Report Number: WUCS-91-29

1991-06-01

The Difficulty of Random Attribute Noise

Sally A. Goldman and Robert H. Sloan

This paper studies the robustness of pac learning algorithms when the instance space is {0,1}n, and the examples are corrupted by purely random noise affecting only the instances (and not the labels). In the past, conflicting results on this subject have been obtained-- the "best agreement" rule can only tolerate small amounts of noise, yet in some cases large amounts of noise can be tolerated. We show the truth lies somewhere between the two alternatives. For uniform attribute noise, in which each attribute is flipped independently at random with the same probability, we present an algorithm that pac learns... **Read complete abstract on page 2.**

The Difficulty of Random Attribute Noise

Sally A. Goldman and Robert H. Sloan

Complete Abstract:

This paper studies the robustness of pac learning algorithms when the instance space is {0,1}n, and the examples are corrupted by purely random noise affecting only the instances (and not the labels). In the past, conflicting results on this subject have been obtained-- the "best agreement" rule can only tolerate small amounts of noise, yet in some cases large amounts of noise can be tolerated. We show the truth lies somewhere between the two alternatives. For uniform attribute noise, in which each attribute is flipped independently at random with the same probability, we present an algorithm that pac learns monomial for an (unknown) noise rate less than 1/2. Contrasting this positive result, we show that nonuniform random attribute noise, where each attribute i is flipped randomly and independently with its own probability pi, is nearly as harmful as malicious noise-- no algorithm can tolerate more than a small amount of such noise.

The Difficulty of Random Attribute Noise

Sally A. Goldman and Robert H. Sloan

WUCS-91-29

June 1991

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
Saint Louis, MO 63130-4899

The Difficulty of Random Attribute Noise

Sally A. Goldman*
Department of Computer Science
Washington University
St. Louis, Missouri 63130

Robert H. Sloan†
Dept. of Electrical Engineering and Computer Science
University of Illinois at Chicago
Chicago, IL 60680

June 6, 1991

Abstract

This paper studies the robustness of pac learning algorithms when the instance space is $\{0,1\}^n$, and the examples are corrupted by purely random noise affecting only the instances (and not the labels). In the past, conflicting results on this subject have been obtained—the "best agreement" rule can only tolerate small amounts of noise, yet in some cases large amounts of noise can be tolerated.

We show that the truth lies somewhere between these two alternatives. For *uniform attribute noise*, in which each attribute is flipped independently at random with the same probability, we present an algorithm that pac learns monomials for any (unknown) noise rate less than 1/2. Contrasting this positive result, we show that *nonuniform random attribute noise*, where each attribute i is flipped randomly and independently with its own probability p_i, is nearly as harmful as malicious noise—no algorithm can tolerate more than a very small amount of such noise.

---

*Supported in part by a GE Foundation Junior Faculty Grant. Part of this research was conducted while the author was at the M.I.T. Laboratory for Computer Science and supported by NSF grant DCR-8607494 and a grant from the Siemens Corporation. Net address: sg@cs.wustl.edu

†Part of this research was conducted while the author was at Harvard and supported by ARO grant DAAL 03-86-K-0171. Net address: sloan@ss1.eecs.uic.edu

1 Introduction

This paper studies the robustness of pac learning algorithms for learning boolean functions (i.e. the instance space is $\{0,1\}^n$). In particular, we examine random noise processes that corrupt examples by independently inverting each attribute bit with a probability given by the noise rate. We assume throughout that the classification of each instance is always correctly reported.

In the past, conflicting results on handling random attribute noise have been obtained. On the one hand, Sloan [11] has show that the "best-agreement" rule can only tolerate very small amounts of random attribute noise—suggesting that random attribute noise may be difficult to overcome. However, by using a different strategy, Shackelford and Volper [10] have obtained an algorithm that tolerates a large amount of random attribute noise (at a *known* noise rate) for learning k-DNF formulas. Thus their result suggests that random attribute noise may be like random classification noise, where large amounts of noise can sometimes be tolerated.

We show that the truth lies somewhere between these two alternatives. In order to fully understand the difficulty of overcoming random attribute noise, we first carefully examine the way in which the noise is formally modeled. All previous work has considered *uniform random attribute noise*, in which each attribute is flipped independently at random with the same probability.

The algorithm of Shackelford and Volper demonstrates that when the exact noise rate is known, at least for learning k-DNF formulas, a large amount of uniform random attribute noise can be tolerated. In this paper we extend that positive result by showing that even if the exact noise rate is *unknown* (only an upper bound on the noise rate is needed), the class of monomials is still efficiently learnable under a large amount of uniform random attribute noise.

Contrasting this positive result, we show that *nonuniform random attribute noise*, where each attribute i is flipped randomly and independently with its own probability p_i (all p_i are less than some given upper bound for the noise rate), is nearly as harmful as malicious noise. That is, no algorithm (regardless of sample complexity or computation time) can tolerate more than a very small amount of nonuniform random attribute noise. On the whole, these results are surprising. Intuitively, one would think that random labeling noise destroys much more information than random attribute noise.

2 Review of pac learning from noisy data

The method most commonly used for pac learning in the presence of noise is to pick a concept that has the best (or at least very good) agreement with a sample of data corrupted by noise. It has been shown that for both discrete [1] and continuous [8] instance spaces, the hypothesis that minimizes disagreements meets the pac criterion when the examples are modified by random labeling noise. Similarly, Kearns and Li [6] have shown that this method of minimizing disagreements can tolerate a small amount of malicious noise in discrete instance spaces. So these results specify the amount of noise that can be tolerated ignoring the issue of computation time. (Of course, if a hypothesis minimizing disagreements can be found in polynomial time then the above techniques produce efficient learning algorithms.) Sloan [11] has extended those results to the case of malicious labeling noise. Finally, as Blumer et al. mention [2], their VC dimension methods can be used to prove that this minimal disagreement method also works for handling small amounts of malicious noise in continuous instance spaces.

In the case of uniform random attribute noise, if one uses the minimal disagreement method, then the minimum error rate obtainable (i.e. the minimum "epsilon") is bounded below by the noise rate [11]. We note that for arbitrary adversarial malicious noise, that is the maximum noise rate that *any* algorithm can tolerate [5].

Although the method of minimizing disagreements is not effective against random attribute noise, there are techniques for coping with uniform random attribute noise. In particular, Shackelford and Volper [10] have an algorithm that tolerates large amounts of random attribute noise for learning k-DNF formulas. That algorithm, however, has one very unpleasant requirement: it must be given the *exact* noise rate as an input. In Section 4, we present a new algorithm for learning monomials that tolerates large amounts of uniform random attribute noise (any noise rate less than 1/2), and only requires some upper bound on the noise rate as an input.

Once we move to nonuniform random attribute noise, however, we can no longer find such algorithms. We show in Section 5 that with nonuniform random attribute noise, the minimum error rate obtainable is bounded below by one-half of the noise rate, regardless of the technique (or computation time) of the learning algorithm.

3 Notation

We assume that the reader is familiar with the model of pac learning introduced by Valiant [12]. Good discussions of the details of the model are given by Kearns et al. and by Haussler et al. [7, 3].

Briefly, a *concept* is a subset of some *instance space* X, and a concept class is some subset of 2^X. An *example* of a concept c is a pair (x, s), where $x \in X$, and s is 1 if $x \in c$ and 0 otherwise. We call a sequence of examples a *sample.*

We assume a fixed (but unknown) probability distribution $\mathbf{D}$ on X. Furthermore, the learner who is trying to learn concept c has available to it a black box or oracle called EX such that each call to EX returns a labeled instance, (x, s) where x is drawn at random from $\mathbf{D}$ and labeled according to c. The learner's goal is the following: Given parameters $0 < \epsilon, \delta \leq 1$, draw a sample from EX, and output some representation of a concept $\hat{c}$ such that

$$\Pr\left[\mathbf{D}(c \bigtriangleup \hat{c}) > \epsilon\right] \leq \delta,$$

where $\bigtriangleup$ denotes symmetric difference, and the probability is over the calls to EX and any coin flips used by the learning algorithm. Such a $\hat{c}$ is called *ϵ-good.*

The ordinary definition of pac learning (from noiseless data) assumes that EX returns correct data. In this paper we are concerned with the case in which our instances come from some noise oracle, instead of the usual noise-free oracle, EX. Each noise oracle represents some noise process being applied to the examples from EX. The output from the noise process is all the learner can observe. The "desired," noiseless output of each oracle would thus be a correctly labeled example (x, s), where x is drawn according to $\mathbf{D}$. We now describe the actual outputs from the following noise oracles: MAL_ν [13], RMC_ν [1], MMC_ν [11], URA_ν [11], and NRA_ν.

- When MAL_ν is called, with probability $1 - \nu$, it does indeed return a correctly labeled (x, s) where x is drawn according to $\mathbf{D}$. With probability ν it returns an example (x, s) about which no assumptions whatsoever may be made. In particular, this example may be maliciously selected by an adversary who has infinite computing power, and has knowledge of the target concept, $\mathbf{D}$, ν, and the internal state of the algorithm calling this oracle. This *malicious noise* oracle models the situation where the learner usually gets a correct example, but some small fraction ν of the time the learner gets noisy examples and the nature of the noise is unknown or unpredictable.

- When RMC_ν is called, it calls EX to obtain some (noiseless) (x, s), and with probability $1 - \nu$, RMC_ν returns (x, s). However, with probability ν, RMC_ν returns $(x, \overline{s})$. This *random misclassification noise* oracle models a benign form of misclassification noise.

- When MMC_ν is called, it also calls EX to obtain some (noiseless) (x, s), and with probability $1-\nu$, MMC_ν returns (x, s). With probability ν, MMC_ν returns (x, l) where l is a label about which no assumption whatsoever may be made. As with MAL_ν we assume an omnipotent, omniscient adversary; but in this case the adversary only gets to choose the label of the example. This *malicious misclassification noise* oracle models a situation in which the only source of noise is misclassification, but the nature of the misclassification is unknown or unpredictable.

- We consider the oracle URA_ν only when the instance space is $\{0, 1\}^n$ (i.e., we are learning boolean functions). The oracle URA_ν calls EX and obtains some $(x_1 \cdots x_n, s)$. URA_ν then adds noise to this example by independently flipping each bit x_i to $\bar{x}_i$ with probability ν for $1 \leq i \leq n$. Note that the label of the "true" example is never altered by URA_ν. This *uniform random attribute noise* oracle models a situation where the attributes of the examples are subject to noise, but that noise is as benign as possible. For example, the attributes might be sent over a noisy channel.

- The oracle NRA_ν also only applies when we are learning boolean functions. This oracle calls EX and obtains some $(x_1 \cdots x_n, s)$. The oracle NRA_ν then adds noise by independently flipping each bit x_i to $\bar{x}_i$ with some fixed probability $\nu_i \leq \nu$ for each $1 \leq i \leq n$. This *nonuniform random attribute noise* oracle provides a more realistic model of random attribute noise than URA_ν.[1]

The noise oracles we will focus on here are URA_ν and NRA_ν.

[1]Technically, NRA_ν specifies a family of oracles, each member of which is specified by n variables, $\nu_1, \ldots, \nu_n$, where $0 \leq \nu_i \leq \nu$.

4 Learning monomials from noisy data

In this section we present an algorithm for learning monomials from data corrupted with uniform random attribute noise. The examples will come from URA_ν, and the only prior information about the noise rate ν that the learner will have is some bound $\nu_b < 1/2$ such that $0 \leq \nu \leq \nu_b$.

The key idea we exploit is the following: Imagine that the literal x_1 is included in the target concept. Then whenever the learner receives a positive instance with the first bit off, it must be that the bit was on in the "noise free" instance, and flipped by the noise oracle. Hence, the ratio

$$\frac{\text{number of times bit 1 is off in positive instance}}{\text{total number of positive instances}} \tag{1}$$

provides a good estimate of the noise rate. Notice also that if x_1 is *not* in the formula, then the ratio specified in (1) still is bounded below by the noise rate. Thus we will estimate the noise rate to be the minimum, over all literals, of the ratio specified in (1) for the literal x_1.

More formally, let the literals be numbered from 1 to $2n$ (Say x_1 is 1, x_2 is 2, ..., $\bar{x}_1$ is $n+1$, ..., and $\bar{x}_n$ is $2n$.), and for each literal i, let

$$\begin{aligned} q_i &= \Pr[\text{Literal } i \text{ is off in a positive instance from EX (noiseless data)}] \\ p_i &= \Pr[\text{Literal } i \text{ is off in a positive instance from } \mathrm{URA}_\nu]. \end{aligned}$$

Our goal is to output a conjunction that contains every literal that is in the target monomial, and no literal with a high value of q_i. Of course, we cannot directly estimate the q_i's since we only see examples from URA_ν. Our method is to accurately estimate all the p_i's using examples from URA_ν and then use these estimates to determine which literals have a high value for q_i and should thus be excluded. Observe that for all i

$$\begin{aligned} p_i &= q_i(1-\nu) + (1-q_i)\nu \\ &= \nu + q_i(1-2\nu) \end{aligned} \tag{2}$$

Thus for any literal i that is in the target monomial, $p_i = \nu$. Furthermore, since $\nu < 1/2$, for any literal i that is not in the target monomial $p_i \geq \nu$. We will show that by accurately estimating all the p_i's we can obtain a good estimate for the noise rate by simply taking the minimum estimated value over all the p_i's. We then output the conjunction of all literals i having values of p_i close to that minimum. The algorithm is specified in full in Figure 1.

Theorem 1 *The monomial algorithm specified in Figure 1 pac learns given data from* URA_ν.

Proof: We now prove that the error of the hypothesis output by our algorithm has error less than ϵ with probability at least $1-\delta$. Let p_+ denote the probability of drawing a positive example from URA_ν. (Note that since the noise process does not affect the labels, p_+ is also the probability of drawing a positive example directly from EX.) By applying Hoeffding's Inequality [4] it is easily shown that if $p_+ \geq \epsilon$ then with probability at least $1-\delta/2$ the algorithm will obtain enough positive examples in step 2. Of course, if $p_+ < \epsilon$ then the hypothesis FALSE is ϵ-good.

We now show that if enough positive examples are obtained in step 2, then the hypothesis output in step 5 is ϵ-good with high probability. Again, by applying Hoeffding's Inequality [4], it is easily shown that the probability that *all* of the estimates $\hat{p}_i$ are within $\epsilon(1-2\nu_b)/8n$ of their true value p_i is at least $1-\delta/2$. That is,

$$\Pr\left[\bigwedge_{(1\leq i\leq 2n)} p_i - \frac{\epsilon(1-2\nu_b)}{8n} < \hat{p}_i < p_i + \frac{\epsilon(1-2\nu_b)}{8n}\right] \geq 1-\delta/2.$$

We now assume that at least m positive examples are obtained in step 2 and that all the $\hat{p}_i$'s are within the above tolerance (these conditions are both satisfied with probability at least $1-\delta$). Then:

1. The estimate $\hat{\nu}$ of ν is accurate; in particular

$$\nu - \frac{\epsilon(1-2\nu_b)}{8n} < \hat{\nu} < \nu + \frac{\epsilon(1-2\nu_b)}{8n}.$$

2. Any literal that is in the target monomial will be placed in the algorithm's hypothesis. Recall that for any literal i that is in the target monomial, $p_i = \nu$ and thus

$$\hat{p}_i < \hat{\nu} + \frac{\epsilon(1-2\nu_b)}{4n}.$$

   Hence for every literal i that is in the target monomial, $\hat{p}_i$ will satisfy the inequality in step 5.

Inputs: ϵ, δ, ν_b (where $\nu \leq \nu_b < 1/2$), access to URA$_\nu$.
Output: Some monomial (possibly "FALSE").

1. Set
$$m = \frac{32n^2}{(1-2\nu_b)^2\epsilon^2} \ln \frac{8n}{\delta}.$$

2. Draw $m' = \max\left\{\frac{2m}{\epsilon}, \frac{2}{\epsilon^2}\ln\frac{3}{\delta}\right\}$ examples from URA$_\nu$.
   - If m positive instances are not obtained, halt and output "FALSE."
   - Otherwise, let S be a sample of m positive examples.

3. For each literal i
   $\hat{p}_i =$ (number of times literal i is off in $S)/m$.

4. $\hat{\nu} = \min_i \hat{p}_i$.

5. Output the AND of all literals i such that
$$\hat{p}_i \leq \hat{\nu} + \frac{\epsilon(1-2\nu_b)}{4n}.$$

Figure 1: Algorithm for pac learning monomials under random attribute noise with an unknown noise rate.

3. No literal that would frequently cause the algorithm's hypothesis to be false when the target monomial is true will be included in the hypothesis. More specifically, we argue that for any literal i that is *not* in the target monomial but is placed in the algorithm's hypothesis $q_i < \epsilon/2n$.

   To prove this we will show that if $q_i \geq \epsilon/2n$ then $\hat{p}_i$ will not satisfy the inequality in step 5. Applying equation (2) we get

$$p_i \geq \nu + \frac{\epsilon(1-2\nu_b)}{2n}.$$

   Finally since $\hat{p}_i$ and $\hat{\nu}$ are within the given tolerance of their true values, it follows that:

$$\hat{p}_i > \hat{\nu} + \frac{\epsilon(1-2\nu_b)}{4n}.$$

   Thus the choice of literals made by the algorithm in step 5 ensures that every literal i in the output formula has $q_i < \epsilon/2n$.

   Now since the literals in the output monomial are a superset of the literals in the target monomial, the algorithm's hypothesis is false whenever the target concept is false. Since every literal i in the output has $q_i < \epsilon/2n$, and there are at most $2n$ literals in the output formula, the probability that the output is false when the target concept is true is at most ϵ.

□

Remark: We can obtain a very similar result for a noise model where with probability $1-\nu$ the example is noise free, and with probability ν a single one of the n bits is picked at random and flipped.

5 Nonuniform random attribute noise

In this section we show that nonuniform random attribute noise makes pac learning almost as difficult as malicious noise. Kearns and Li [6] showed that for any nontrivial concept class, it is impossible to pac learn to accuracy ϵ with examples from MAL_ν unless

$$\nu < \epsilon/(1+\epsilon).$$

Our result for nonuniform random attribute noise is similar, with a slightly weaker bound.

Labeled Instance	$\mathbf{D}_1$	$\mathbf{D}_2$
$(00, -)$	$(1-\nu)/2$	$(1-\nu)/2$
$(00, +)$	0	0
$(01, -)$	$\nu/2$	0
$(01, +)$	0	$\nu/2$
$(10, -)$	0	$\nu/2$
$(10, +)$	$\nu/2$	0
$(11, -)$	0	0
$(11, +)$	$(1-\nu)/2$	$(1-\nu)/2$

Table 1: Two induced noise-free probability distributions obtained from distribution $\mathbf{D}$ for concept class C.

Theorem 2 *Let C be any concept class over the domain $\{0,1\}^n$ that includes the concepts x_i and x_j for some $i \neq j$. It is possible to pac learn C to accuracy ϵ with examples from NRA$_\nu$ only if $\nu < 2\epsilon$.*

Proof: We use the method of induced distributions [6].

Say C contains the concepts x_1 and x_2. In what follows, we will put zero probability weight on instances containing 1's in positions 3 through n, and thus may assume that $\nu_k = 0$ for $3 \leq k \leq n$. In fact, we can assume that our entire instance space is $\{00, 01, 10, 00\}$, since all instances seen will have 0's in all other attributes.

Fix some value of ν in the range $0 \leq \nu < 1/2$. Consider a distribution $\mathbf{D}$ which assigns weight $(1-\nu)/2$ to 00 and 11 and weight $\nu/2$ to 01 and 10. In Table 1 we show the two noise-free probability distributions on examples obtained by labeling the instances drawn from $\mathbf{D}$ according to concept x_1 or concept x_2.

Now consider what happens under the following two learning problems:

1. For the first learning problem let x_1 be the target concept, let $\mathbf{D}$ be the distribution on instances (so $\mathbf{D}_1$ is the noise-free distribution on examples), and let the noise oracle be NRA$_\nu$ with $\nu_1 = \nu$ and $\nu_2 = 0$.

2. For the second learning problem, let x_2 be the target concept, let $\mathbf{D}$ be the distribution (so $\mathbf{D}_2$ is the noise-free distribution on examples), and let the noise oracle be NRA$_\nu$ with $\nu_1 = 0$ and $\nu_2 = \nu$.

These two learning problems have an identical probability distribution on the observed (noisy) samples. Therefore, no pac learning algorithm has any basis for distinguishing between these two scenarios. Thus the learning algorithm must output a concept c such that:

$$\mathbf{D}(c_1 \triangle c) < \epsilon \text{ and } \mathbf{D}(c_2 \triangle c) < \epsilon.$$

By the triangle inequality it follows that:

$$\mathbf{D}(c_1 \triangle c) + \mathbf{D}(c_2 \triangle c) \geq \mathbf{D}(c_1 \triangle c_2).$$

Thus the learning algorithm must a concept c such that $\mathbf{D}(c_1 \triangle c_2) < 2\epsilon$. Finally, since $\mathbf{D}(c_1 \triangle c_2) = \nu$, no learning algorithm can succeed unless $\nu < 2\epsilon$. □

Remark: Notice that the bound of Theorem 2 is an information-theoretic representation-independent hardness result. No algorithm, regardless of either its sample size or computation time, can escape this bound. Furthermore, we have made no assumptions on the representation class from which the hypothesis may be selected—this bound on the tolerable noise rate holds for *any* hypothesis the learner may output.

6 Final thoughts

We have studied the robustness of pac learning algorithms under several forms of random attribute noise. We presented a new algorithm for learning monomials that tolerates large amounts of uniform random attribute noise (any noise rate less than 1/2), and only requires some upper bound on the noise rate as input. An intriguing open question is whether one can pac learn k-DNF formulas under uniform random attribute noise for an *unknown* but bounded noise rate.

However, we feel that our negative result for the more realistic nonuniform random attribute noise oracle makes it clear that, in general, under the pac learning model random attribute noise is quite harmful. This result was surprising to the authors. One expects to only be able to tolerate a small amount of truly malicious noise—it is obviously the worst sort of noise possible. Yet, one would expect that labeling noise would be worse than random attribute noise. Indeed, in one empirical test (of the ID-3 system), that is exactly what was found [9]. Yet, in spite of both these empirical results and our intuition, we have shown that in the pac model random attribute noise (when it is nonuniform) is significantly more harmful than random labeling noise. In fact, nonuniform random attribute noise is significantly more harmful than *malicious* labeling noise generated by a powerful adversary [11], and nearly as harmful as truly malicious noise.

Acknowledgments

We would like to thank Les Valiant for suggesting this line of research. We thank Dana Angluin for pointing out an error (and providing a correction to this error) in our original statement of Theorem 2.

References

[1] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

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

[3] David Haussler, Michael Kearns, Nick Littlestone, and Manfred K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*. To appear.

[4] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[5] Michael Kearns. Thoughts on hypothesis boosting. (Unpublished), December 1988.

[6] Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, Chicago, Illinois, May 1988.

[7] Michael Kearns, Ming Li, Leonard Pitt, and Leslie Valiant. On the learnability of boolean formulae. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, pages 285–295, New York, New York, May 1987.

[8] Philip D. Laird. *Learning from Good and Bad Data*. Kluwer international series in engineering and computer science. Kluwer Academic Publishers, Boston, 1988.

[9] J. Ross Quinlan. The effect of noise on concept learning. In *Machine Learning, An Artificial Intelligence Approach (Volume II)*, chapter 6, pages 149–166. Morgan Kaufmann, 1986.

[10] George Shackelford and Dennis Volper. Learning k-DNF with noise in the attributes. In *First Workshop on Computatinal Learning Theory*, pages 97–103, Cambridge, Mass. August 1988. Morgan Kaufmann.

[11] Robert H. Sloan. Types of noise in data for concept learning. In *First Workshop on Computational Learning Theory*, pages 91–96. Morgan Kaufmann, 1988.

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

[13] Leslie G. Valiant. Learning disjunctions of conjunctions. In *Proceedings IJCAI-85*, pages 560–566. International Joint Committee for Artificial Intelligence, Morgan Kaufmann, August 1985.