

Agnostic Learning and Structural Risk Minimization

1 Introduction

While the Occam theorem gives us a powerful tool for proving the PAC learnability of certain concept classes, it is predicated on several requirements which limit its applicability.

In this section we discuss cases going beyond these restrictions; What if the target concept $c \notin H$ or if $|H| = \infty$ or the VC dimension of H is ∞ . This lecture is given in terms of discrete classes for simplicity but similar arguments hold for the case of finite or infinite VCD.

2 Target concept $c \notin H$

If $c \notin H$ it is not certain that we can find a hypothesis consistent with the example set. In this case we must refine our goal to do the best we can with H .

For any fixed concept c , distribution D , and hypothesis h let $err_C(h) = Pr_D[h(x) \neq c(x)]$. Define $h^*(c) \in H$ to be $argmin_{h \in H}(err(h))$, so $h^*(c)$ is the best hypothesis in H when learning c under D .

Definition: Agnostic PAC Learning An algorithm \mathcal{A} Agnostically PAC learns a concept class H if $\forall D, \forall \delta, \forall \epsilon$, and $\forall c$, \mathcal{A} uses IID examples labeled by c and, with probability $1 - \delta$, produces a hypothesis h which has $err(h) \leq err(h^*) + \epsilon$. If \mathcal{A} runs in time polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ then we say \mathcal{A} efficiently agnostically PAC learns \mathcal{A} .

Assuming a finite size of the hypothesis class H , we can define an algorithm analogous to Occam Algorithms. This is known in the literature as “minimizing disagreements” or as “empirical risk minimization” (ERM). In the following explanation we use the notation $\frac{\# \text{ of mistakes on } S}{|S|}$ where S is the sample set used in learning.

Definition: ERM Algorithm An ERM algorithm for Agnostic PAC learning operates by taking a sample and finding $h = argmin_{h \in H}(e\hat{r}(h))$.

Theorem: if \mathcal{A} is an ERM algorithm for H and it uses $m \geq \frac{2}{\epsilon^2} \ln \frac{2|H|}{\delta}$ IID examples, then \mathcal{A} agnostically PAC learns H

Proof: We will prove this theorem through the use of the following lemma.

Lemma 1: if $\forall h \in H |err(h) - \hat{err}(h)| \leq \frac{\epsilon}{2}$ then the hypothesis h output by \mathcal{A} satisfies the condition $err(h) \leq err(h^*) + \epsilon$

Pick any bad hypothesis h_{bad} such that $err(h_{bad}) > err(h^*) + \epsilon$. From the antecedent of Lemma 1 we have that $\hat{err}(h_{bad}) \geq err(h_{bad}) - \frac{\epsilon}{2}$ and $err(h^*) \geq \hat{err}(h^*) - \frac{\epsilon}{2}$ implying that $\hat{err}(h_{bad}) \geq err(h_{bad}) - \frac{\epsilon}{2} \geq err(h^*) + \epsilon - \frac{\epsilon}{2} \geq \hat{err}(h^*)$. Since the ERM algorithm picks the hypothesis with the smallest empirical error, it would have then chosen h^* over h_{bad} , and therefore no h defying the error bounds of agnostic PAC learning could be chosen.

All that is required to complete the proof is then to show that Lemma 1 is satisfied with probability $\geq 1 - \delta$. The probability that a single hypothesis violates the condition of Lemma 1 can be bounded using the Chernoff bound:

$$Pr[|err(h) - \hat{err}(h)| \geq \frac{\epsilon}{2}] < 2e^{-2m(\frac{\epsilon}{2})^2}$$

Using $m \geq \frac{2}{\epsilon^2} \ln \frac{2|H|}{\delta}$ gives $Pr < \frac{\delta}{|H|}$ and by the union bound over all $h \in H$ we have a probability of failure $\leq \delta$

The requirement that \mathcal{A} produces the hypothesis which absolutely minimizes the error be produced may be relaxed slightly without losing Agnostic PAC learnability.

Definition: Let h_{best} be the hypothesis from H which has minimum empirical error on a given sample S . An algorithm \mathcal{A} is an Almost ERM algorithm for a concept class H if it produces a hypothesis $h \in H$ which has $\hat{err}(h) \leq \hat{err}(h_{best}) + \frac{\epsilon}{2}$.

Theorem: if \mathcal{A} is an Almost ERM algorithm then \mathcal{A} is an Agnostic PAC learner.

The proof follows the same lines as that for ERM algorithms.

3 Structural Risk Minimization

Another requirement of the Occam Theorem is that the size of the hypothesis class be finite or have finite VC dimension. In this section we consider the case where the hypothesis class is of potentially infinite size and the number of examples is fixed.

First, let us define our hypothesis class as consisting of several subsets of hypotheses H_i indexed such that $H_0 \subseteq H_1 \subseteq H_2 \subseteq \dots$ and $H = \bigcup_0^\infty H_i$. Our goal is to pick a H_i and an $h \in H_i$ as our hypothesis. This classical problem is known as the model selection problem.

Simply selecting the smallest indexed hypothesis with zero empirical error will favor large models, however, and lead to overfitting. To compensate for this we will give each H_i 's best hypothesis a error tolerance ϵ_i which is based on not only $|H_i|$ but also i itself, and pick the one which minimizes $\hat{err}(h_i) + \epsilon_i$. This algorithmic approach can be applied in different contexts and is known as Structural Risk Minimization. In our case we can make the the following choices:

- Let $\epsilon_i = \sqrt{\frac{2}{m} \ln \frac{2|H_i|}{\delta} + \frac{\ln i}{m}}$
- Let h_i be some hypothesis in H_i

- Using the Chernoff bound, we can show that $Pr[|err(h_i) - e\hat{r}(h_i)| > \epsilon_i] \leq \frac{\delta}{2i^2|H_i|}$
- Using a union bound, we can bound the probability that the error estimate for any h_i in any H_i is off by more than ϵ_i . Thus we sum over h_i in H_i , and over i to yield $Pr \leq \delta$.
- Therefore, with probability $\geq 1 - \delta$, for all i , and for all $h_i \in H_i$ $|e\hat{r}(h_i) - err(h_i)| \leq \epsilon_i$.

Assuming this condition holds, the overall quality of the chosen hypothesis can be shown to be close to optimal, similar to the requirements of agnostic learning. Let h^* be the actual best hypothesis, and let the least of the H_i 's containing it be H_M . Let \hat{h}_i refer to the hypothesis with the smallest empirical error in H_i . Finally, suppose the algorithm chooses some $h_k \in H_k$.

- Assuming $\forall h \in H, |err(h) - e\hat{r}(h)| \leq \epsilon_i$ we have that $err(h^*) > e\hat{r}(h^*) - \epsilon_M$.
- Note that by definition of \hat{h}_M , we have $\forall h \in H_M, e\hat{r}(\hat{h}_M) \leq e\hat{r}(h)$. This implies that $err(h^*) \geq e\hat{r}(\hat{h}_M) - \epsilon_M$
- We next note that $[e\hat{r}(\hat{h}_k) + \epsilon_k] \leq [e\hat{r}(\hat{h}_M) + \epsilon_M]$ because SRM chooses the hypothesis which minimizes $e\hat{r}(h_i) + \epsilon_i$.
- Therefore $err(h^*) \geq e\hat{r}(\hat{h}_M) - \epsilon_M = [e\hat{r}(\hat{h}_M) + \epsilon_M] - 2\epsilon_M \geq [e\hat{r}(\hat{h}_k) + \epsilon_k] - 2\epsilon_M \geq err(\hat{h}_k) - 2\epsilon_M$ where in the last step we used again the approximation guarantee (this time for \hat{h}_k).

We have therefore shown the following result.

Theorem: $\forall D \forall \epsilon \forall \delta$ and $\forall c$, given a hypothesis class H let h^* be the best hypothesis in H and let M be the index of the least hypothesis set H_M , then with probability $1 - \delta$ the output of the SRM algorithm h satisfies $err(h) \leq err(h^*) + 2\epsilon_M$