# The Optimal PAC Algorithm

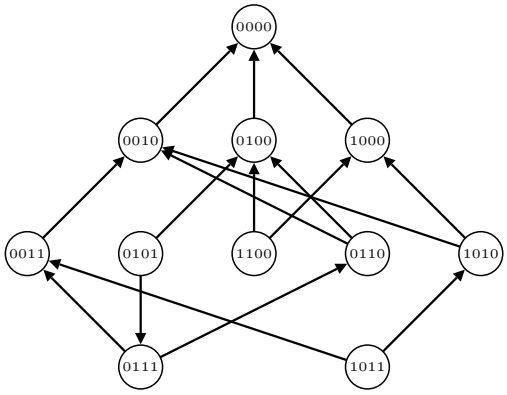
Manfred K. Warmuth

UC Santa Cruz

Assume we are trying to learn a concept class $C$ of VC dimension $d$ with respect to an arbitrary distribution. There is PAC sample size bound that holds for any algorithm that always predicts with some consistent concept in the class $C$ (BEHW89): $O(\frac{1}{\epsilon}(d \log \frac{1}{\epsilon} + \log \frac{1}{\epsilon}))$, where $\epsilon$ and $\delta$ are the accuracy and confidence parameters. Thus after drawing this many examples (consistent with any concept in $C$), then with probability at least $1 - \delta$, the error of the produced concept is at most $\epsilon$. Here the examples are drawn with respect to an arbitrary but fixed distribution $D$, and the accuracy is measured with respect to the same distribution. There is also a lower bound that holds for any algorithm (EHKV89): $\Omega(\frac{1}{\epsilon}(d + \log \frac{1}{\delta}))$. It means that at least this many examples are required for any algorithm to achieve error at most $\epsilon$ with probability at least $1 - \delta$. The lower bound is realized by distributions on a fixed shattered set of size $d$.

**Conjecture**: The one-inclusion graph algorithm of HLW94 always achieves the lower bound. That is after receiving $O(\frac{1}{\epsilon}(d + \log \frac{1}{\epsilon}))$ examples, its error is at most $\epsilon$ with probability at least $1 - \delta$.

The one-inclusion graph for a set of $t+1$ unlabeled examples uses the following subset of the $(t + 1)$-dimensional hypercube as its vertex set: all bit patterns in $\{0, 1\}^{t+1}$ produced by labeling the $t+1$ examples with a concept in $C$. There is an edge between two patterns if they are adjacent in the hypercube (i.e. Hamming distance one).




An orientation of a one-inclusion graph is an orientation of its edges so that the maximum out-degree of all the vertices is minimized. In HLW94 it is shown how to do this using a network flow argument. The minimum maximum out-degree can be shown to be at most at most $d$, the VC dimension of $C$.

J. Shawe-Taylor and Y. Singer (Eds.): COLT 2004, LNAI 3120, pp. 641–642, 2004.

The one-inclusion graph algorithm is formulated as a prediction algorithm: When given $t$ examples labeled with a concept in $C$ and one more unlabeled example, the algorithm produces a binary prediction on the unlabeled example.[1] How does this algorithm predict? It creates and orients the one-inclusion graph for all $t+1$ examples. If there is a unique extension of the $t$ labeled examples to a labeling of the last example, then the one-inclusion graph algorithm predicts with that labeling. However, if there are two labels possible for the unlabeled example (i.e. the unlabeled example corresponds to an edge), then the algorithm predicts with the label of the bit pattern at the head of the oriented edge.

The expected error[2] of the one-inclusion graph algorithm is at most $\frac{d}{t+1}$ (HLW94), and it has been shown that this bound is within a factor of $1+o(1)$ of optimal (LLS02). On the other hand, predicting with an arbitrary consistent hypothesis, can lead to an expected error of $\Omega(\frac{d\log n}{t}\log\frac{t}{d}))$ (HLW94). So in this open problem we conjecture that the one-inclusion algorithm is also optimal in the PAC model.

For special cases of intersection closed concept classes, the *closure algorithm* has been shown to have the optimum $O(\frac{1}{\epsilon}(d+\log\frac{1}{\delta}))$ bound (AO04). This algorithm is can be seen as an instantiation of the one-inclusion graph algorithm (the closure algorithm predicts with an orientation of the one-inclusion graph with maximum out-degree at most $d$). There are cases that show that the upper bound of $O(\frac{1}{\epsilon}(d\log\frac{1}{\epsilon}+\log\frac{1}{\delta}))$ that holds for any algorithm that predicts with a consistent hypothesis cannot be improved (e.g. AO04). However all such cases that we are aware of seem to predict with orientations of the one-inclusion graph that have unnecessarily high out-degree.

## References

P. Auer and R. Ortner. A new PAC bound for intersection-closed concept classes. Appearing concurrently in this COLT 2004 proceedings.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.

D. Haussler, N. Littlestone, and M. 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.

Y. Li, P. M. Long, and A. Srinivasan. The one-inclusion graph algorithm is near optimal for the prediction model of learning. *Transaction on Information Theory*, 47(3):1257–1261, 2002.

[1] Prediction algorithms implicitly represent hypotheses. For any fixed set of $t$ labeled examples, the predictions on the next unlabeled example define a hypothesis. However, as for the algorithm discussed here, this hypothesis is typically not in $C$.

[2] This is the same as the probability of predicting wrong on the unlabeled example.