

# Contents

<b>24 Queueing Models</b>	<b>1573</b>
24.1 Background	1573
24.2 Terminology	1575
24.2.1 Input/Arrival Processes	1578
24.2.2 Output/Service Processes	1579
24.2.3 Queue Discipline	1581
24.2.4 Joining a Queue	1581
24.3 Theoretical Framework	1582
24.3.1 Kendall-Lee Notation	1582
24.3.2 Birth-Death Processes	1583
24.3.3 Little's Queueing Formula	1584
24.4 $M/M/1$ Queueing Systems	1585
24.4.1 Basics	1585
24.4.2 Limited Capacity	1587
24.5 $M/M/c$ Queueing Systems	1589
24.6 Exercises	1592
Chapter References	1592
<b>25 Bayesian Data Analysis</b>	<b>1593</b>
25.1 Plausible Reasoning	1593
25.1.1 Rules of Probability	1594
25.1.2 Bayes' Theorem	1596
25.1.3 Bayesian Inference Basics	1598
25.1.4 Bayesian Data Analysis	1600
25.2 Simple Examples	1601
25.2.1 The Mysterious Coin	1601
25.2.2 The Salary Question	1603
25.2.3 Money (\$ Bill Y'All)	1607
25.3 Prior Distributions	1614
25.3.1 Conjugate Priors	1614
25.3.2 Uninformative Priors	1615
25.3.3 Informative Priors	1616
25.3.4 Maximum Entropy Priors	1620
25.4 Posterior Distributions	1623
25.4.1 High-Density Regions	1624
25.4.2 MCMC Methods	1626
25.4.3 The MH Algorithm	1626
25.5 Additional Topics	1633
25.5.1 Uncertainty	1633
25.5.2 Bayesian A/B Testing	1635
25.6 Exercises	1639
Chapter References	1642

<b>26</b>	<b>Anomaly Detection and Outlier Analysis</b>	<b>1643</b>
26.1	Overview . . . . .	1643
26.1.1	Basic Notions & Concepts . . . . .	1643
26.1.2	ML Framework . . . . .	1648
26.1.3	Motivating Example . . . . .	1655
26.2	Quantitative Approaches . . . . .	1658
26.2.1	Distance Methods . . . . .	1658
26.2.2	Density Methods . . . . .	1669
26.3	Qualitative Approaches . . . . .	1683
26.3.1	AVF Algorithm . . . . .	1684
26.3.2	Greedy Algorithm . . . . .	1685
26.4	High-Dimensional Data . . . . .	1686
26.4.1	Definitions and Challenges . . . . .	1687
26.4.2	Projection Methods . . . . .	1687
26.4.3	Subspace Methods . . . . .	1697
26.4.4	Ensemble Methods . . . . .	1698
26.5	Exercices . . . . .	1702
	Chapter References . . . . .	1703
<b>27</b>	<b>Text Analysis and Text Mining</b>	<b>1705</b>
27.1	Introduction . . . . .	1705
27.1.1	Case Study: BOTUS . . . . .	1705
27.1.2	Text Analysis . . . . .	1710
27.1.3	TM vs. NLP . . . . .	1711
27.2	Basics of Text Analysis . . . . .	1713
27.2.1	Text Collection . . . . .	1715
27.2.2	Text Representation . . . . .	1716
27.2.3	Text Processing . . . . .	1716
27.2.4	Text Statistics . . . . .	1721
27.2.5	Text Visualization . . . . .	1723
27.3	Text Mining Tasks . . . . .	1724
27.3.1	Classification . . . . .	1725
27.3.2	Clustering . . . . .	1728
27.3.3	Sentiment Analysis . . . . .	1729
27.4	Examples . . . . .	1734
27.4.1	NHL Game Recaps I . . . . .	1734
27.4.2	Shakespeare vs. Marlowe . . . . .	1749
27.4.3	The Play's the Thing . . . . .	1765
27.4.4	Ham or Spam . . . . .	1775
27.4.5	NHL Game Recaps II . . . . .	1789
27.4.6	The Scottish Play . . . . .	1794
27.4.7	Regular Expressions . . . . .	1808
27.4.8	Movie Reviews . . . . .	1812
27.5	Exercises . . . . .	1820
	Chapter References . . . . .	1821

<b>28 Mining Data Streams</b>	<b>1823</b>
28.1 Overview	1823
28.1.1 Motivating Examples	1823
28.1.2 Basic Notions	1824
28.2 Change Detection and Maintaining Statistics	1831
28.2.1 Change Detection	1831
28.2.2 Maintaining Statistics	1834
28.3 Clustering	1839
28.3.1 Basics and Challenges	1839
28.3.2 Approaches	1839
28.3.3 Evaluation	1841
28.3.4 Algorithms	1842
28.4 Classification	1844
28.4.1 Basics and Challenges	1844
28.4.2 Approaches	1846
28.4.3 Ensemble Classifiers	1849
28.5 Frequent Itemset Mining	1850
28.6 Examples	1856
28.6.1 Obtaining Statistics	1856
28.6.2 Bloom Filter	1857
28.6.3 Sampling (Reservoir)	1860
28.6.4 Sampling (Hash Function)	1862
28.6.5 Fading Window	1863
28.6.6 ADWIN	1865
28.6.7 PID	1867
28.6.8 Histogram Drift	1869
28.7 Exercises	1870
Chapter References	1870

## List of Figures

24.1 Components of a generic queueing system	1575
24.2 Poisson and exponential distributions	1577
24.3 Erlang random variables	1579
24.4 Single line at bank with three tellers – $M/M/3/FCFS/20/\infty$	1583
24.5 Birth-death process	1583
24.6 Schematics of steady state vs. transient behaviour	1584
24.7 Generic $M/M/c$ queue	1590
25.1 Deductive vs. inductive reasoning	1594
25.2 4 priors for the fair coin problem	1602
25.3 Posteriors for a different numbers of tosses; 4 priors, same data	1604
25.4 Two priors for the salary problem	1605
25.5 Posteriors for the salary problem – one per priors	1606
25.6 Marginal posteriors for the salary problem – one per priors	1606
25.7 Catch-and-release schematics in the simple model	1607

25.8	Catch-and-release schematics in the brittle model . . . . .	1610
25.9	Catch-and-release schematics in the expert model . . . . .	1612
26.1	A school of fish . . . . .	1645
26.2	Tukey's boxplot test . . . . .	1647
26.3	Multi-modal supply chain corridor . . . . .	1649
26.4	Illustration of how to derive the various monthly fluidity indicators . . . . .	1649
26.5	Conceptual time series decomposition . . . . .	1650
26.6	Marine transit CV data, from 2010 to 2013 . . . . .	1651
26.7	Diagnostic plot for marine transit CV data . . . . .	1651
26.8	Adjusted plot for marine transit CV data . . . . .	1651
26.9	Oversampling, undersampling, and hybrid strategy for anomaly detection . . . . .	1654
26.10	Generating artificial cases with SMOTE and DRAMOTE . . . . .	1655
26.11	Illustration of autoencoder compression/reconstruction for anomaly detection . . . . .	1656
26.12	2D visualization of various similarity metrics . . . . .	1665
26.13	Low-density areas as outlier nurseries . . . . .	1669
26.14	Illustration of $k$ -local density . . . . .	1670
26.15	Algorithm: LOF . . . . .	1671
26.16	Illustration of reachability . . . . .	1672
26.17	Illustration of DBSCAN main concepts . . . . .	1673
26.18	Algorithm: DBSCAN . . . . .	1674
26.19	DBSCAN clustering outcomes . . . . .	1677
26.20	HDBSCAN/OPTICS clustering outcomes . . . . .	1678
26.21	Algorithm: IsoTree . . . . .	1680
26.22	Isolation Forest schematics . . . . .	1681
26.23	Algorithm: IsoForest . . . . .	1682
26.24	3-way, 2-way, and 1-way tables for the artificial example . . . . .	1683
26.25	Algorithm: AVF . . . . .	1685
26.26	Data analytical tasks that are not aligned with PCA . . . . .	1690
26.27	Algorithm: FB . . . . .	1697
26.28	Algorithm: SE . . . . .	1699
26.29	Algorithm: IE . . . . .	1700
27.1	T3's Trump and Dump process . . . . .	1707
27.2	Examples of @realDonaldTrump tweets . . . . .	1708
27.3	BOTUS reporting on its trades (part 1) . . . . .	1709
27.4	BOTUS reporting on its trades (part 2) . . . . .	1710
27.5	A poutine (on the left); something else (on the right) . . . . .	1712
27.6	Syntactic parsing of a sentence using the Stanford parser . . . . .	1713
27.7	Abridged syntactic parsing of a sentence using the Enju English parser . . . . .	1716
27.8	TDM/DTM for a hypothetical corpus . . . . .	1720
27.9	Text visualizations (examples) . . . . .	1723
27.10	Text mining and NLP pipeline . . . . .	1724
28.1	Different types of histograms . . . . .	1836
28.2	Maintaining a histogram: updating layer 1 . . . . .	1837
28.3	Spam-filtering decision tree . . . . .	1846
28.4	The Hoeffding tree algorithm . . . . .	1847
28.5	A review of $k$ NN classification . . . . .	1849
28.6	Frequent datastream pattern mining FDPM-1 . . . . .	1854
28.7	Illustration of a sliding window and transactions . . . . .	1855

## List of Tables

25.1	Deductive vs. inductive syllogisms . . . . .	1594
26.1	Confusion matrix for an anomaly detection problem . . . . .	1652
26.2	Metric values for various supervised anomaly detection models . . . . .	1653
27.1	Penn treebank tagset (part 1) . . . . .	1717
27.2	Penn treebank tagset (part 2) . . . . .	1718
27.3	The 37 universal syntactic relations used in Universal Dependencies v2 . . . . .	1718
27.4	Universal dependency relations, alphabetical listing . . . . .	1719
28.1	Layer 1 histogram for a data stream . . . . .	1837
28.2	Equal-width layer 2 histogram for a data stream . . . . .	1837
28.3	Equal-frequency layer 2 histogram for a data stream . . . . .	1838
28.4	Possible itemsets in the supermarket example . . . . .	1850
28.5	Transactions in the supermarket example . . . . .	1851
28.6	Popular frequent itemset mining data streams algorithm . . . . .	1853