

Agnostic Learning

We now turn to what is known as *agnostic learning* (or *unrealizable learning*). In this model, the target function $f : X \rightarrow \{0, 1\}$ is unknown, as in the “standard” PAC model, but furthermore, nothing is known about the class F that f belongs to. As in the standard PAC model, there is also an unknown distribution D over X , and the learning algorithm receives examples $(x, f(x))$, where each $x \in X$ is selected independently according to D . Here too the algorithm receives two parameters, ϵ and δ , and is required to output a hypothesis h from a hypothesis class H .

Let

$$\epsilon_{f,D}^{\text{opt}}(H) \stackrel{\text{def}}{=} \min_{g \in H} \{\text{err}_{f,D}(g)\} \quad (1)$$

be the minimum error (with respect to the target function f and the underlying distribution D) that any function g in H can achieve. (Note that if $f \in H$ as in the standard model, then $\epsilon_{f,D}^{\text{opt}}(H) = 0$). The goal of the agnostic learning algorithm is, with probability at least $1 - \delta$, to output a hypothesis $h \in H$, such that

$$\text{err}_{f,D}(h) \leq \epsilon_{f,D}^{\text{opt}}(H) + \epsilon \quad (2)$$

We next prove a theorem analogous to Occam’s razor in this model (and in particular assume that H has finite size).

Theorem 1 *Let A be an algorithm, that given any labeled sample, finds a function $h \in H$ that minimizes the empirical error on the sample. That is, given a sample $S = \{(a^1, b^1), \dots, (a^m, b^m)\}$, the algorithm A finds a hypothesis $h \in H$ for which $\hat{\epsilon}_S(h) \stackrel{\text{def}}{=} \frac{|\{i: h(a^i) \neq b^i\}|}{m}$ is minimized.*

If we give A a sample of size $m \geq \frac{2}{\epsilon^2} \ln(|H|/\delta)$, that is distributed according to an unknown distribution D , and labeled by an arbitrary unknown function f , then, with probability at least $1 - \delta$, the hypothesis h satisfies Equation (2).

Before we prove the theorem, we dwell a bit on its implications. On one hand, this is a strong statement: No matter what the function is, if you have a way of minimizing the empirical error then you have an agnostic learning algorithm where the sample complexity of the algorithms is polynomial in $1/\epsilon$ and logarithmic in the size of $|H|$. What are the caveats? Similarly to the discussion we had concerning the “original” Occam’s razor, one possible difficulty with applying the theorem, is that H may not be finite. Here too the VC-dimension of H can replace $\ln |H|$ (roughly). As already noted in the case of the noisy version of Occam, a more major problem is a computational one. In many cases, there is no efficient algorithm for minimizing the empirical error (that is, the problem is NP-hard).

To prove the theorem, we define for every function $g \in H$, and for every $1 \leq j \leq m$ a random variable χ_g^j which equals 1 if $g(a^j) \neq f(a^j)$, and is 0 otherwise (recall that a^j denotes the j^{th} example

in the sample. By definition of the χ_g^j 's, $\hat{\epsilon}_S(g) = \frac{1}{m} \sum_{j=1}^m \chi_g^j$. Let $h^{\text{opt}} \in H$ be a hypothesis in H such that $\text{err}_{f,D}(h^{\text{opt}}) = \epsilon_{f,D}^{\text{opt}}(H)$. That is, h^{opt} is an optimal hypothesis in H in terms of approximating f with respect to D (if there is more than one such optimal hypothesis, then h^{opt} is selected arbitrarily). Let

$$B_{f,D,\epsilon} = \left\{ g \in H : \text{err}_{f,D}(g) > \epsilon_{f,D}^{\text{opt}}(h) + \epsilon \right\} \quad (3)$$

be the collection of *bad* hypotheses in H (that is, those we do not want our algorithm to select).

We next show that with probability at least $1 - \delta$, we have:

1. $\frac{1}{m} \sum_{j=1}^m \chi_{h^{\text{opt}}}^j \leq \epsilon_{f,D}^{\text{opt}} + \epsilon/2$;
2. $\frac{1}{m} \sum_{j=1}^m \chi_g^j > \epsilon_{f,D}^{\text{opt}} + \epsilon/2$, for every $g \in B_{f,D,\epsilon}$.

It directly follows that with probability at least $1 - \delta$, the empirical error of every bad hypothesis is larger than the empirical error of the optimal one in H , and hence the hypothesis selected (for minimizing the empirical error) is necessarily not bad, as required (though of course it is not necessarily an optimal one).

Consider the first item. By definition of h^{opt} , we have that $\Pr[\chi_{h^{\text{opt}}}^j = 1] = \epsilon_{f,D}^{\text{opt}}$. By an additive Chernoff bound,

$$\Pr \left[\frac{1}{m} \sum_{i=1}^m \chi_{h^{\text{opt}}}^i > \epsilon_{f,D}^{\text{opt}} + \epsilon/2 \right] \leq e^{-2m(\epsilon/2)^2} \leq \delta/|H| \quad (4)$$

We turn to the second item. For any given $g \in B_{f,D,\epsilon}$, we have that (for every j), $\Pr[\chi_g^j = 1] > \epsilon_{f,D}^{\text{opt}} + \epsilon$. By an additive Chernoff bound,

$$\Pr \left[\frac{1}{m} \sum_{i=1}^m \chi_g^i \leq \epsilon_{f,D}^{\text{opt}} + \epsilon/2 \right] \leq \Pr \left[\frac{1}{m} \sum_{j=1}^m \chi_g^j < \text{err}_{f,D}(g) - \epsilon/2 \right] \quad (5)$$

$$\leq e^{-2m(\epsilon/2)^2} = \delta/|H| \quad (6)$$

The probability that both $\frac{1}{m} \sum_{j=1}^m \chi_{h^{\text{opt}}}^j \leq \epsilon_{f,D}^{\text{opt}} + \epsilon/2$, and $\frac{1}{m} \sum_{j=1}^m \chi_g^j > \epsilon_{f,D}^{\text{opt}} + \epsilon/2$, for every $g \in B_{f,D,\epsilon}$, equals 1 minus the probability that either $\frac{1}{m} \sum_{j=1}^m \chi_{h^{\text{opt}}}^j > \epsilon_{f,D}^{\text{opt}} + \epsilon/2$, or $\frac{1}{m} \sum_{j=1}^m \chi_g^j \leq \epsilon_{f,D}^{\text{opt}} + \epsilon/2$, for some $g \in B_{f,D,\epsilon}$. By a union bound, since $|B_{f,D,\epsilon}| \leq |H| - 1$, the latter probability is at most δ , and we are done.