

The Mathematical Engineering of Deep Learning

Book Draft

Benoit Liquet, Sarat Moka and Yoni Nazarathy

June 15, 2022

Bibliography

- [1] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [3] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
- [4] Jeremy Howard and Sylvain Gugger. *Deep Learning for Coders with fastai and PyTorch*. O’Reilly Media, 2020.
- [5] Luciano Ramalho. *Fluent Python*. O’Reilly Media, Incorporated, 2021.
- [6] Dirk P Kroese, Zdravko Botev, Thomas Taimre, and Radislav Vaisman. *Data science and machine learning: Mathematical and statistical methods*. CRC Press, 2019.
- [7] Yoni Nazarathy and Hayden Klok. *Statistics with Julia*. Springer, 2021.
- [8] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

- [9] Pierre Lafaye de Micheaux, Rémy Drouilhet, and Benoit Liquet. *The R software: Fundamentals of programming and statistical analysis*. Springer, 2013.
- [10] JJ Allaire. *Deep Learning with R*. Simon and Schuster, 2018.
- [11] Abhijit S Pandya and Robert B Macy. *Pattern recognition with neural networks in C++*. CRC press, 1995.
- [12] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O’Reilly Media, 2019.
- [13] Joe Papa. *PyTorch Pocket Reference: Building and Deploying Deep Learning Model*. O’Reilly Media, 2021.
- [14] Andrew Ng. Machine learning yearning. URL: [http://www.mlyearning.org/\(96\)](http://www.mlyearning.org/(96)), 139, 2017.
- [15] Neil Savage. How ai and neuroscience drive each other forwards. *Nature*, 571(7766):S15–S15, 2019.
- [16] Demis Hassabis, Dhharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.
- [17] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural Networks*, 111:47–63, 2019.
- [18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [19] Charu C Aggarwal et al. *Neural networks and deep learning*, volume 10. Springer, 2018.
- [20] Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, 2015.
- [21] Ovidiu Calin. *Deep learning architectures*. Springer, 2020.

- [22] Alan M Turing and J Haugeland. *Computing machinery and intelligence*. MIT Press Cambridge, MA, 1950.
- [23] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [24] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [26] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.
- [27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [28] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.
- [29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [30] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- [31] Demis Hassabis. Artificial intelligence: Chess match of the century. *Nature*, 544(7651):413–414, 2017.

- [32] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- [33] Stephen Boyd and Lieven Vandenberghe. *Introduction to applied linear algebra: vectors, matrices, and least squares*. Cambridge university press, 2018.
- [34] Gilbert Strang. *Linear algebra and learning from data*. Wellesley-Cambridge Press Cambridge, 2019.
- [35] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.
- [36] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *arXiv preprint arXiv:2007.06823*, 2020.
- [37] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.
- [38] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.
- [39] Siddhant Garg and Goutham Ramakrishnan. Advances in quantum deep learning: An overview. *arXiv preprint arXiv:2005.04316*, 2020.
- [40] David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.
- [41] Christopher M Bishop. *Pattern recognition*, volume 128. 2006.
- [42] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

- [43] Andriy Burkov. *The hundred-page machine learning book*, volume 1. Andriy Burkov Quebec City, QC, Canada, 2019.
- [44] Andreas Lindholm, Niklas Wahlström, Fredrik Lindsten, and Thomas B. Schön. *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022.
- [45] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- [46] Douglas C Montgomery. *Design and analysis of experiments*. John wiley & sons, 2017.
- [47] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.
- [48] James O Berger and Robert L Wolpert. The likelihood principle: A review, generalizations, and statistical implications. *Lecture Notes—Monograph Series*, 6, 1988.
- [49] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.
- [50] Yuhong Yang. Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950, 2005.
- [51] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.
- [52] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, 2016.
- [53] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

- [54] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.
- [55] Jeffrey S Simonoff. *Smoothing methods in statistics*. Springer Science & Business Media, 2012.
- [56] Frank Emmert-Streib, Salisou Moutari, and Matthias Dehmer. A comprehensive survey of error measures for evaluating binary decision making in data science. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(5):e1303, 2019.
- [57] David J Hand. Assessing the performance of classification methods. *International Statistical Review*, 80(3):400–414, 2012.
- [58] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.
- [59] Nancy A Obuchowski and Jennifer A Bullen. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol*, 63(7):07TR01, mar 2018.
- [60] C Van Rijsbergen. Information retrieval (book 2nd ed), 1979.
- [61] Vladimir N Vapnick. *Statistical learning theory*. Wiley, New York, 1998.
- [62] Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 2012.
- [63] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [64] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [65] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

- [66] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- [67] Gauss, c. f. (1809). *theoria motus corporum coelestium*. perthes, hamburg. translation reprinted as *theory of the motions of the heavenly bodies moving about the sun in conic sections*. dover, new york, 1963.
- [68] Stephen M Stigler. Gauss and the invention of least squares. *the Annals of Statistics*, pages 465–474, 1981.
- [69] Eliakim H Moore. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.*, 26:394–395, 1920.
- [70] Arne Bjerhammar. *Application of calculus of matrices to method of least squares: with special reference to geodetic calculations*. Elander, 1951.
- [71] Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge University Press, 1955.
- [72] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
- [73] Gene Howard Golub. Least squares, singular values and matrix approximations. *Aplikace matematiky*, 13(1):44–51, 1968.
- [74] L DO Q. Numerically efficient methods for solving least squares problems. *Pennsylvania State University*, 2012.
- [75] Edward R Mansfield and Billy P Helms. Detecting multicollinearity. *The American Statistician*, 36(3a):158–160, 1982.
- [76] Subhash C Narula and John F Wellington. The minimum sum of absolute errors regression: A state of the art survey. *International Statistical Review/Revue Internationale de Statistique*, pages 317–326, 1982.
- [77] Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, pages 799–821, 1973.

- [78] Melissa A Hardy. *Regression with dummy variables*, volume 93. Sage, 1993.
- [79] Claude Lemaréchal. Cauchy and the gradient method. *Doc Math Extra*, 251(254):10, 2012.
- [80] Katherine Mary Malan. A survey of advances in landscape analysis for optimisation. *Algorithms*, 14(2):40, 2021.
- [81] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.
- [82] Petre Stoica and Yngve Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.
- [83] Gerda Claeskens, Nils Lid Hjort, et al. Model selection and model averaging. *Cambridge Books*, 2008.
- [84] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [85] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143:143, 2015.
- [86] Andrey Nikolayevich Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.
- [87] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.
- [88] Quan Zhou, Wenlin Chen, Shiji Song, Jacob Gardner, Kilian Weinberger, and Yixin Chen. A reduction of the elastic net to support vector machines with an application to gpu computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

- [89] Tzu-Tsung Wong and Po-Yang Yeh. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1586–1594, 2019.
- [90] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745, 2009.
- [91] Yoonsuh Jung. Multiple predicting k-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, 30(1):197–215, 2018.
- [92] Leo Breiman, JH Friedman, RA Olshen, and CJ Stone. Classification and regression trees. wadsworth & brooks. *Cole Statistics/Probability Series*, 1984.
- [93] Xinchuan Zeng and Tony R Martinez. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12, 2000.
- [94] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.
- [95] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.
- [96] Absalom E Ezugwu, Abiodun M Ikotun, Olaide O Oyelade, Laith Abualigah, Jeffery O Agushaka, Christopher I Eke, and Andronicus A Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.
- [97] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.
- [98] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

- [99] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.
- [100] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.
- [101] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.
- [102] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.
- [103] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Linear algebra*, pages 134–151. Springer, 1971.
- [104] Alan Kaylor Cline and Inderjit S Dhillon. Computation of the singular value decomposition. 2006.
- [105] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
- [106] Chester I Bliss. The method of probits. *Science*, 79(2037):38–39, 1934.
- [107] Jan Salomon Cramer. The origins of logistic regression. 2002.
- [108] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [109] Albert B Novikoff. On convergence proofs for perceptrons. Technical report, STANFORD RESEARCH INST MENLO PARK CA, 1963.
- [110] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.
- [111] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

- [112] Ralitzza Gueorguieva, Robert Rosenheck, and Daniel Zelterman. Dirichlet component regression and its applications to psychiatric data. *Computational statistics & data analysis*, 52(12):5344–5355, 2008.
- [113] Rafiq H Hijazi and Robert W Jernigan. Modelling compositional data using dirichlet regression models. *Journal of Applied Probability & Statistics*, 4(1):77–91, 2009.
- [114] Jacob C Douma and James T Weedon. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, 10(9):1412–1430, 2019.
- [115] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2003.
- [116] Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.
- [117] Joseph M Hilbe. *Logistic regression models*. Chapman and hall/CRC, 2009.
- [118] Amnon Shashua and Anat Levin. Ranking with large margin principle: Two approaches. *Advances in neural information processing systems*, 15, 2002.
- [119] Daniel Commenges and Helene Jacqmin-Gadda. *Dynamical biostatistical models*, volume 86. CRC Press, 2015.
- [120] David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. CRC Press, 1979.
- [121] Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200, 1992.
- [122] David Charte, Francisco Charte, Salvador García, María J del Jesus, and Francisco Herrera. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*, 44:78–96, 2018.
- [123] Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*, 2018.

- [124] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988.
- [125] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.
- [126] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [127] Mykel J Kochenderfer and Tim A Wheeler. *Algorithms for optimization*. Mit Press, 2019.
- [128] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.
- [129] Rangarajan K Sundaram et al. *A first course in optimization theory*. Cambridge university press, 1996.
- [130] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [131] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- [132] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.
- [133] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization?, 2019.
- [134] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.
- [135] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic

- segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [136] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [137] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.
- [138] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.
- [139] Henry W Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017.
- [140] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.
- [141] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Mądry. How does batch normalization help optimization? In *Proceedings of the 32nd international conference on neural information processing systems*, pages 2488–2498, 2018.
- [142] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.
- [143] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2019.
- [144] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical

- report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
- [145] Daniel A Roberts, Sho Yaida, and Boris Hanin. The principles of deep learning theory. *arXiv preprint arXiv:2106.10165*, 2021.
- [146] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.
- [147] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [148] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- [149] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [150] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [151] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [152] Mingxing Tan and Quoc V Le. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*, 2021.
- [153] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

- [154] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruysen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. The visual task adaptation benchmark. 2019.
- [155] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018.
- [156] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [157] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
- [158] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [159] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [160] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.
- [161] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [162] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio.

- Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [163] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [164] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
- [165] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *International conference on machine learning*, pages 2342–2350. PMLR, 2015.
- [166] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [167] Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7):e0180944, 2017.
- [168] Zhen Shen, Wenzheng Bao, and De-Shuang Huang. Recurrent neural network for predicting transcription factor binding sites. *Scientific reports*, 8(1):1–10, 2018.
- [169] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, et al. The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv:1804.09849*, 2018.
- [170] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [171] Andrea Galassi, Marco Lippi, and Paolo Torrioni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

- [172] Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*, 2020.
- [173] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [174] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [175] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pre-training for language understanding. *Advances in neural information processing systems*, 32, 2019.
- [176] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [177] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [178] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.
- [179] James Bergstra and Yoshua Bengio. Random search for hyperparameter optimization. *Journal of machine learning research*, 13(2), 2012.
- [180] Genevieve B Orr and Klaus-Robert Müller. *Neural networks: tricks of the trade*. Springer, 2003.
- [181] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

- [182] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [183] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [184] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.
- [185] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [186] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [187] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [188] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [189] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.
- [190] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.
- [191] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The odds are odd: A statistical test for detecting adversarial examples. In *International Conference on Machine Learning*, pages 5498–5507. PMLR, 2019.

- [192] Matthew Botvinick, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 23(5):408–422, 2019.
- [193] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.
- [194] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019.
- [195] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [196] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. *arXiv preprint arXiv:1811.12560*, 2018.
- [197] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [198] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.
- [199] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.
- [200] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.
- [201] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

- [202] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pages 1407–1416. PMLR, 2018.
- [203] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Index

- F_β , 51
- k -times continuously differentiable, 391, 392
- 2 by 2 confusion matrix, 47
- 80-20 splitting strategy, 47

- absolute error loss, 57
- accuracy, 48
- acquisition function, 327
- activation function, 112
- Activation functions, 323
- active learning, 39
- Adadelta, 177
- Adagrad, 172
- ADAM, 174
- AdaMax, 178
- adaptive moment estimation, 174
- adaptive subgradient, 172
- adversarial autoencoders, 149
- Adversarial machine learning, 340
- Adversarial Training, 335
- affine discrete time linear dynamical system, 67
- agent, 38
- Akaike information criterion (AIC), 40
- AlexNet, 13, 19, 277
- algorithms, 36
- annotated, 21
- arctan, 219
- Area Under the Curve (AUC), 49
- Armijo condition, 182

- artificial general intelligence (AGI), 17
- artificial intelligence, 2
- artificial intelligence (AI), 3
- artificial neural network, 1
- artificial neuron, 112
- attention, 13
- auto-encoders, 14
- autoencoder, 134
- automatic differentiation, 192
- average-pooling, 270

- back propagation algorithm, 12, 221
- backward mode, 192
- Bagging, 241
- balanced, 25
- batch gradient descent, 160
- batch-normalization, 12
- Bayesian information criterion (BIC), 40
- Bayesian neural networks, 34
- Bayesian Optimisation, 324
- bias, 52, 103
- bias and variance tradeoff, 77
- bias vector, 204
- bias-variance tradeoff, 70
- bias-variance-noise decomposition equation, 77
- big data (analytics), 3
- binary cross entropy, 107
- binomial distribution, 102

- bottleneck, 134
- categorical, 22
- categorical cross entropy, 123
- categorical distribution, 118
- Cauchy-Schwartz, 384
- causal modeling, 34
- centered data matrix, 88
- centers, 85
- central processing units (CPUs), 18
- centroids, 85
- channels, 13
- CIFAR-10 dataset, 25
- classification, 4, 37, 42
- cloud computing, 19
- clustering, 37, 84
- code, 134
- column-major, 22
- composability, 7
- computational graph, 193
- computed features, 14
- computer scientists, 9
- confidence bands, 43
- confusion matrix, 63
- consistent, 73
- constrained, 152
- constrained optimization, 140
- constraint set, 152
- continuous, 385
- continuously differentiable, 390
- contractive autoencoders, 149
- control theory, 10
- converge, 385
- convex function, 155
- convex set, 154
- convexity, 154
- convolution, 250, 251
- convolutional kernel, 261
- Convolutional layers, 248
- convolutional neural network architecture, 12
- Convolutional neural networks, 247
- correlated features, 89
- covariance, 89
- cross entropy, 396
- cross validation, 81
- curse of dimensionality, 15
- data, 36
- data matrix, 88
- data mining, 3
- data reduction, 37
- data science, 3
- data scientists, 9
- de-mean, 88
- de-noising, 144
- decaying step factor, 180
- decision rule, 46
- decision trees, 42, 52
- decoder, 134
- decoders, 14
- deep, 113
- deep fakes, 14
- Deep feedforward networks, 203
- deep learning (DL), 3
- deep reinforcement learning (Deep RL), 15
- deep variational inference, 34
- degree, 130
- denoising autoencoder, 146
- derivative, 390
- descent direction method, 159
- descent step, 160
- design matrix, 53
- differentiable, 390
- dimension, 22
- directional derivative, 387
- Dirichlet regression, 149
- discriminative models, 51
- discriminator, 14, 341

- Dropout, 239, 240
- dropout, 12
- dummy, 59
- dying Relu, 218

- Early Stopping, 169, 321
- Eckart–Young–Mirsky theorem, 95
- elastic net, 79
- elbow, 87
- elu, 219
- encoder, 134
- encoders, 14
- engineered feature, 45
- epoch, 168
- error, 43, 54
- ethical aspects, 34
- Euclidean distance, 384
- Euclidean norm, 384
- expected generalization performance, 74
- Expected Improvement (EI), 327
- exploding , 230
- Exploding gradients, 230
- expression swell, 192

- F_1 score, 50
- False Negative, 47
- False Positive, 47
- false positive rate, 49
- fashion MNIST dataset, 25
- fast.ai, 6
- fastai, 33
- feature engineering, 213
- feature extraction, 14
- Feature extractor, 330, 332
- feature vectors, 21
- feedforward neural networks, 203
- feedforward pass, 204, 207
- filtering, 248
- fine-tune, 329
- Fine-Tuning, 330
- Fine-Tuning and/or Freeze, 331
- finite sum problem, 156
- First hidden layer, 323
- first Wolfe condition, 182
- Flux.jl, 33
- forward mode, 192
- Fruits 360, 6
- full SVD, 56, 93
- fully connected, 271
- fully connected feedforward neural network, 10
- fully connected layers, 12

- game theory, 15
- Gaussian Process (GP), 325
- general fully connected neural network, 201
- generalization ability, 70
- generalization error, 72
- generalization gap, 74
- generalization performance, 72
- generalize, 78
- Generalized Additive Model, 45
- generalized linear model, 103
- Generalized Linear Models, 45
- Generative adversarial network (GAN), 14
- Generative Adversarial Networks (GAN), 335
- Generative modelling, 340
- generative modelling, 38
- generative models, 51
- generator, 14, 341
- geometric transformations, 335
- global minimum, 153
- GPT-3, 17
- gradient, 386
- gradient boosted trees, 52
- gradient descent, 63
- Gram matrix, 56

- graphical processing units (GPUs),
18
- Grid Search, 324
- Hessian, 389
- Hessian matrix, 114
- hidden layer, 134
- high dimensional, 19
- hold out set, 39
- Huber error loss, 57
- hyper-parameter, 79
- Hyper-parameters, 320
- hyperbolic tangent, 217
- hypothesis tests, 43

- i.i.d, 105
- identifiable, 118
- identity activation function, 112
- image processing, 12
- ImageNet, 277, 329
- ImageNet challenge, 26
- ImageNet database, 4, 25
- imitation game, 16
- impulse response, 252
- indicator, 59
- information radius, 348
- information theory, 10
- inherent noise, 77
- inner product, 384
- input layer, 202
- interaction terms, 45
- intercept, 52, 103
- interpolation on the latent space,
147
- interpretable models, 109
- interpretation, 43, 108

- Jacobian, 388
- Jensen-Shannon distance, 397
- Jensen-Shannon divergence, 397
- Julia, 33

- K-fold cross validation, 81
- k-fold cross validation, 40, 324
- K-means, 84
- k-nearest neighbours, 52
- Keras, 6
- KL divergence, 395
- knee point, 87
- Kullback–Leibler divergence, 395

- Label Classifier component, 332
- Label Domain Classifier, 332
- labelling, 85
- labels, 21
- language models, 17
- LASSO, 80
- latent space, 144, 341
- latent variable, 104
- leaky ReLU, 218, 323
- learning, 36
- Learning rate, 320
- learning rate, 64, 161
- learning to rank, 149
- least absolute shrinkage and selec-
tion operator, 80
- least squares problem, 54
- likelihood, 40
- likelihood function, 104
- linear approximation, 391, 393
- linear autoencoder, 139
- linear classifiers, 111
- linear discriminant analysis (LDA),
51
- linear model, 10
- Linear Time Invariant, 251
- linearly separable, 149
- link function, 103
- loading vector, 90
- local minimum, 153
- localization, 7
- locally convex, 155

- Locally Estimated Scatterplot Smoothing (LOESS), 45
- log odds, 102
- log-likelihood, 105
- log-sum-exp, 158
- logistic, 216
- logistic distribution, 104
- logistic function, 103
- logistic regression, 11, 46
- logit, 102
- loss function, 42, 54
- loss landscape, 68
- low rank approximation, 94
- LSTM long short term memory, 13

- machine learning, 2
- machine learning (ML), 3
- machine learning engineers, 9
- manual annotation process, 23
- Maple, 33
- Mathematica, 33
- mathematical engineering, 10
- MATLAB, 33
- max pooling, 13
- max-pooling, 270
- maximum a posteriori probability, 126
- maximum apostriori probability, 204
- maximum likelihood estimation, 104
- maximum likelihood estimation method, 57
- mean computation, 85
- mean square error, 41
- Mini-Batch Size:, 320
- mini-batches, 167
- mixed models, 45
- MLPs, 202
- MNIST database, 25
- model bias, 70, 78
- model misspecification, 45
- model parameters, 43
- model selection, 70
- model specific hyper-parameters, 320
- model variance, 70, 78
- models, 36
- momentum, 170, 171
- monomial, 130
- Moore-Penrose pseudo-inverse, 56
- multi-class classification, 42
- multi-class classification problem, 60
- multi-class logistic regression, 117
- multi-index, 392
- multi-layer perceptrons, 12
- multilayer network, 202
- multilayer perceptrons, 202, 203
- multinomial distribution, 118
- multinomial logistic regression, 117
- multinomial regression model, 117
- multinomial softmax regression, 11
- multivariate chain rule, 390

- Nadam, 176
- Nadaraya–Watson kernel regression, 45
- naive Bayes classifier, 51
- narrow tasks, 18
- natural language processing (NLP), 13, 26
- negative definite, 389
- negative predictive value, 48
- negative semidefinite, 389
- Nesterov momentum, 176
- neural network, 1
- noise features, 89
- nominal categorical variables, 22
- non-interprettable, 109
- non-linear PCA, 143

- normal equations, 55
- Number of Epochs:, 321
- Number of hidden units, 323
- Number of layers, 323
- numerical, 22
- numerical differentiation, 192

- objective function, 152
- odds, 102
- odds ratio, 108
- one vs. all, 60
- one vs. one, 60
- one vs. rest, 60
- one-hot encoding, 58
- optimizer hyper-parameters, 320
- ordinal categorical variable, 22
- ordinal regression, 149
- over-fitting, 70
- overcomplete, 149
- overfit, 78

- Padding, 263
- parameter estimates, 43
- partial derivative, 386
- peaks function, 154
- perceptron, 12, 149
- performance function, 71
- performance metrics, 71
- piece wise affine function, 219
- pooling layers, 270
- positive (semi) definiteness, 389
- positive definite, 389
- positive predictive value, 48
- positive semidefinite, 389
- power product, 130
- precision, 48
- prediction, 37
- PReLU, 218
- preprocessing, 40
- Principal Component Analysis, 84
- principal components, 89

- probability, 10
- probit regression model, 104
- pure mathematics, 10
- Python, 6, 33
- PyTorch, 6
- PyTorch Lightning, 33

- quadratic approximation, 391, 393
- quadratic loss, 54
- quantum deep learning, 34

- R statistical computing system, 33
- Random deletion, 335
- random forests, 52
- Random Insertion, 335
- Random Search, 324
- Random swap, 335
- ranking learning, 149
- real world, 36
- recall, 48
- Receiver Operating Characteristic (ROC) curve, 49
- rectified linear unit, 217
- recurrent neural network, 13
- recurrent neural networks, 203
- reduced SVD, 56, 93
- reference level, 59
- regression, 6, 37, 42
- regression parameter, 52, 103
- Regularization, 239, 242
- regularization, 70, 79
- regularization hyperparameter, 321
- regularization parameter, 79
- regularization term, 79
- Regularized tuning parameter:, 321
- reinforcement learning, 15, 38
- relative entropy, 395
- RELU, 232
- ReLU, 218, 230, 232
- ReLU, 323
- researchers, 10

- residual, 54
- ridge regression, 80
- RMSprop, 172
- robust autoencoders, 149
- Root Mean Square Propagation, 172
- row-major, 22

- saddle points, 155
- sample correlation, 69
- sample correlation matrix, 89
- sample covariance matrix, 88
- sample mean, 40
- sample standard deviation, 41
- sample variance, 40
- scientific machine learning, 19
- scree plot, 96
- seen data, 39, 70
- self-supervised learning, 39
- selu, 219
- semantic segmentation, 7
- semi-supervised learning, 39
- sensitivity, 48
- sentiment, 23
- shallow, 113
- shallow neural network, 112
- sigmoid, 204, 211, 216, 230, 323
- sigmoid function, 103
- simple linear regression model, 43
- single layer, 112
- Singular Value Decomposition, 93
- singular value decomposition, 56
- singular values, 93
- singular vectors, 93
- skip connections, 12
- Sobel filter, 248
- softmax, 204
- softmax activation function, 120
- softmax logistic regression, 117
- softmax regression, 117
- softplus, 219
- softsign, 219
- space transformations, 335
- sparse autoencoders, 149
- specificity, 48
- spectral decomposition, 94
- square loss, 54
- standardization of the data, 40
- standardized samples, 41
- statistical learning, 3
- statisticians, 9
- statistics, 3, 10
- step, 216
- Step decay, 320
- step size, 160
- Stirling's approximation, 131
- stochastic gradient descent, 164
- strictly convex, 155
- supervised learning, 21, 36
- support vector machines (SVM), 52
- surrogate model, 325
- swish, 219
- symbolic differentiation, 192
- Synonym Replacement, 335

- tanh, 217, 230, 323
- tasks, 6
- Taylor polynomial, 391
- Taylor's theorem, 391
- tensor processing units (TPUs), 19
- Tensorflow, 6
- termination condition, 64, 160, 168, 175
- test set, 24
- testing data, 39
- testing set, 39
- threshold, 204
- Tikhonov regularization, 80
- train set, 24
- train-validate split, 81

train-validate-test split, 81
trainable convolutions, 13
training, 3
training data, 39
training set, 39, 81
training time, 77
Transfer Learning, 328, 329
transfer learning, 7, 38
transformers, 13
triangle inequality, 385
True Negative, 47
True Positive, 47
tuple, 30
Turing test, 16

unbiased estimator, 73
unconstrained, 152
under-fitting, 70, 78
undercomplete, 149
unit, 139
univariate, 42
unseen data, 39, 70
unseen input data, 3
Unsupervised Domain Adaptation,
332
unsupervised learning, 36

validation set, 40, 81
vanilla gradient descent, 160
vanishing, 230
vanishing gradient, 230
variational autoencoders, 149
Vectors, 30
VGG19, 249
VGG19 model, 4
visual cortex, 16

weight decay, 321
weight initialization, 12
weight matrix, 204
weight vector, 52, 103

Wisconsin Breast Cancer Dataset,
46