Computational Learning Theory (Books)

[1] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999 (in press).

[2] M. Anthony and N. Biggs. *Computational Learning Theory*. Cambridge Tracts in Theoretical Computer Science (30). Cambridge University Press, 1992.

[3] Luc Devroye, Lázsló Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics: Stochastic Modelling and Applied Probability (31). Springer, 1996.

[4] M. J. Kearns and U. Vazirani. *Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, 1995.

[5] Balas K. Natarajan. *Machine Learning: A Theoretical Approach*. Morgan Kaufmann, San Mateo, CA, 1991.

[6] Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[7] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[8] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

[9] M. Vidyasagar. *A Theory of Learning and Generalization, with Applications to Neural Networks and Control Systems*. Springer, New York, 1997.

VC Theory

[10] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Inform. Proc. Lett.*, 24:377–380, April 1987.

[11] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[12] A. Ehrenfeucht, David Haussler, Michael J. Kearns, and Leslie G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989. First appeared in Proc. 1st Annu. Workshop on Comput. Learning Theory, 1988.

[13] David Haussler, N. Littlestone, and Manfred K. Warmuth. Predicting $\{0, 1\}$ functions on randomly drawn points. *Information and Computation*, 115(2):284–293, 1994. Was in FOCS88, COLT88, and Univ. of California at Santa Cruz TR UCSC-CRL-90-54.

[14] Balas K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[15] Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984.

[16] Vladimir N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probab. and its Applications*, 16(2):264–280, 1971.

VC-Dimension Bounds

[17] Eric B. Baum and David Haussler. What size net gives valid generalization? *Neural Computation*, 1(1):151–160, 1989.

[18] Yossi Erlich, Dan Chazan, Scott Petrack, and Avraham Levy. Lower bound on VC-dimension by local shattering. *Neural Computation*, 9(4):771–776, 1997.

[19] Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers. *Machine Learning*, 18(2/3):131–148, 1995.

[20] Marek Karpinksi and Angus J. Macintyre. Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and System Sciences*, 54:169–176, 1997.

[21] Pascal Koiran and Eduardo D. Sontag. Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54(1):190–198, February 1997.

[22] Wolfgang Maass. Neural nets with superlinear VC-dimension. *Neural Computation*, 6(5):877–884, 1994.

[23] A. Sakurai. Tighter bounds of the VC-dimension of three-layer networks. In *Proceedings of the 1993 World Congress on Neural Networks*, volume 3, pages 540–543, Hillsdale, NJ, 1993. Erlbaum.

[24] R. S. Wenocur and R. M. Dudley. Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33:313–318, 1981.

Learning Real-Valued Functions

[25] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):616–631, 1997.

[26] Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

[27] Peter L. Bartlett, Sanjeev R. Kulkarni, and S. Eli Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 43(5):1721–1724, 1997.

[28] Peter L. Bartlett, Philip M. Long, and Robert C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proc. 7th Annu. ACM Workshop on Comput. Learning Theory*, pages 299–310. ACM Press, New York, NY, 1994.

[29] Kevin L. Buescher and P. R. Kumar. Simultaneous learning of concepts and simultaneous estimation of probabilities. In *Proc. 4th Annu. Workshop on Comput. Learning Theory*, pages 33–42, San Mateo, CA, 1991. Morgan Kaufmann.

[30] R. M. Dudley. Universal Donsker classes and metric entropy. *Annals of Probability*, 15(4):1306–1326, 1987.

[31] Leonid Gurvits and Pascal Koiran. Approximation and learning of convex superpositions. In *Computational Learning Theory: Eurocolt '95*, pages 222–236. Springer-Verlag, 1995.

[32] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. Comput.*, 100(1):78–150, September 1992.

[33] Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proc. of the 31st Symposium on the Foundations of Comp. Sci.*, pages 382–391. IEEE Computer Society Press, Los Alamitos, CA, 1990.

[34] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

[35] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. In *Proc. 9th Annu. Conf. on Comput. Learning Theory*, pages 140–146. ACM Press, New York, NY, 1996.

[36] Balas K. Natarajan. Occam's razor for functions. In *Proc. 6th Annu. Workshop on Comput. Learning Theory*, pages 370–376. ACM Press, New York, NY, 1993.

[37] David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[38] Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *Siam J. Computing*, 26(3):751–763, 1997.

Large Margin Classification, Boosting/ARCing, Support Vector Machines

[39] Peter L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[40] Peter L. Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In Bernhard Schölkopf, Chris J. C. Burges, and Alex J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*. MIT Press, 1998. to appear.

[41] B. E. Boser, I. M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. 5th Annu. Workshop on Comput. Learning Theory*, pages 144–152. ACM Press, New York, NY, 1992.

[42] Leo Breiman. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California at Berkeley, 1996. ftp://ftp.stat.berkeley.edu/pub/users/breiman/arc97.ps.Z.

[43] Harris Drucker and Corinna Cortes. Boosting decision trees. In *Advances in Neural Information Processing Systems*, volume 8, pages 479–485, 1996.

[44] Marcus Frean and Tom Downs. A simple cost function for boosting. Technical report, University of Queensland, 1998.

[45] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, September 1995. Also appeared in COLT90.

[46] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Eurocolt '95*, pages 23–37. Springer-Verlag, 1995.

[47] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1996.

[48] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for AdaBoost. Technical Report Neuro COLT Technical Report TR-1998-21, GMD FIRST, 1998. (http://www.first.gmd.de/raetsch/Neurocolt_Margin.ps.gz).

[49] J. Ross Quinlan. Bagging, boosting, and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 725–730, 1996.

[50] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[51] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proc. 14th International Conference on Machine Learning*, pages 322–330. Morgan Kaufmann, 1997. (Also to appear in *Annals of Statistics*.)

[52] John Shawe-Taylor, Peter Bartlett, Bob Williamson, and Martin Anthony. Structural risk minimisation over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[53] Bob Williamson, Alex Smola, and Bernhard Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report Neuro COLT Technical Report TR-1998-19, Australian National University, 1998.

URLs: Computational Learning Theory, Boosting, and Support Vector Machines

http://wwwsyseng.anu.edu.au/lsg/ (includes tutorial slides)
http://www.cse.ucsc.edu/research/ml/ (COLT)
http://www.research.att.com/šchapire/boost.html (boosting)
http://svm.first.gmd.de (SVMs)
http://svm.research.bell-labs.com/SVMdoc.html (SVMs)
http://www.dcs.rhbnc.ac.uk/research/compint/machlearn.shtml (SVMs and COLT)