

# Some Learning Theory Results

## General Bounds on the Number of Examples

As in the book (pp. 669ff), I will use the following notation and assumptions.

Let  $\mathbf{X}$  be the set of all possible examples.

Let  $D$  be the probability distribution from which examples are drawn.

Let  $\mathbf{H}$  be the set of possible hypotheses, i.e., the learning algorithm is guaranteed to return a hypothesis from  $\mathbf{H}$ .

Let  $m$  be the number of examples in the training set, where each example is independently drawn from  $\mathbf{X}$  according to  $D$ .

**Theorem 1** *If the algorithm finds a hypothesis  $h \in \mathbf{H}$  with zero error on the training examples, and if*

$$m \geq \frac{1}{\epsilon} \ln \frac{|\mathbf{H}|}{\delta}$$

*then, with probability at least  $1 - \delta$ , the error of  $h$  on the distribution  $D$  is at most  $\epsilon$ .*

**Proof:** If a hypothesis  $h$  has error greater than  $\epsilon$  on the distribution  $D$ , then the probability that  $h$  will be consistent with the  $m$  examples is less than  $(1 - \epsilon)^m$ . The number of such hypotheses is at most  $|\mathbf{H}|$ . The probability of a disjunction of events is less than or equal than the sum of the probabilities of the events. Thus, the probability that some hypothesis with error more than  $\epsilon$  is consistent with the examples is less than  $|\mathbf{H}|(1 - \epsilon)^m$ . If this value is less than or equal to  $\delta$ , then the theorem is proved. In the following sequence of inequalities, the previous inequality implies the following inequality. This uses the inequality  $\ln(1 + x) \leq x$ .

$m \geq \frac{1}{\epsilon} \ln \frac{ \mathbf{H} }{\delta} = \frac{1}{\epsilon} (\ln  \mathbf{H}  - \ln \delta)$	because $\ln(xy) = (\ln x) + (\ln y)$
$\epsilon m \geq \ln  \mathbf{H}  - \ln \delta$	multiply both sides by $\epsilon$
$-\epsilon m \leq \ln \delta - \ln  \mathbf{H} $	negating both sides reverses the inequality
$m \ln(1 - \epsilon) \leq \ln \delta - \ln  \mathbf{H} $	because $\ln(1 + x) \leq x$
$\ln(1 - \epsilon)^m \leq \ln \delta - \ln  \mathbf{H} $	because $n \ln x = \ln x^n$
$\ln  \mathbf{H}  + \ln(1 - \epsilon)^m \leq \ln \delta$	add $\ln  \mathbf{H} $ to both sides
$\ln  \mathbf{H} (1 - \epsilon)^m \leq \ln \delta$	because $(\ln x) + (\ln y) = \ln(xy)$
$ \mathbf{H} (1 - \epsilon)^m \leq \delta$	because $\ln x \leq \ln y$ implies $x \leq y$

**End Proof.**

Often, it will be the case that there will no hypothesis in  $\mathbf{H}$  that correctly classifies all the training examples (or it will be very hard to find). Based on inequalities that are referred to as Chernoff Bounds or Hoeffding's Inequality, it is possible to show the following theorem.

**Theorem 2** *If the algorithm finds a hypothesis  $h \in \mathbf{H}$  with  $\alpha$  error on the training examples, and if*

$$m \geq \frac{1}{2\epsilon^2} \ln \frac{2|\mathbf{H}|}{\delta}$$

*then, with probability at least  $1 - \delta$ , the error of  $h$  on the distribution  $D$  is at most  $\alpha + \epsilon$ .*

In many cases, such as perceptrons and neural networks discussed in Chapter 20, the set of hypotheses is infinite. When this happens, there is a quantity called the Vapnik-Chervonenkis dimension that can be used to bound the number of examples. Usually, the VC dimension closely corresponds to the number of free parameters. Very roughly, this number takes the place of the  $\ln |\mathbf{H}|$  term in the above inequalities. For example, the VC dimension of the perceptron is the number of attributes plus one. The VC dimension of a neural network is  $O(w \ln w)$ , where  $w$  is the number of weights.

Two good books for looking at these issues in more detail are: M. Anthony and M. Briggs (1992), *Computational Learning Theory*, Cambridge University Press; and M. J. Kearns and U. V. Vazirani (1994), *An Introduction to Computational Learning Theory*, MIT Press.

## Bound for Decision Lists

Here, I demonstrate a bound on the number of decision list hypotheses. This bound can be substituted into the above theorems to bound the number of examples needed for learning decision lists.

**Theorem 3** *The size of the hypothesis space of 1-DL( $n$ ) (decision lists over  $n$  binary attributes where each test includes 1 attribute) is bounded by:*

$$|1\text{-DL}(n)| \leq (4n)^{2n}$$

**Proof:** Each node in a 1-DL( $n$ ) decision list consists of one literal mapped to either positive or negative. To construct a node, first choose one of the  $n$  attributes, followed by choosing whether to negate the attribute, followed by choosing whether to map to positive or negative. This leads to a total of  $4n$  possible nodes.

A 1-DL( $n$ ) decision list does not need to be longer than  $2n$  because there are only  $2n$  literals. There are  $4n$  choices for each of at most  $2n$  nodes; thus, we have

$$|1\text{-DL}(n)| \leq (4n)^{2n}$$

**End Proof.**

Note that  $\ln(4n)^{2n} = 2n \ln(4n)$ , so substituting  $2n \ln(4n)$  for  $\ln |\mathbf{H}|$  in the above theorems leads to upper bounds on the number of examples needed for learning decision lists.