Proximities	92
7.7.1	Outliers	92
7.7.2	Prototypes	93
7.8	Disadvantages of Random Forests	93
7.9	Next Steps	94
8	Testing the Algorithm and the Network	95
8.1	Test Set	95
8.2	Overfit	96
8.3	Underfit	96
8.4	Determining the Number of Degrees	96
8.5	Determining λ	97
8.6	Increasing Data Count	98
8.6.1	High Bias Case	98
8.6.2	High Variance Case	99
8.7	The Underlying Mathematics (Optional)	99
8.8	Utilizing the Bias vs Variance Information	101

8.9	Derived Data	101
8.10	Approach	102
8.11	Test Data	102
9	(Artificial) Neural Networks	103
9.1	Logistic Regression Extended to Form Neural Network	103
9.2	Neural Network as Oversimplified Brain	105
9.3	Visualizing Neural Network Equations	106
9.4	Matrix Formulation of Neural Network	107
9.5	Neural Network Representation	108
9.6	Starting to Design a Neural Network	109
9.7	Training the Network	110
9.7.1	Chain Rule	111
9.7.2	Components of Gradient Computation	112
9.7.3	Gradient Computation Through Backpropagation	114
9.7.4	Updating Weights	115
9.8	Vectorization	116
9.9	Controlling Computations	116
9.10	Next Steps	116
10	Natural Language Processing	117
10.1	Complexity of NLP	117
10.2	Algorithms	119
10.2.1	Rule-Based Processing	119
10.2.2	Tokenizer	119
10.2.3	Named Entity Recognizers	120
10.2.4	Term Frequency-Inverse Document Frequency (<i>tf-idf</i>)	121
10.2.5	Word Embedding	122
10.2.6	Word2vec	123
11	Deep Learning	127
11.1	Recurrent Neural Networks	127
11.1.1	Representation of RNN	129
11.1.2	Backpropagation in RNN	132
11.1.3	Vanishing Gradients	133
11.2	LSTM	134
11.3	GRU	136
11.4	Self-Organizing Maps	138
11.4.1	Representation and Training of SOM	138
12	Principal Component Analysis	141
12.1	Applications of PCA	141
12.1.1	Example 1	141
12.1.2	Example 2	142
12.2	Computing PCA	143
12.2.1	Data Representation	143
12.2.2	Covariance Matrix	143

12.2.3	Diagonal Matrix	144
12.2.4	Eigenvector	144
12.2.5	Symmetric Matrix	144
12.2.6	Deriving Principal Components	144
12.2.7	Singular Value Decomposition (SVD)	145
12.3	Computing PCA	145
12.3.1	Data Characteristics	145
12.3.2	Data Preprocessing	146
12.3.3	Selecting Principal Components	146
12.4	PCA Applications	148
12.4.1	Image Compression	148
12.4.2	Data Visualization	149
12.5	Pitfalls of PCA Application	151
12.5.1	Overfitting	151
12.5.2	Model Generation	152
12.5.3	Model Interpretation	152
13	Anomaly Detection	153
13.1	Anomaly vs Classification	153
13.2	Model	154
13.2.1	Distribution Density	155
13.2.2	Estimating Distribution Parameters	156
13.2.3	Metric Value	156
13.2.4	Finding ϵ	157
13.2.5	Validating and Tuning the Model	157
13.3	Multivariate Gaussian Distribution	158
13.3.1	Determining Feature Mean	159
13.3.2	Determining Covariance	159
13.3.3	Computing and Applying the Metric	160
13.4	Anomalies in Time Series	160
13.4.1	Time Series Decomposition	161
13.4.2	Time Series Anomaly Types	161
13.4.3	Anomaly Detection in Time Series	163
14	Recommender Systems	169
14.1	Features Known	169
14.1.1	User's Affinity Toward Each Feature	170
14.2	User's Preferences Known	171
14.2.1	Characterizing Features	172
14.3	Features and User Preferences Both Unknown	173
14.3.1	Collaborative Filtering	173
14.3.2	Predicting and Recommending	175
14.4	New User	176
14.4.1	Shortcomings of the Current Algorithm	177
14.4.2	Mean Normalization	178
14.5	Tracking Changes in Preferences	178

15	Convolution	181
15.1	Convolution Explained	181
15.2	Object Identification Example	183
15.2.1	Exact Shape Known	183
15.2.2	Exact Shape Not Known	184
15.2.3	Breaking Down Further	184
15.2.4	Unanswered Questions	185
15.3	Image Convolution	185
15.4	Preprocessing	187
15.5	Post-Processing	188
15.6	Stride	189
15.7	CNN	190
15.8	Matrix Operation	191
15.9	Refining the Filters	192
15.10	Pooling as Neural Network	193
15.11	Character Recognition and Road Signs	193
15.12	ADAS and Convolution	193
16	Components of Reinforcement Learning	195
16.1	Key Participants of a Reinforcement Learning System	195
16.1.1	The Agent	195
16.1.2	The Environment	198
16.1.3	Interaction Between Agent and Environment	199
16.2	Environment State Transitions and Actions	200
16.2.1	Deterministic Environment	200
16.2.2	Stochastic Environment	201
16.2.3	Markov States and MDP	202
16.3	Agent's Objective	203
16.4	Agent's Behavior	204
16.5	Graphical Notation for a Trajectory	205
16.6	Value Function	205
16.6.1	State-Value Function	206
16.6.2	Action-Value Function	206
16.7	Methods for Finding Optimal Policies	208
16.7.1	Agent's Awareness of MDP	208
16.7.2	Model-Based and Model-Free Reinforcement Learning	210
16.7.3	On-Policy and Off-Policy Reinforcement Learning	210
16.8	Policy Iteration Method for Optimal Policy	210
16.8.1	Computing Q-function for a Given Policy	211
16.8.2	Policy Iteration	211
17	Reinforcement Learning Algorithms	213
17.1	Monte Carlo Learning	213
17.1.1	State Value Estimation	213
17.1.2	Action Value Estimation	215

17.2	Estimating Action Values with TD Learning	215
17.3	Exploration vs Exploitation Trade-Off	217
17.3.1	ϵ -greedy Policy	217
17.4	Q -learning	218
17.5	Scaling Through Function Approximation	219
17.5.1	Approximating the Q -function in Q -learning	220
17.6	Policy-Based Methods	220
17.6.1	Advantages of Policy Gradient Methods	221
17.6.2	Parameterized Policy	221
17.6.3	Training the Model	222
17.6.4	Monte Carlo Gradient Methods	223
17.6.5	Actor-Critic Methods	223
17.6.6	Reducing Variability in Gradient Methods	224
17.7	Simulation-Based Learning	225
17.8	Monte Carlo Tree Search (MCTS)	227
17.8.1	Search Tree	227
17.8.2	Monte Carlo Search Tree	228
17.8.3	MCTS Algorithm	231
17.8.4	Pseudo Code for MCTS Algorithm	233
17.8.5	Parallel MCTS Algorithms	235
17.9	MCTS Tree Values for Two-Player Games	235
17.10	Alpha Zero	236
17.10.1	Overview	236
17.10.2	Aspects of Alpha Zero	238
17.10.3	MCTS Search	239
18	Designing a Machine Learning System	243
18.1	Pipeline Systems	243
18.1.1	Ceiling Analysis	244
18.2	Data Quality	244
18.2.1	Unstructured Data	246
18.2.2	Getting Data	246
18.3	Improvisations over Gradient Descent	247
18.3.1	Momentum	247
18.3.2	RMSProp	248
18.3.3	ADAM (Adaptive Moment Estimation)	249
18.4	Software Stacks	250
18.4.1	TensorFlow	250
18.4.2	MXNet	251
18.4.3	pyTorch	252
18.4.4	The Microsoft Cognitive Toolkit	252
18.4.5	Keras	253
18.5	Choice of Hardware	253
18.5.1	Traditional Computer Systems	253
18.5.2	GPU	254

18.5.3	FPGAs	255
18.5.4	TPUs	255
Bibliography		257
Index		259