



Why

We want to talk about an algorithm which can learn from a hypothesis class, regardless of the underlying distribution.

Definition

Let (X, \mathcal{X}) and (Y, \mathcal{Y}) be measurable input and output spaces.

A hypothesis class \mathcal{H} of measurable functions from X to Y is *probably approximately correct learnable* (or *PAC learnable*) if (a) there exists an inductor $\mathcal{A} : (X \times Y)^n \rightarrow \mathcal{H}$, so that (b) for every underlying measure μ and labeling function $f : X \rightarrow Y$ (c) for every $\varepsilon, \delta \in (0, 1)$ (d) there exists $m_0 \in \mathbf{N}$ so that (e) for all $m \geq m_0$

$$\mu^m [\mu [\mathcal{A}((x_i, f(x_i))_{i=1}^n)(\xi) \neq f(\xi)] \leq \varepsilon] \geq 1 - \delta \quad (1)$$

where $x \in X^m$ and $\xi \in X$. In this case we say that the inductor (or learning algorithm) \mathcal{A} *PAC learns* \mathcal{H} .

We interpret this as follows: “no matter the underlying distribution and correct labeling function, if someone specifies an accuracy and confidence we can tell them the number of samples they need so that the inductor outputs a hypothesis which is probably approximately correct.”

Some authors require that the hypothesis class be realizable with respect to the underlying distribution and correct labeling function. This is natural, because if the hypothesis class includes the correct labeling function f , then it is realizable. In this case they refer to the above definition as the *agnostic PAC model*. We emphasize here that there this definition includes the notion of realizability. In other words. We emphasize again that this definition contains to “approximation parameters.” The accuracy parameter ε corresponds to the “approximately correct” piece and the confidence parameters δ corresponds to the “probably” piece.

Sample complexity

Note that the existence of an m_0 above for each ε and δ is equivalent to requiring that there exists $m_0 : (0, 1)^2 \rightarrow \mathbf{N}$ so that for all $m \geq m_0(\varepsilon, \delta)$ the condition in Equation 1 holds. There may exist multiple such m_0 , so we define $\tilde{m} : (0, 1)^2 \rightarrow \mathbf{N}$ so that $\tilde{m}(\varepsilon, \delta)$ is the smallest integer so that Equation (1) holds. We call \tilde{m} the *sample complexity*. Clearly it is a function of ε and δ . It also depends on the hypothesis class \mathcal{H} and the learning algorithm \mathcal{A} .

