
Bibliography

1. C. Aggarwal. Data mining: The textbook. *Springer*, 2015.
2. C. Aggarwal. Machine learning for text. *Springer*, 2018.
3. C. Aggarwal. Recommender systems: The textbook. *Springer*, 2016.
4. C. Aggarwal. Outlier analysis. *Springer*, 2017.
5. C. C. Aggarwal and S. Sathe. Outlier Ensembles: An Introduction. *Springer*, 2017.
6. C. Aggarwal. Neural networks and deep learning: A textbook. *Springer*, 2018.
7. C. Aggarwal. On the effects of dimensionality reduction on high dimensional similarity search. *ACM PODS Conference*, pp. 256–266, 2001.
8. R. Ahuja, T. Magnanti, and J. Orlin. Network flows: theory, algorithms, and applications. *Prentice Hall*, 1993.
9. A. Azran. The rendezvous algorithm: Multiclass semi-supervised learning with markov random walks. *ICML*, pp. 49–56, 2007.
10. M. Bazaraa, H. Sherali, and C. Shetty. Nonlinear programming: theory and algorithms. *John Wiley and Sons*, 2013.
11. I. Bayer. Fastfm: a library for factorization machines. *arXiv preprint arXiv:1505.00641*, 2015. <https://arxiv.org/pdf/1505.00641v2.pdf>
12. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1), pp. 183–202, 2009.
13. S. Becker, and Y. LeCun. Improving the convergence of back-propagation learning with second order methods. *Proceedings of the 1988 connectionist models summer school*, pp. 29–37, 1988.
14. J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, pp. 281–305, 2012.
15. D. Bertsekas. Nonlinear programming. *Athena scientific*, 1999.
16. D. Bertsimas and J. Tsitsiklis. Introduction to linear optimization. *Athena Scientific*, 1997.
17. S. Bhagat, G. Cormode, and S. Muthukrishnan. Node classification in social networks. *Social Network Data Analytics*, Springer, pp. 115–148. 2011.

18. C. M. Bishop. Pattern recognition and machine learning. *Springer*, 2007.
19. C. M. Bishop. Neural networks for pattern recognition. *Oxford University Press*, 1995.
20. E. Bodewig. Matrix calculus. *Elsevier*, 2014.
21. P. Boggs and J. Tolle. Sequential quadratic programming. *Acta Numerica*, 4, pp. 1–151, 1995.
22. S. Boyd and L. Vandenberghe. Convex optimization. *Cambridge University Press*, 2004.
23. S. Boyd and L. Vandenberghe. Applied linear algebra. *Cambridge University Press*, 2018.
24. S. Brin, and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1–7), pp. 107–117, 1998.
25. A. Brouwer and W. Haemers. Spectra of graphs. *Springer Science and Business Media*, 2011.
26. A. Bryson. A gradient method for optimizing multi-stage allocation processes. *Harvard University Symposium on Digital Computers and their Applications*, 1961.
27. C. Chang and C. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 27, 2011.
<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
28. O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5), pp. 1155–1178, 2007.
29. F. Chung. Spectral graph theory. *American Mathematical Society*, 1997.
30. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3), pp. 273–297, 1995.
31. N. Cristianini, and J. Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. *Cambridge University Press*, 2000.
32. Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *NIPS Conference*, pp. 2933–2941, 2014.
33. S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 41(6), pp. 391–407, 1990.
34. C. Deng. A generalization of the Sherman Morrison Woodbury formula. *Applied Mathematics Letters*, 24(9), pp. 1561–1564, 2011.
35. C. Ding, T. Li, and W. Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics and Data Analysis*, 52(8), pp. 3913–3927, 2008.
36. N. Draper and H. Smith. Applied regression analysis. *John Wiley & Sons*, 2014.
37. D. Du and P. Pardalos (Eds). Minimax and applications, *Springer*, 2013.
38. J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, pp. 2121–2159, 2011.
39. R. Duda, P. Hart, and D. Stork. Pattern classification. *John Wiley and Sons*, 2012.
40. D. Easley, and J. Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. *Cambridge University Press*, 2010.
41. C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), pp. 211–218, 1936.

42. A. Emmott, S. Das, T. Dietterich, A. Fern, and W. Wong. Systematic Construction of Anomaly Detection Benchmarks from Real Data. *arXiv:1503.01158*, 2015.
<https://arxiv.org/abs/1503.01158>
43. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM Computer Communication Review*, pp. 251–262, 1999.
44. R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9, pp. 1871–1874, 2008.
<http://www.csie.ntu.edu.tw/~cjlin/liblinear/>
45. R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7: pp. 179–188, 1936.
46. P. Flach. Machine learning: the art and science of algorithms that make sense of data. *Cambridge University Press*, 2012.
47. C. Freudenthaler, L. Schmidt-Thieme, and S. Rendle. Factorization machines: Factorized polynomial regression models. *GPSDAA*, 2011.
48. J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), pp. 432–441, 2008.
49. M. Garey, and D. S. Johnson. Computers and intractability: A guide to the theory of NP-completeness. *New York, Freeman*, 1979.
50. E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. *ACM SIGIR Conference*, pp. 601–602, 2005.
51. H. Gavin. The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems, 2011.
<http://people.duke.edu/~hpgavin/ce281/lm.pdf>
52. G. Golub and C. F. Van Loan. Matrix computations, *John Hopkins University Press*, 2012.
53. I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. *MIT Press*, 2016.
54. I. Goodfellow, O. Vinyals, and A. Saxe. Qualitatively characterizing neural network optimization problems. *arXiv:1412.6544*, 2014. [Also appears in *ICLR*, 2015]
<https://arxiv.org/abs/1412.6544>
55. A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. *ACM KDD Conference*, pp. 855–864, 2016.
56. T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. *Springer*, 2009.
57. T. Hastie, R. Tibshirani, and M. Wainwright. Statistical learning with sparsity: the lasso and generalizations. *CRC Press*, 2015.
58. K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.
59. M. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6), 1952.
60. G. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40(1–3), pp. 185–234, 1989.
61. G. Hinton. Neural networks for machine learning, *Coursera Video*, 2012.
62. K. Hoffman and R. Kunze. Linear algebra, Second Edition, *Pearson*, 1975.

63. T. Hofmann. Probabilistic latent semantic indexing. *ACM SIGIR Conference*, pp. 50–57, 1999.
64. C. Hsieh, K. Chang, C. Lin, S. S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. *ICML*, pp. 408–415, 2008.
65. Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. *IEEE ICDM*, pp. 263–272, 2008.
66. H. Yu and B. Wilamowski. Levenberg–Marquardt training. *Industrial Electronics Handbook*, 5(12), 1, 2011.
67. R. Jacobs. Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1(4), pp. 295–307, 1988.
68. T. Jaakkola, and D. Haussler. Probabilistic kernel regression models. *AISTATS*, 1999.
69. P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. *ACM Symposium on Theory of Computing*, pp. 665–674, 2013.
70. C. Johnson. Logistic matrix factorization for implicit feedback data. *NIPS Conference*, 2014.
71. H. J. Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10), pp. 947–954, 1960.
72. D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. <https://arxiv.org/abs/1412.6980>
73. M. Knapp. Sines and cosines of angles in arithmetic progression. *Mathematics Magazine*, 82(5), 2009.
74. D. Koller and N. Friedman. Probabilistic graphical models: principles and techniques. *MIT Press*, 2009.
75. Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 8, pp. 30–37, 2009.
76. A. Langville, C. Meyer, R. Albright, J. Cox, and D. Duling. Initializations for the nonnegative matrix factorization. *ACM KDD Conference*, pp. 23–26, 2006.
77. D. Lay, S. Lay, and J. McDonald. Linear Algebra and its applications, *Pearson*, 2012.
78. Q. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Ng, On optimization methods for deep learning. *ICML Conference*, pp. 265–272, 2011.
79. D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.
80. C. J. Lin, R. C. Weng, and S. S. Keerthi. Trust region newton method for logistic regression. , 9(Apr), 627–650. *Journal of Machine Learning Research*, 9, pp. 627–650, 2008.
81. T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3), pp. 225–231, 2009.
82. B. London and L. Getoor. Collective classification of network data. *Data Classification: Algorithms and Applications*, CRC Press, pp. 399–416, 2014.
83. D. Luenberger and Y. Ye. Linear and nonlinear programming, *Addison-Wesley*, 1984.
84. U. von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4), pp. 395–416, 2007.
85. S. Marsland. Machine learning: An algorithmic perspective, *CRC Press*, 2015.
86. J. Martens. Deep learning via Hessian-free optimization. *ICML Conference*, pp. 735–742, 2010.

87. J. Martens and I. Sutskever. Learning recurrent neural networks with hessian-free optimization. *ICML Conference*, pp. 1033–1040, 2011.
88. J. Martens, I. Sutskever, and K. Swersky. Estimating the hessian by back-propagating curvature. *arXiv:1206.6464*, 2016.
<https://arxiv.org/abs/1206.6464>
89. J. Martens and R. Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. *ICML Conference*, 2015.
90. P. McCullagh. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pp. 109–142, 1980.
91. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.
<https://arxiv.org/abs/1301.3781>
92. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *NIPS Conference*, pp. 3111–3119, 2013.
93. T. Minka. A comparison of numerical optimizers for logistic regression. *Unpublished Draft*, 2003.
94. T. Mitchell. Machine learning, *McGraw Hill*, 1997.
95. K. Murphy. Machine learning: A probabilistic perspective, *MIT Press*, 2012.
96. G. Nemhauser, A. Kan, and N. Todd. Nondifferentiable optimization. *Handbooks in Operations Research and Management Sciences*, 1, pp. 529–572, 1989.
97. Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27, pp. 372–376, 1983.
98. A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *NIPS Conference*, pp. 849–856, 2002.
99. J. Nocedal and S. Wright. Numerical optimization. *Springer*, 2006.
100. N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3), pp. 127–239, 2014.
101. J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. *EMNLP*, pp. 1532–1543, 2014.
102. J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Method: Support Vector Learning*, MIT Press, pp. 85–208, 1998.
103. B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. *ACM KDD Conference*, pp. 701–710, 2014.
104. E. Polak. Computational methods in optimization: a unified approach. *Academic Press*, 1971.
105. B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4), pp. 838–855, 1992.
106. N. Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1), pp. 145–151, 1999.
107. S. Rendle. Factorization machines. *IEEE ICDM Conference*, pp. 995–100, 2010.
108. S. Rendle. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology*, 3(3), 57, 2012.

109. F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386, 1958.
110. D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by back-propagating errors. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pp. 318–362, 1986.
111. T. Schaul, S. Zhang, and Y. LeCun. No more pesky learning rates. *ICML Conference*, pp. 343–351, 2013.
112. B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5), pp. 1299–1319, 1998.
113. B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7), pp. 1443–1472, 2001.
114. J. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. *Technical Report, CMU-CS-94-125*, Carnegie-Mellon University, 1994.
115. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), pp. 888–905, 2000.
116. N. Shor. Minimization methods for non-differentiable functions (Vol. 3). *Springer Science and Business Media*, 2012.
117. A. Singh and G. Gordon. A unified view of matrix factorization models. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 358–373, 2008.
118. B. Schölkopf and A. J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. *Cambridge University Press*, 2001.
119. J. Solomon. Numerical Algorithms: Methods for Computer Vision, Machine Learning, and Graphics. *CRC Press*, 2015.
120. N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, pp. 1329–1336, 2004.
121. G. Strang. The discrete cosine transform. *SIAM review*, 41(1), pp. 135–147, 1999.
122. G. Strang. An introduction to linear algebra, Fifth Edition. *Wellesley-Cambridge Press*, 2016.
123. G. Strang. Linear algebra and its applications, Fourth Edition. *Brooks Cole*, 2011.
124. G. Strang and K. Borre. Linear algebra, geodesy, and GPS. *Wellesley-Cambridge Press*, 1997.
125. G. Strang. Linear algebra and learning from data. *Wellesley-Cambridge Press*, 2019.
126. J. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (5500), pp. 2319–2323, 2000.
127. A. Tikhonov and V. Arsenin. Solution of ill-posed problems. *Winston and Sons*, 1977.
128. M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1), pp. 1–118, 2016.
<https://github.com/madeleineudell/LowRankModels.jl>
129. G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. *Advances in Kernel Methods-Support Vector Learning*, 6, pp. 69–87, 1999.
130. H. Wendland. Numerical linear algebra: An introduction. *Cambridge University Press*, 2018.
131. P. Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. *PhD thesis, Harvard University*, 1974.

132. B. Widrow and M. Hoff. Adaptive switching circuits. *IRE WESCON Convention Record*, 4(1), pp. 96–104, 1960.

133. C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. *NIPS Conference*, 2000.

134. S. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1), pp. 3–34, 2015.

135. T. T. Wu, and K. Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1), pp. 224–244, 2008.

136. H. Yu, F. Huang, and C. J. Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1–2), pp. 41–75, 2011.

137. H. Yu, C. Hsieh, S. Si, and I. S. Dhillon. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. *IEEE ICDM*, pp. 765–774, 2012.

138. R. Zafarani, M. A. Abbasi, and H. Liu. Social media mining: an introduction. *Cambridge University Press*, 2014.

139. M. Zeiler. ADADELTA: an adaptive learning rate method. *arXiv:1212.5701*, 2012. <https://arxiv.org/abs/1212.5701>

140. T. Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46, 1–3, pp. 81–129, 2002.

141. Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the Netflix prize. *Algorithmic Aspects in Information and Management*, pp. 337–348, 2008.

142. J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Advances in neural information processing systems*, 2002.

143. X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. *ICML Conference*, pp. 912–919, 2003.

144. <https://www.csie.ntu.edu.tw/~cjlin/libmf/>