Complements to

‘Pattern Recognition and Neural Networks’

by B.D. Ripley

Cambridge University Press, 1996, ISBN 0-521-46086-7

These complements provide further details, and references which appeared (or came to my attention) after the book was completed in June 1995. Minor corrections can be found in the Errata list.

Chapter 1: Introduction

Page 4:

The book by Przytula & Prasanna (1993) discusses in detail the parallel implementation of neural networks.

Page 16:

Langley (1996) provides a book-length introduction to one viewpoint on machine learning. Langley & Simon (1995) and Bratko & Muggleton (1995) discuss applications of machine learning with claimed real-world benefits.

Valentin *et al.* (1994) survey recent developments in face recognition.

Arbib (1995) provides many short sketches of topics over a very wide range of neural networks, both artificial and biological.

Chapter 2: Statistical Decision Theory

Page 41:

Lauritzen (1996, Chapter 6) gives an extensive treatment of conditional Gaussian distributions, and Edwards (1995) has a more practically-oriented account.

Page 65–66:

There has been a lot of interest in combining classifiers produced by the *same* method on *different* training sets. *Bagging* (Breiman, 1994, 1996a) is an abbreviation for ‘bootstrap aggregating’; the proposal is to take an unweighted average of the predictions of, say 100, classifiers trained on training sets formed by resampling with replacement from the original training set. Breiman (1996b) motivates this for *unstable* methods such as classification trees in which a small change in the training set can lead to a large change in the classifier. (‘Plug-in’ neural network fitting with multiple local minima may also be unstable.)

A variant on this idea which has been suggested many times is to add ‘noise’ to the training set, randomly perturbing either the feature vectors $\mathbf{x}$ or the classes (or both). Further along this line, we could model the joint distribution of $(\mathbf{X}, C)$ and create new training sets from this distribution. Bagging can be seen as the rather extreme form of this procedure in which the model is the empirical distribution. (Krogh & Vedelsby, 1995, use cross-validation rather than re-sampling, and consider designing training sets weighted towards areas where the existing classifiers are prone to disagree.)

Boosting (Schapire, 1990; Freund, 1990; Drucker *et al.*, 1993, 1994; Freund, 1995; Freund & Schapire, 1995, 1996a) is altogether more ambitious. The idea is to *design* a series of training sets and use a combination of classifiers trained on these sets. (Majority voting and linear combinations have both been used.) The training sets are chosen sequentially, with the weights for each example being modified based on the success of the classifier trained on the previous set. (The weight of the examples which were classified incorrectly is increased relative to those which were classified correctly.) For classifiers that do not accept weights on the examples, resampling with probabilities proportional to the weights can be used.

There have been a number of practical tests of boosting, including Drucker (1996); Drucker & Cortes (1996); Freund & Schapire (1996b); Breiman (1996c) and Quinlan (1996a). Each of Freund & Schapire, Breiman and Quinlan compare bagging and boosting for classification trees (although only Quinlan used weighting rather than weighted re-sampling for boosting), and all find that boosting usually out-performs bagging but can fail so badly as to make the boosted classifier worse that the original.

The original motivation for boosting was to produce a combined classifier that could fit the training set perfectly by concentrating on the regions of the feature space which the current classifier found hard. As both Breiman and Quinlan point out, this is a very different aim from producing a classifier with good generalization in problems with overlapping classes, a problem which requires boosting beyond the point needed to fit the training set perfectly. It seems likely that greater study of boosting will lead to algorithms designed to boost generalization error.

Cesa-Bianchi *et al.* (1993, 1996) consider how to learn how to do as well as the best of a class of classifiers over a sequence of examples, and derive bounds for the excess errors made whilst learning which is best. This is a different goal from the usual one in combining classifiers, which is to do better than even the best classifier by exploiting the strengths of all. The algorithm used is boosting-like, maintaining weights for each expert rather than each example.

Page 80:

Chernoff's bound follows from the results of that paper, but in this precise form is due to Angluin & Valiant (1979).

Page 82:

The constants in these results can be refined slightly by using probability inequalities which are specific to binomial distributions. (One difficulty is that the references, like Anthony & Shawe-Taylor, 1993, omit the proofs of these inequalities, which I find far subtler than the ideas which *are* proved. The improvements were omitted because of the lack of accessible proofs to which to refer the reader.)

In (2.50) the factor $1/8$ in the exponent can be removed. Parrondo & Van der Broeck (1993) give

$$\mathsf{Pr}\Big\{\sup_{g\in\mathcal{F}} |\widehat{\mathsf{pmc}}(g) - \mathsf{pmc}(g)| > \epsilon\Big\} \leqslant 6e^{2\epsilon}\Delta(2n)\exp[-n\epsilon^2]$$

and Vapnik (1995, pp. 66, 85) quotes the constant 4 as in (2.50), although proofs are deferred to the forthcoming Vapnik (1996).

The improvements of Parrondo & Van der Broeck rely on

(i) Replacing (2.55) by

$$\mathsf{Pr}\{\sup_{g\in\mathcal{F}} |\widehat{\eta}_1 - \eta| > \epsilon\} \leqslant 2\,\mathsf{Pr}\{\sup_{g\in\mathcal{F}} |\widehat{\eta}_1 - \widehat{\eta}_2| > \epsilon - \tfrac{1}{n}\}.$$

This follows from the binomial inequality

$$\mathsf{Pr}\Big\{\pm\big[\widehat{\eta}_2(\widehat{g}) - \eta(\widehat{g})\big] \leqslant \tfrac{1}{n} \;\Big|\; \widehat{g}\Big\} \geqslant \tfrac{1}{2} \tag{1}$$

for each choice of sign, and so, conditionally on the first sample

$$\begin{aligned}
&\mathsf{Pr}\{|\widehat{\eta}_1(\widehat{g}) - \widehat{\eta}_2(\widehat{g})| > \epsilon - \tfrac{1}{n}\}\\
&\leqslant \mathsf{Pr}\{\widehat{\eta}_1(\widehat{g}) > \eta + \epsilon, \widehat{\eta}_2(\widehat{g}) < \eta + \tfrac{1}{n}\} + \mathsf{Pr}\{\widehat{\eta}_1(\widehat{g}) < \eta - \epsilon, \widehat{\eta}_2(\widehat{g}) > \eta - \tfrac{1}{n}\}\\
&\leqslant \tfrac{1}{2} I\{|\widehat{\eta}_1(\widehat{g}) - \eta| > \epsilon\}
\end{aligned}$$

Inequality (1) says that for a binomial distribution the median lies within one of the mean; it seems well-known to combinatorists, none of whom was able to give me a reference.

(ii) replacing the use of Hoeffding's inequality at the top of page 87 by an inequality of Vapnik (1982):

$$\mathsf{Pr}\Big\{\Big|\sum_{i=1}^{n} Y_i\Big| > n\epsilon\Big\} \leqslant 3e^{-n\epsilon^2}.$$

The bound for proposition 2.5 may be improved to $2\Delta(2n)2^{-n\epsilon}$ by the same ideas. (The statement in Parrondo & Van der Broeck, 1993, does not follow from their proof and appears to be wrong.)

Chapter 3: Linear Discriminant Analysis

Page 116:

Analytical evidence that optimizing the number of errors made by a perceptron is hard is provided by Höffgen *et al.* (1995) who showed that the problem of determining if there is a solution with at most $k \geqslant 1$ misclassifications is NP-hard. (For $k = 0$ it is reducible to a linear programming program as shown on this page, so solvable in polynomial time.)

Page 118:

Minsky & Papert showed that the weights needed in a perceptron could be large for some realistic problems. As for binary inputs the perceptron learning algorithm only changes one coordinate of $\mathbf{a}$ by unity at each step, large weights entail slow convergence.

Prior to Minsky & Papert, Muroga *et al.* (1961) had shown that the weights in a linearly-separating perceptron with n binary inputs could be chosen to be integers less than $(n+1)^{(n+1)/2}2^{-n}$, and Muroga (1965) showed that there were problems where integer weights of $\Omega(2^n)$[1] are required. Hampson & Volper (1986) showed that in some sense an average problem needs integer weights of $\Omega(2^{n/2})$. More recently Håstad found an example which needs integer weights at least as large as $n^{n/2}2^{-n}e^{O(2^{0.585})}$. These results are reviewed and proved by Parberry (1994).

[1]The notation $\Omega(g(n))$ is defined on page 178.

Other learning rules are less restricted. We saw that Mansfield's method makes at most $O(p^3 \log p)$ mistakes before finding a linearly separating solution, a bound reduced to $O(p^2 \log p)$ by Maass & Turán (1994) by using a more recent method in convex optimization. (They also show that every method makes at least $\binom{p}{2}$ mistakes are necessary on some problems with p binary inputs.).

Page 120:

Support-vector machines (Cortes & Vapnik, 1995; Vapnik, 1995). If two classes are linearly separable, there will be a continuum of weight vectors $\mathbf{a}$ which give rise to separating hyperplanes. Amongst these we can choose a hyperplane with maximal distance to the nearest example, achieved by minimizing $\|\mathbf{a}\|^2$ whilst insisting that $\mathbf{za} \geqslant 1$. Finding this hyperplane is a quadratic programming problem, and the usual Kuhn-Tucker optimality conditions show that there will be a subset of examples $\mathbf{z}_i$ (known as *support vectors*) for which $\mathbf{z}_i\mathbf{a} = 1$ and that the optimal $\mathbf{a}$ is a linear combination of these $\mathbf{z}_i$.

The advantage is choosing the optimal hyperplane is to reduce the VC-dimension of the space of solutions (which is proportional to a bound on $\|\mathbf{a}\|^2$). If the two classes *are* linearly separable then (Vapnik, 1995, Theorem 5.2) the expected error rate on future examples is bounded by the expected number of support vectors divided by $n - 1$. Thus finding a small number of support vectors might indicate good generalization properties.

Of course, linear separation in the original feature space is quite rare, but as for *generalized linear discrimination* (page 121) we can expand the feature space by using polynomials or even radial-basis function networks and sigmoidal functions. These can give rise to very large feature spaces, but generalization may remain acceptable if the number of support vectors remains small, which was the case in the experiments reported by Vapnik (1995, section 5.7).

By jointly minimizing the sum of the degree of error (page 116) and $\|\mathbf{a}\|^2$ these ideas can be extended to non-separable two-class problems (Cortes & Vapnik, 1995).

Chapter 4: Flexible Discriminants

Page 125:

Further examples of using additive models with interacxtion terms in logistic regression are given by Wahba *et al.* (1996); Kooperberg *et al.* (1996); Stone *et al.* (1997).

Chapter 5: Feed-forward Neural Networks

Page 157:

According to Werbos (1995), the weight-decay penalty $\sum_{ij} w_{ij}^2$ was also proposed by Werbos (1987).

Page 165:

We saw that a local minimum which fits well is not important in the Bayesian integration if it corresponds to a sharp peak of the posterior density of the weights. Hochreiter & Schmidhuber (1995, 1996) add a penalty to the optimization to encourage exploring broad peaks. However, they still use local optimization, and their penalty is better regarded as an elaborate form of weight decay.

Page 172:

Marchand *et al.* (1990) had another early construction algorithm.

Page 179:

This VC-dimension result for threshold-unit neural networks was anticipated by Cover (1968).

Page 180:

The description of the $O(W \log W)$ and $\Omega(W \log W)$ results fails to make clear what is allowed to vary; these results apply to networks where both the number of input units and the number of hidden units are allowed to increase. In that case Sakurai (1993) has $\Omega(W \log W)$ results for networks with one hidden layer and real inputs (whereas Maass considered binary inputs).

For sigmoidal neural networks, Karpinski & Macintyre (1995a,b) showed that the VC-dimension is $O(W^4)^2$, and Koiran & Sontag (1996) showed that $\Omega(W^2)$. Bartlett & Williamson (1996) bound the VC-dimension for a single-hidden-layer network by $2W \log_2(24e\,WD)$ if the inputs are restricted to $\{-D, \ldots, D\}$.

Chapter 6: Non-parametric Methods

Page 197–198:

Lowe (1995) proposes a distance-weighted nearest-neighbour method with a Gaussian kernel ($\exp -d_j^2/\sigma^2$) where σ is proportional to the average distance to the, say, first 5 neighbours, and d_j^2 is weighted by weights for each feature. (Thus a quadratic metric with diagonal A is used.) Both the constant of proportionality and the metric are chosen by (leave-one-out) cross-validation. Like that of Fukunaga & Flick (1984) this is a global method, but Lowe mentions that the feature weights could be chosen locally, as was done by Atkeson (1991) in a control problem. The restriction to a diagonal metric places the pre-processing of the features at a premium.

Friedman (1994) and Hastie & Tibshirani (1996a,b) look for a local metric. Friedman works in the spirit of Short & Fukunaga (1980, 1981), using a local measure of relevance and choosing a hyper-rectangle (corresponding to an L_∞ distance with a diagonal weighting matrix) by recursive partitioning on this relevance measure. The proposal of Hastie & Tibshirani is much simpler; they use Euclidean distance on the linear discriminants for points within a small neighbourhood (defined by this metric and so chosen recursively).

Page 200–201:

Further references on 'memory-based reasoning' are Waltz (1990); Aha *et al.* (1991); Cost & Salzberg (1993) and Rachlin *et al.* (1994).

Chapter 7: Tree-structured Classifiers

Page 236–7:

Debate has continued in the machine-learning community on how to choose amongst attributes at a split. Buntine & Niblett (1992) reviewed approaches at that time. More recently, some authors have spotted (Catlett, 1991; Auer *et al.*, 1995; Dougherty *et al.*,

[2] with an explicit bound for a single-output network which has a leading term of $(WM)^2/2$ where M is the number of sigmoidal units

1995) that discretizing a continuous attribute can improve the performance of a classification tree induced by Quinlan's C4.5 and rarely makes the performance worse. This might arise since for a continuous attribute X the only splits allowed are of the form $X \leqslant t$, whereas for a nominal attribute *any* split of the values into two groups is allowed, and this can result in splitting a discretized continuous attribute into a series of intervals and their complement. Indeed, Holte (1993) and Auer *et al.* (1995) have shown that very shallow trees (one level or two levels respectively) can be rather effective on many common problems if splits of this sort are allowed.

This has led to discussion of algorithms to find 'efficient' and 'optimal' multi-way splits of continuous attributes (Fayyad & Irani, 1993; Fulton *et al.*, 1995; Elomaa & Rousu, 1996). The idea is to partition a continuous attribute into $K > 2$ disjoint intervals, and have K daughter nodes, in such a way as to maximize the reduction in impurity. (Elomaa & Rousu, 1996 point out errors and inefficiencies in the earlier algorithms.)

Quinlan (1996b) disputes whether this is the actual cause of the increased performance, and proposes amendments to C4.5 which in his experiments regain its advantage over discretization. There are several small changes, but the major effect is to introduce a bias against continuous attributes, as the wide choice of threshold t gives a selection bias in favour of continuous attributes. (If there are N values of X in the training set, the information gain of $X \leqslant t$ is reduced by $\log_2(N-1)$.)

Page 240:
There has been renewed interest in taking linear combinations of variables at nodes, sometimes known as *perceptron*, *neural*, *multivariate* or *oblique* trees: Utgoff (1989); Heath *et al.* (1993); Murthy *et al.* (1993, 1994); Brodley & Utgoff (1995).

Page 241:
Craven & Shavlik (1996) consider using trees to 'explain' the classifier modelled by a neural network, that is approximates a neural network by a classification tree, trained by querying the neural network as an oracle. (This is related to Angluin's ideas on page 7.)

Page 242:
There has been further work on averaging multiple decision trees. One idea is to replace the pruning of a tree by a weighted average over all prunings. Although there are very many such prunings, averaging a prediction over prunings amounts to a weighted average over potential terminal nodes (usually all nodes) and so is feasible; see Willems *et al.* (1993, 1995); Oliver & Hand (1995); Helmbold & Schapire (1995, 1996). Other combination ideas are given by Ali & Pazzani (1995); Ho (1995) and Shlien (1990).

Chapter 8: Belief Networks

Page 245:
Applications of Bayesian belief networks are considered in in the March 1995 issue of *Communications of the ACM* by (Heckerman & Wellman, 1995; Burnell & Horvitz, 1995; Heckerman *et al.*, 1995; Fung & Del Favarro, 1995).

Page 245, 258–262:
Almond (1995) is a thesis-length treatment of graphical computations for Dempster-Shafer belief functions, which have probability calculations as a special case. There are many possible join trees, and Almond considers other constructions which may

produce trees better-suited to the computational complexity of belief-function computations. Also, as belief functions do not have a division operator, the calculations have to be organised to exclude the denominators from the product in the numerator. [Despite its claimed date, Almond's book was published several months into 1996, with 1993 papers cited as 'to appear' in its references.]

Page 258–262:
Lauritzen (1996, Chapter 2) gives a self-contained account of these graph-theoretical properties based on decomposability. Shafer (1996) will give an in-depth account of join trees and their rôle in probabilistic expert systems. [It is due in August 1996.]

Chapter 9: Unsupervised Methods

Page 304:
The experiments of Wang & Oja (1993) comparing a five-layer auto-associator with principal component analysis found that the latter gave better generalization.

Glossary

Page 354:
Benveniste *et al.* (1990) and Ljung *et al.* (1992) are books on stochastic approximation.

References

Aha, D. W., Kibler, D. & Albert, M. K. (1991) Instance-based learning algorithms. *Machine Learning* **6**(1), 37–66.

Ali, K. M. & Pazzani, M. J. (1995) On the link between error correlation and error reduction in decision tree ensembles. Technical Report 95-38, Department of Information and Computer Science, University of California at Irvine.

Almond, R. G. (1995) *Graphical Belief Modeling*. London: Chapman & Hall. ISBN 0-412-06661-0. [Despite the date, this book was actually published in May 1996].

Angluin, D. & Valiant, L. G. (1979) Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences* **18**, 155–193.

Anthony, M. & Shawe-Taylor, J. (1993) A result of Vapnik with applications. *Discrete Applied Mathematics* **47**, 207–217. [Erratum (1994) **52**, 211 (the proof of theorem 2.1 is corrected)].

Arbib, M. A. (ed.) (1995) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press. ISBN 0-262-01148-4.

Atkeson, C. G. (1991) Using locally weighted regression for robot learning. In *Proceedings of the IEEE Conference on Robotics and Automation (Sacremento, CA, 1991)*, pp. 958–963. IEEE Press.

Auer, P., Holte, R. C. & Maass, W. (1995) Theory and application of agnostic PAC-learning with small decision trees. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 21–29. San Francisco: Morgan Kaufmann. Also NeuroCOLT Technical Report NC-TR-96-034 (Feb 1996).

Bartlett, P. L. & Williamson, R. C. (1996) The VC dimension and pseudodimension of two-layer neural networks with discrete inputs. *Neural Computation* **8**(3), 625–628.

Benveniste, A., Métivier, M. & Priouret, P. (1990) *Adaptive Algorithms and Stochastic Approximations*. Berlin: Springer.

Bratko, I. & Muggleton, S. (1995) Applications of inductive logic programming. *Communications of the Association for Computing Machinery* **38**(11), 65–70.

Breiman, L. (1994) Bagging predictors. Technical Report 421, Department of Statistics, University of California at Berkeley.

Breiman, L. (1996a) Bagging predictors. *Machine Learning* .

Breiman, L. (1996b) The heuristics of instability in model selection. *Annals of Statistics* .

Breiman, L. (1996c) Bias, variance and arcing classifiers. Technical report, Department of Statistics, University of California at Berkeley.

Brodley, C. E. & Utgoff, P. E. (1995) Multivariate decision trees. *Machine Learning* **19**, 45–77.

Buntine, W. & Niblett, T. (1992) A further comparison of splitting rules for decision-tree induction. *Machine Learning* **8**, 75–86.

Burnell, L. & Horvitz, E. (1995) Structure and chance: Melding logic and probability for software debugging. *Communications of the ACM* **38**(3), 31–41, 57.

Catlett, J. (1991) On changing continuous attributes into ordered discrete attributes. In *Proceedings of the European Working Session on Learning – EWSL-91*, ed. Y. Kodratoff, pp. 164–178. Berlin: Springer.

Cesa-Bianchi, M., Freund, Y., Helmbold, D. P., Haussler, D., Schapire, R. E. & Warmuth, M. K. (1993) How to use expert advice. In *Proceedings of the Twenth-Fifth ACM Symposium on the Theory of Computing (San Diego, CA, 1993)*, pp. 382–391. New York: ACM Press.

Cesa-Bianchi, M., Freund, Y., Helmbold, D. P., Haussler, D., Schapire, R. E. & Warmuth, M. K. (1996) How to use expert advice. *Journal of the ACM* .

Cortes, C. & Vapnik, V. (1995) Support-vector networks. *Machine Learning* **20**, 273–297.

Cost, S. & Salzberg, S. (1993) A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning* **10**, 57–78.

Cover, T. M. (1968) Capacity problems for linear machines. In *Pattern Recognition*, ed. L. Kanal, pp. 283–289. Thompson.

Craven, M. W. & Shavlik, J. W. (1996) Extracting tree-structured representations of trained networks. In Touretzky *et al.* (1996), pp. 24–30. ISBN 0-262-20107-0.

Dasarathy, B. V. (ed.) (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Computer Society Press.

Dougherty, J., Kohavi, R. & Sahami, M. (1995) Supervised and unsupervised discretization of continuous features. In *Proceedings of the Twelfth International Conference on Machine Learning*, eds A. Prieditis & S. Russell, pp. 194–202. San Francisco: Morgan Kaufmann.

Drucker, H. (1996) Fast decision tree ensembles for optical character recognition. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrival*.

Drucker, H. & Cortes, C. (1996) Boosting decision trees. In Touretzky *et al.* (1996), pp. 479–485. ISBN 0-262-20107-0.

Drucker, H., Schapire, R. & Simard, P. (1993) Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence* **7**(4), 705–719.

Drucker, H., Cortes, C., Jaeckel, L. D., LeCun, Y. & Vapnik, V. (1994) Boosting and other ensemble methods. *Neural Computation* **6**(6), 1289–1301.

Edwards, D. (1995) *Introduction to Graphical Modelling*. Springer.

Elomaa, T. & Rousu, J. (1996) Finding optimal multi-splits for numerical attributes in decision tree learning. NeuroCOLT Technical Report Series NC-TR-96-041, Department of Computer Science, University of Helsinki.

Fayyad, U. M. & Irani, K. B. (1993) Multi-interval discretization of continuous-valued attributes in decision tree generation. In *Proceedings of the the Thirteenth International Joint Conference on Artificial Intelligence (Chambery, France, 1993)*, pp. 1022–1027. San Francisco: Morgan Kaufmann.

Freund, Y. (1990) Boosting a weak learning algorithm by majority. In *Proceedings of the Third Workshop on Computational Learning Theory*, pp. 202–216. Morgan Kaufmann.

Freund, Y. (1995) Boosting a weak learning algorithm by majority. *Information and Computation* **121**(2), 256–285.

Freund, Y. & Schapire, R. E. (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, pp. 23–37. Springer.

Freund, Y. & Schapire, R. E. (1996a) Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*.

Freund, Y. & Schapire, R. E. (1996b) Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*.

Friedman, J. H. (1994) Flexible metric nearest neighbor classification. Technical report, Department of Statistics, Stanford University.

Fukunaga, K. & Flick, T. E. (1984) An optimal global nearest neighbor metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 314–318. [Reprinted in Dasarathy (1991)].

Fulton, T., Kasif, S. & Salzberg, S. (1995) Efficient algorithms for finding multi-way splits for decision trees. In *Proceedings of the Twelfth International Conference on Machine Learning*, eds A. Prieditis & S. Russell, pp. 244–251. San Francisco: Morgan Kaufmann.

Fung, R. & Del Favarro, B. (1995) Applying Bayesian networks to information retrival. *Communications of the ACM* **38**(3), 42–48, 57.

Hampson, S. E. & Volper, D. J. (1986) Linear function neurons: structure and training. *Biological Cybernetics* .

Hastie, T. & Tibshirani, R. (1996a) Discriminant adaptive nearest neighbor classification and regression. In Touretzky *et al.* (1996), pp. 409–415. ISBN 0-262-20107-0.

Hastie, T. & Tibshirani, R. (1996b) Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**, 607–618.

Heath, D., Kasif, S. & Salzberg, S. (1993) Learning oblique decision trees. In *Proceedings of the the Thirteenth International Joint Conference on Artificial Intelligence (Chambery, France, 1993)*, pp. 1002–1007. San Francisco: Morgan Kaufmann.

Heckerman, D. & Wellman, M. P. (1995) Bayesian networks. *Communications of the ACM* **38**(3), 26–30.

Heckerman, D., Breese, J. S. & Rommelse, K. (1995) Decision-theoretic troubleshooting. *Communications of the ACM* **38**(3), 49–57.

Helmbold, D. P. & Schapire, R. E. (1995) Predicting nearly as well as the best pruning of a decision tree. In *Proceedings of the Eight Annual Conference on Computational Learning Theory*, pp. 61–68. New York: ACM Press.

Helmbold, D. P. & Schapire, R. E. (1996) Predicting nearly as well as the best pruning of a decision tree. *Machine Learning* .

Ho, T. K. (1995) Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, pp. 278–282. IEEE Computer Society Press.

Hochreiter, S. & Schmidhuber, J. (1995) Simplifying neural nets by discovering flat minima. In Tesauro *et al.* (1995), pp. 529–536. ISBN 0-262-20104-6.

Hochreiter, S. & Schmidhuber, J. (1996) Flat minima. *Neural Computation* .

Holte, R. C. (1993) Very simple classification rules perform well on most commonly used datasets. *Machine Learning* **11**, 63–91.

Höffgen, K.-U., Simon, H.-U. & Van Horn, K. S. (1995) Robust trainability of single neurons. *Journal of Computer and System Sciences* **50**(1), 114–125.

Karpinski, M. & Macintyre, A. (1995a) Bounding VC-dimension of neural networks: Progress and prospects. In *Proceedings of the Second European Conference on Computational Learning Theory (Barcelona, Spain)*, ed. P. Vitanyi, number 904 in Lecture Notes in Artificial Intelligence, pp. 337–341. Berlin: Springer.

Karpinski, M. & Macintyre, A. (1995b) Polynomial bounds for VC dimension of sigmoidal neural networks. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing (Las Vegas)*, pp. 200–208. ACM Press.

Koiran, P. & Sontag, E. D. (1996) Neural networks with quadratic VC dimension. In Touretzky *et al.* (1996), pp. 197–203. ISBN 0-262-20107-0.

Kooperberg, C., Bose, S. & Stone, C. J. (1996) Polychotomous regression. *Journal of the American Statistical Association* .

Krogh, A. & Vedelsby, J. (1995) Neural network ensembles, cross validation, and active learning. In Tesauro *et al.* (1995), pp. 231–238. ISBN 0-262-20104-6.

Langley, P. (1996) *Elements of Machine Learning*. San Francisco: Morgan Kaufmann.

Langley, P. & Simon, H. A. (1995) Applications of machine learning and rule induction. *Communications of the ACM* **38**(11), 54–64.

Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon Press. ISBN 0-19-852219-3.

Ljung, L., Pflug, H. & Walk, H. (1992) *Stochastic Approximation and Optimization of Random Systems*. Berlin: Birkhäuser.

Lowe, D. G. (1995) Similarity metric learning for a variable-kernel classifier. *Neural Computation* **7**(1), 72–85.

Maass, W. & Turán, G. (1994) How fast can a threshold gate learn? In *Computational Learning Theory and Natural Learning Systems: Constraints and Prospects*, eds S. J. Hanson, G. A. Drastal & R. L. Rivest, volume I, pp. 381–414. MIT Press.

Marchand, M., Golea, M. & Rujan, P. (1990) A convergence theorm for sequential learning in two-layer perceptrons. *Europhysics Letters* **11**, 487–492.

Muroga, S. (1965) Lower bounds of the number of threshold functions and a maximum weight. *IEEE Transactions on Electronic Computers* **14**, 136–148.

Muroga, S., Toda, I. & Takasu, S. (1961) Theory of majority decision elements. *Journal of the Franklin Institute* **271**, 376–418.

Murthy, S. K., Kasif, S., Salzberg, S. & Beigel, R. (1993) OC1: randomized induction of oblique decision tress. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (Washington, DC, 1993)*, pp. 322–327. AAAI Press. ISBN 0-262-51071-5.

Murthy, S. K., Kasif, S. & Salzberg, S. (1994) A system for the induction of oblique decision trees. *Journal of Artificial Intelligence Research* **2**, 1–33.

Oliver, J. J. & Hand, D. J. (1995) On pruning and averaging decision trees. In *Machine Learning: Proceedings of the Twelfth International Conference*, pp. 430–437. Morgan Kaufmann.

Parberry, I. (1994) *Circuit Complexity and Neural Networks*. Cambridge, MA: MIT Press. ISBN 0-262-16148-6.

Parrondo, J. M. R. & Van der Broeck, C. (1993) Vapnik-Chervonenkis bounds for generalization. *J. Phys. A* **26**, 2211–2223.

Przytula, K. W. & Prasanna, V. K. (1993) *Parallel Digital Implementation of Neural Networks*. Englewood Cliffs, NJ: Prentice Hall.

Quinlan, J. R. (1996a) Bagging, boosting, and C4.5. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Quinlan, J. R. (1996b) Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research* **4**, 77–90.

Rachlin, J., Kasif, S., Salzberg, S. & Aha, D. (1994) Towards a better understanding of memory-based and Bayesian classifiers. In *Proceedings of the Eleventh International Conference on Machine Learning (New Brunswick, NJ)*, pp. 242–250.

Sakurai, A. (1993) Tighter bounds of the VC-dimension of three-layer networks. In *Proceedings of the 1993 World Congress on Neural Networks*, volume 3, pp. 540–543. Hillsdale, NJ: Erlbaum.

Schapire, R. E. (1990) The strength of weak learnability. *Machine Learning* **5**(2), 197–227.

Shafer, G. (1996) *Probabilistic Expert Systems*. Number 67 in CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PA: SIAM. ISBN 0-89871-373-0.

Shlien, S. (1990) Multiple binary decision tree classifiers. *Pattern Recognition* **23**(7), 757–763.

Short, R. D. & Fukunaga, K. (1980) A new nearest neighbor distance measure. In *Proceedings of the Fifth IEEE International Conference on Pattern Recognition (Miami Beach, 1980)*, pp. 81–86. Los Alamitos, CA: IEEE Computer Society Press.

Short, R. D. & Fukunaga, K. (1981) The optimal distance measure for nearest neighbor classification. *IEEE Transactions on Information Theory* **27**, 622–627. [Reprinted in Dasarathy (1991)].

Stone, C. J., Hansen, M., Kooperberg, C. & Truong, Y. K. (1997) Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics* .

Tesauro, G., Touretzky, D. S. & Leen, T. K. (eds) (1995) *Advances in Neural Information Processing Systems 7. Proceedings of the 1994 Conference*. Cambridge, MA: MIT Press. ISBN 0-262-20104-6.

Touretzky, D. S., Moser, M. C. & Hasselmo, M. E. (eds) (1996) *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*. Cambridge, MA: MIT Press. ISBN 0-262-20107-0.

Utgoff, P. E. (1989) Perceptron trees: a case study in hybrid concept representations. *Connection Science* **1**(4), 377–391.

Valentin, D., Abdi, H., O'Toole, A. J. & Cottrell, G. (1994) Connectionist models of face processing: A survey. *Pattern Recognition* **27**, 1208–1230.

Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik, V. N. (1996) *Statistical Learning Theory*. New York: Wiley.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1996) Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *Annals of Statistics* **23**(6), 1865–1895.

Waltz, D. (1990) Memory-based reasoning. In *Natural and Artificial Parallel Computation*, eds M. Arbib & J. Robinson, pp. 251–276. Cambridge, MA: MIT Press.

Wang, L. & Oja, E. (1993) Image compression by MLP and PCA neural networks. In *Proceedings of the Eight Scandinavian Conference on Image Analysis (Tromsø, Norway)*, pp. 1317–1324.

Werbos, P. (1987) Learning how the world works. In *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, pp. 320–310. New York: IEEE Press.

Werbos, P. J. (1995) Backpropagation: Basics and new developments. pp. 134–139 of Arbib (1995).

Willems, F. M. J., Shtarkov, Y. M. & Tjalkens, T. J. (1993) Context tree weighting: A sequential universal source coding procedure for FSMX sources. In *Proceedings of the 1993 IEEE International Symposium on Information Theory*, p. 59. IEEE Press.

Willems, F. M. J., Shtarkov, Y. M. & Tjalkens, T. J. (1995) The context-tree weighting method: Basic properties. *IEEE Transactions on Information Theory* pp. 653–664.