



Why

We want to talk about an inductor's error under a probabilistic supervised data model.

Definition

Let (X, \mathcal{X}, μ) and $f : X \rightarrow Y$ be a probabilistic supervised data model. Let $\mathcal{H} \subset (X \rightarrow Y)$ be a hypothesis class.

Let $\varepsilon, \delta \in (0, 1)$. An inductor $A : (X \times Y)^n \rightarrow \mathcal{H}$ is (ε, δ) -*probably approximately correct* for μ and f if

$$\mu^n \left[\text{error}_{\mu, f} (A((x_i, f(x_i))_{i=1}^n)) \leq \varepsilon \right] \geq 1 - \delta. \quad (1)$$

To show that an algorithm is probably approximately correct, we bound by δ the probability of the inductor producing a predictor which exceeds ε error.

A hypothesis class \mathcal{H} is *probably approximately correct learnable* (or *PAC learnable*) if (a) for every underlying measure μ , labeling function $f : X \rightarrow Y$, error $1 - \varepsilon$ and confidence $1 - \delta$, (b) there exists $m_0 \in \mathbf{N}$ and a sequence of inductors $(A^m : (X \times Y)^m \rightarrow \mathcal{H})_{m=m_0}^\infty$ so that (c) A^m is (ε, δ) -probably approximately correct for all $m \geq m_0$. If the inductors are subfamilies of a family of inductors $(A^n)_{n \in \mathbf{N}}$ then we say that this family of inductors *PAC learns* \mathcal{H} .

We interpret these conditions as follows. No matter the underlying distribution and labeling function, given an accuracy and confidence, we can specify the number of samples required to make the algorithm “succeed.” By which we mean that (“with high probability”) the predictor’s error will be “small.”

Sample complexity

If a family of inductors $\mathcal{A} = (A_n)_n$ PAC learns \mathcal{H} , then condition (b) above is equivalent to the existence function mapping (ε, δ) to \mathbf{N} . Let

m_0^* be the smallest integer such that condition (b) holds. We call the function $(\varepsilon, \delta) \mapsto m_0^*(\varepsilon, \delta)$ the *sample complexity* of the family \mathcal{A} on the hypothesis class \mathcal{H} .

Other terminology

For (what seem to be) historical reasons, some authors refer to the definition above as *agnostic PAC learning* since it drops a common condition (developed further in these sheets) relating the hypothesis class to the correct labeling function f .

