PAC learnability of a finite concept class

An example

To see that PAC learning is a "natural" way of addressing learnablity when the concept class has a finite numbers of concepts we consider a specific case in which there are three examples e_a, e_b, e_c, and each example can be either positive or negative. In addition, the set H of hypotheses contains only five hypotheses:

$$h_1 = \left\{ \begin{array}{cc} e_a & + \\ e_b & + \\ e_c & + \end{array} \right\},\ h_2 = \left\{ \begin{array}{cc} e_a & + \\ e_b & + \\ e_c & - \end{array} \right\},\ h_3 = \left\{ \begin{array}{cc} e_a & + \\ e_b & - \\ e_c & + \end{array} \right\},\ h_4 = \left\{ \begin{array}{cc} e_a & + \\ e_b & - \\ e_c & - \end{array} \right\},\ h_5 = \left\{ \begin{array}{cc} e_a & - \\ e_b & - \\ e_c & + \end{array} \right\}$$

Let c, the target concept, and D, the probability distribution on the examples of c be:

$$c = \left\{ \begin{array}{cc} e_a & + \\ e_b & - \\ e_c & + \end{array} \right\} \qquad \begin{array}{l} D(e_a) = 0.45 \\ D(e_b) = 0.54 \\ D(e_c) = 0.01 \end{array}$$

Observe the following:

The probability that h_1 wrongly classifies a randomly chosen example is 0.54.
The probability that h_2 wrongly classifies a randomly chosen example is 0.55.
The probability that h_3 wrongly classifies a randomly chosen example is 0.
The probability that h_4 wrongly classifies a randomly chosen example is 0.01.
The probability that h_5 wrongly classifies a randomly chosen exaple is 0.45.

Given a set of m training examples $e_1, \ldots e_m$ of the target concept c, a learning algorithm will produce as output a hypothesis $h \in H$ that correctly classifies all the training examples. Such an algorithm is called *a consistent algorithm*. We are interested in the guaranteed performance of a consistent algorithm, that is its ability to generalize and correctly classify future examples.

QUESTION: What should be m to guarantee that h has an error of less than 0.1 in classifying future examples ($\epsilon = 0.1$)?

ANSWER: Since we do not know how the examples are chosen, an arbitrary large (finite) set of training examples may not suffice. For example, all m examples can be the same as e_a.

QUESTION: What should be m to guarantee that h has an error of less than 0.1 in classifying future examples ($\epsilon = 0.1$) if it is known that the examples are drawn at random according to the probability distribution D?

ANSWER: Since it is possible that the chosen sample of examples is uncharacteristic of the distribution D, (for example, all m examples can be the same as e_a), we must conclude that an arbitrary large (finite) set of training examples may not suffice.

QUESTION: What should be m to guarantee with confidence of at least 95% ($\delta = 0.05$) that h has an error of less than 0.1 in classifying future examples ($\epsilon = 0.1$) if it is known that the examples are drawn at random according to the probability distribution D?

ANSWER: In this case we can, indeed, compute m. To see how, first notice that the requirement of $\epsilon = 0.1$ means that the learned hypothesis should be either h_3 or h_4. We have to choose m such that the chosen hypothesis is h_1, h_2, or h_5 with probability of less than δ. Observe the following:

The probability that h_1 correctly classifies 1 randomly chosen example is $(1 - 0.54)$.
The probability that h_1 correctly classifies 2 randomly chosen examples is $(1 - 0.54)^2$.
The probability that h_1 correctly classifies m randomly chosen examples is $(1 - 0.54)^m$.

Similarly:

The probability that h_2 correctly classifies m randomly chosen examples is $(1 - 0.55)^m$.
The probability that h_5 correctly classifies m randomly chosen examples is $(1 - 0.45)^m$.

Using the fact that $Prob(A \vee B \vee C) \leq Prob(A) + Prob(B) + Prob(C)$ we observe that:

The probability that either h_1 or h_2 or h_5 correctly classifies m randomly chosen examples is at most $(1-0.54)^m + (1-0.55)^m + (1-0.45)^m$.

Since the answer to our problem requires that h_1, h_2, h_5 are chosen with probability of at most δ we can determine m by solving for m such that:

$$0.46^m + 0.45^m + 0.55^m < \delta = 0.05.$$

Since solving the above inequality is difficult, we solve a simpler inequality which may give a slightly larger value for m:

$$3 \cdot 0.55^m < 0.05$$

(Notice that a solution to this inequality is also a solution to the other inequality.)
Therefore, we can choose m such that:

$$m > \frac{1}{\log \frac{1}{0.55}}(\log \frac{1}{0.05} + \log 3) = 6.8$$

So the number of examples m should be 7. Similar calculations can be done even if D is unknown. The results are given by the following theorem:

Theorem: A consistent algorithm that produces a hypothesis h from a class of r hypotheses needs no more than m examples to achieve accuracy of $1-\epsilon$ with a confidence of $1-\delta$, where m is determined by:

$$m \geq \frac{1}{\epsilon}\left(\ln(r) + \ln(\frac{1}{\delta})\right) = \frac{1}{\epsilon}\ln(r/\delta).$$

Proof: Without loss of generality let $h_1, \ldots, h_k$ be the hypotheses with error larger than ϵ. The probability that any one of these hypotheses is consistent with m randomly chosen examples is less than $(1-\epsilon)^m$. The probability that there is a hypothesis from $h_1, \ldots, h_k$ consistent with m randomly chosen examples is less than $k(1-\epsilon)^m$. Since $k \leq r$ it follows that it is enough to choose m such that:

$$r(1-\epsilon)^m \leq \delta,$$

which implies that

$$m \geq \frac{1}{-\log(1-\epsilon)}\left(\log(r) + \log(\frac{1}{\delta})\right)$$

To prove the theorem it is enough to show that $\frac{1}{\epsilon} \geq \frac{1}{-\ln(1-\epsilon)}$, which is the same as showing that $f(\epsilon) = e^{-\epsilon} + \epsilon - 1$ is nonnegative for $\epsilon \geq 0$. And this follows from the fact that f is monotone increasing, with $f(0) = 0$. QED