

Contents

29 (Social) Network Data Analysis	1873
29.1 Case Studies	1876
29.1.1 BlueDot	1876
29.1.2 6 Degrees of Kevin Bacon	1877
29.2 Preliminaries	1878
29.2.1 Graph Theory	1879
29.2.2 Properties	1881
29.2.3 Network Data	1883
29.2.4 Dyads, Triads, Interactions	1886
29.2.5 Measures	1887
29.2.6 Visualizations	1889
29.2.7 Roles and Position	1893
29.2.8 Ethics	1893
29.3 Methods	1896
29.3.1 Communities	1896
29.3.2 Paths and Connectivity	1897
29.4 Dynamic Networks	1900
29.5 Networks and ML	1901
29.5.1 GNN	1901
29.5.2 Feature Extraction	1902
29.5.3 Network Data in NLP	1903
29.5.4 Graph Clustering	1903
29.6 Networks and Big Data	1904
29.7 Applications	1905
29.7.1 Social Networks	1905
29.7.2 Ecological Networks	1907
29.7.3 Transport Networks	1910
29.7.4 BookCrossing Network	1914
29.7.5 Hockey Network	1916
29.8 Examples	1922
29.8.1 Simple Example	1922
29.8.2 Erdős Network	1935
29.8.3 BookCrossing Network	1936
29.8.4 Hockey Network	1941
29.9 Exercises	1947
Chapter References	1950
30 What's the Big Deal with Big Data?	1955
30.1 Danish Medical Data	1956
30.2 Motivation	1960
30.2.1 Big Data in Practice	1960
30.2.2 Big Data Across Domains	1960
30.2.3 Value, Cost, Decisions	1961
30.2.4 Common Misconceptions	1961

30.2.5	Big Data vs. Small Data	1962
30.2.6	Data Sources	1962
30.2.7	5-V Paradigm	1963
30.2.8	Failure Modes	1964
30.3	Proceeding With Caution	1965
30.3.1	Regime Change	1965
30.3.2	Ethics and Values	1966
30.3.3	Truly Large Numbers	1967
30.3.4	Practical Guidance	1967
30.4	Distributed Computing	1968
30.4.1	One-Machine Breakdowns	1968
30.4.2	Distributed vs. Parallel	1969
30.4.3	Intuitive Analogies	1970
30.4.4	When It Helps	1972
30.5	Hardware Solutions	1973
30.5.1	Device-Level Parallelism	1974
30.5.2	Scaling	1974
30.5.3	Cost vs. Performance	1975
30.5.4	Benchmarking	1976
30.6	Software Solutions	1979
30.6.1	R Tools	1979
30.6.2	Python Tools	1981
30.6.3	Frameworks	1982
30.6.4	Selecting a Framework	1983
30.7	Practical Considerations	1984
30.7.1	MapReduce and Hadoop	1984
30.7.2	Apache Spark	1986
30.7.3	Parquet & Column Storage	1987
30.7.4	Working With Clusters	1988
30.7.5	AMS EMR Workflow	1992
30.8	ML at Scale in Spark	1997
30.8.1	Naïve Bayes	1997
30.8.2	k -Means and Initialization	2004
30.8.3	Streaming k -Means	2009
30.8.4	Regression/Regularization	2012
30.8.5	Tree-Based Methods	2016
30.8.6	Class Imbalance	2020
30.9	Exercises	2024
	Chapter References	2027

31 A Deep Learning Launchpad 2029

31.1	Brief Overview of Tensors	2030
31.1.1	Tensor Products	2031
31.1.2	Tensor Decomposition	2035
31.1.3	Python Examples	2036
31.2	Deep Networks	2042
31.2.1	Getting Started	2044
31.2.2	Activation Functions	2046
31.2.3	Weight Initialization	2048
31.3	Regularization	2048
31.3.1	Weight Decay	2048

31.3.2	Early Stopping2049
31.3.3	Dropout2049
31.4	Stochastic Descent2050
31.4.1	Momentum2050
31.4.2	Nesterov Momentum2051
31.4.3	AdaGrad2051
31.4.4	RMSprop2052
31.4.5	AdaDelta2052
31.4.6	Adam2053
31.4.7	Yogi2053
31.5	CNN2054
31.5.1	What is Convolution?2054
31.5.2	How it is Used2055
31.5.4	Image Class Example2058
31.6	RNN2064
31.6.1	Vector-to-Sequence2065
31.6.2	Sequence-to-Vector2065
31.6.3	Sequence-to-Sequence2065
31.6.4	Encode-Decode Models2066
31.6.5	Bi-Directional RNN2066
31.6.6	Long-Term Memory2067
31.6.7	Music Generation2068
31.7	Specialized Architectures2076
31.7.1	GAN2077
31.7.2	Autoencoders2079
31.7.3	Transformers2082
31.8	Additional Examples2083
31.8.1	Learning XOR2084
31.8.2	Iris Dataset2086
31.8.3	Boston Dataset2089
31.8.4	MNIST Dataset2092
31.8.5	Sunspot Dataset2100
31.9	Exercises2105
	Chapter References2107
32	Natural Language Processing	2109
32.1	Introduction2112
32.1.1	Learning Basics2113
32.1.2	Train/Test/Validate2114
32.1.3	Linear Models2114
32.1.4	Training as Optimization2125
32.1.5	Regularization2126
32.1.6	Gradient Descent2128
32.1.7	Linear Model Limitations2129
32.1.8	FFNN2130
32.1.9	Dropout2131
32.2	Natural Language Data2132
32.2.1	Syntax vs. Semantics2132
32.2.2	Classification Problems2138
32.2.3	Features for NLP Problems2139
32.2.4	Linguistic Annotation2143

32.2.5	Text Features	2146
32.3	NLP Tasks	2148
32.4	From Text to Inputs	2153
32.4.1	One-Hot Encoduing	2153
32.4.2	Dense Encoding	2154
32.4.3	Combining Dense Vectors	2155
32.4.4	Example: POS Tagging	2157
32.4.5	Example: Arc Parsing	2158
32.5	Language Modeling	2160
32.5.1	Limitations	2162
32.5.2	Neural Language Models	2162
32.6	Word Embeddings	2163
32.6.1	Random Initialization	2163
32.6.2	Pre-Trained Embeddings	2164
32.6.3	Algorithms	2165
32.6.4	Word2Vec	2167
32.6.5	Choice of Contexts	2168
32.6.6	Using Embeddings	2171
32.6.7	Pitfalls	2174
32.7	Other NLP Concepts	2175
32.7.1	Sentence Inference	2175
32.7.2	Large Language Models	2177
32.7.3	Topic Models	2185
32.8	Examples	2199
32.8.1	Sentiment Analysis	2199
32.8.2	Tweet Classification	2202
32.8.3	NLTK Intro	2226
32.8.4	Text Generation	2243
32.8.5	Topic Modeling	2250
32.8.6	Summarizing	2254
32.9	Exercises	2265
	Chapter References	2266

List of Figures

29.2	Six Degrees of Kevin Bacon (subset)	1878
29.3	Simple friendship graph	1879
29.4	Undirected friendship graph and directed social media graph	1880
29.5	Weighted graph	1880
29.6	Vertex degrees	1881
29.7	Path length between vertices	1882
29.8	Connected and disconnected graphs	1882
29.9	Strongly connected and weakly connected graphs	1883
29.10	Common data representations	1884

29.11	Attributed network	1885
29.12	Network dyads	1886
29.13	Network triads	1886
29.14	Network visualizations I	1890
29.15	Network visualizations II	1891
29.16	Network visualizations III	1892
29.17	Structural positions	1894
29.18	Module partition and population dynamics	1909
29.19	Construction process of Dalian bus and metro composite network	1912
29.20	Key Dalian bus and metro stations	1913
29.21	Map of Nairobi’s informal minibus	1913
29.22	Top 120 strongest nodes in the filtered BookCrossing subgraph	1915
29.23	Top 120 strongest nodes in the filtered BookCrossing subgraph, with Louvain communities	1917
29.24	Senators players’ network for the first preseason game of 2024–25	1920
29.25	Senators players’ network for the 2024–25 season	1921
29.26	OOTS cast of characters	1923
29.27	OOTS social network	1929
30.1	Diagnoses in the Danish medical dataset	1957
30.2	Diagnosis trajectory clusters in the Danish medical dataset	1959
30.3	Parallel vs. distributed computing	1969
30.4	Parallel counting by aggregation	1970
30.5	Bottlenecks limit speed-ups	1971
30.6	Trends in the cost of computing	1979
30.7	The MapReduce pattern	1984
30.8	Creating the EMR cluster	1992
30.9	Setting up instances	1993
30.10	Creating a key pair	1993
31.1	Examples of tensors	2031
31.2	Two-layer perceptron.	2043
31.3	Deep network.	2043
31.4	The ReLU activation function.	2046
31.5	The sigmoid and hyperbolic tangent functions.	2047
31.6	Dropout regularization meme.	2050
31.7	The convolution of an input X and kernel Y	2054
31.8	Stride of 3 skips	2056
31.9	Stride of 2 skips	2056
31.10	Various kernel paddings	2057
31.11	Various kernel sizes	2057
31.12	A deep convolutional network with four convolutional layers and two dense layers	2057
31.13	“Unrolling” in a recurrent network.	2064
31.14	Schematics for a simple autoencoder.	2079
31.15	A variational autoencoder.	2081
31.16	Linear non-separability of the XOR function	2084
32.1	A sample conversation with ELIZA	2110
32.2	A conversation with SHRDLU	2110
32.3	Sigmoid function	2124
32.4	A single-neuron artificial neural network.	2130
32.5	An MLP with two hidden layers.	2130

32.6	A constituency (phrase structure) tree for the sample sentence	2144
32.7	A dependency tree for the sample sentence.	2144
32.8	Semantic role labeling for the sample sentence.	2145
32.9	Deriving the governor and object from a syntactic tree.	2152
32.10	Markov chain built from a short sequence of lyrics to <i>Imagine</i>	2246

List of Tables

29.1	Summary table: types of graphs	1880
29.2	Top 10 strongest book–book connections	1914
29.3	Neon green community	1918
29.4	Dark blue community	1918
29.5	Light neon green community	1918
29.6	Summary table of Senators statistics from the first preseason game of 2024–25	1919
29.7	Summary table of Senators statistics from the 2024–25 season	1922
32.1	Basic word orders and their approximate prevalence	2132