

A Refined Analysis of PAC Learning via the Disagreement Coefficient: Working Notes

Steve Hanneke
Machine Learning Department
Carnegie Mellon University
shanneke@cs.cmu.edu

April 17, 2009

1 Definitions

We will work in the PAC model with concept space \mathbb{C} under distribution \mathcal{D} (over instance space \mathcal{X}). In particular, there is always a target function $f \in \mathbb{C}$ with $er(f) = 0$.

Note that the known general upper bound for this problem is that, if the VC dimension of \mathbb{C} is d , then with probability $1 - \delta$, every classifier in \mathbb{C} consistent with n random samples has error rate at most

$$4 \frac{d \ln(2en/d) + \ln(4/\delta)}{n}. \quad (1)$$

This is due to Vapnik (1982). There is a slightly different bound (for a different learning strategy) of

$$\propto \frac{d \log(1/\delta)}{n} \quad (2)$$

proven by Haussler, Littlestone, and Warmuth (1994). It is also known that one cannot get a bound always smaller than

$$\propto \frac{d + \log(1/\delta)}{n}$$

for any concept space (Vapnik, 1982). The question we are concerned with here is deriving upper bounds that are closer to this lower bound than either (1) or (2) in some cases.

The *disagreement coefficient*, defined in (Hanneke, 2007), and later used in (Dasgupta, Hsu, and Monteleoni, 2007; Balcan, Hanneke, and Wortman, 2008; Hanneke, 2009), is defined as follows.

Definition 1. For any measurable classifier f , and $r > 0$, define

$$B(f, r) = \{h \in \mathbb{C} : \mathbb{P}[h(X) \neq f(X)] \leq r\},$$

and for any $V \subseteq \mathbb{C}$, let

$$DIS(V) = \{x \in \mathcal{X} : \exists h_1, h_2 \in V \text{ s.t. } h_1(x) \neq h_2(x)\}$$

denote the region of disagreement of V . Then the disagreement coefficient of f with respect to \mathbb{C} and \mathcal{D} , denoted θ_f , is defined as¹

$$\theta_f = \sup_{r > r_n} \frac{\mathbb{P}(\text{DIS}(B(f, r)))}{r}.$$

For our purposes, we can take $r_n = \frac{d + \log(1/\delta)}{n}$ (or, for a coarser analysis, we could take $r_n = 0$).

In particular, note that $\theta_f \leq \frac{1}{r_n}$ always. However, it is sometimes much smaller, or even constant.

2 Error Rates for Any Consistent Classifier

For simplicity and to focus on the nontrivial cases, the results in this section will be stated for the case where $\mathbb{P}(\text{DIS}(\mathbb{C})) > 0$. The $\mathbb{P}(\text{DIS}(\mathbb{C})) = 0$ case is trivial, since every $h \in \mathbb{C}$ has $er(h) = 0$ there.

Theorem 1. *Let d be the VC dimension of concept space \mathbb{C} , and let $V_n = \{h \in \mathbb{C} : \forall i \leq n, h(x_i) = f(x_i)\}$, where $f \in \mathbb{C}$ is the target function (i.e., $er(f) = 0$), and $(x_1, x_2, \dots, x_n) \sim \mathcal{D}^n$ is a sequence of i.i.d. training examples. Then for any $\delta \in (0, 1)$, with probability $\geq 1 - \delta$, $\forall h \in V_n$,*

$$er(h) \leq \frac{24}{n} \left(d \ln(880\theta_f) + \ln \frac{12}{\delta} \right). \quad (3)$$

Proof. Since $\mathbb{P}(\text{DIS}(\mathbb{C})) > 0$ by assumption, $\theta_f > 0$ (and $d > 0$ also follows). As above, let $V_m = \{h \in \mathbb{C} : \forall i \leq m, h(x_i) = f(x_i)\}$, and define $\text{radius}(V_m) = \sup_{h \in V_m} er(h)$. We will prove the result by induction on n . As a base case, note that the result clearly holds for $n \leq d$, as we always have $er(h) \leq 1$.

Now suppose $n \geq d + 1 \geq 2$, and suppose the result holds for any $m < n$; in particular, consider $m = \lfloor n/2 \rfloor$. Thus, for any $\delta \in (0, 1)$, with probability $\geq 1 - \delta/3$,

$$\text{radius}(V_m) \leq \frac{24}{m} \left(d \ln(880\theta_f) + \ln \frac{36}{\delta} \right).$$

Note that $r_n < r_m$, so we can take this inequality to hold for the θ_f defined with r_n as well. If $\text{radius}(V_m) \leq r_n$, then we clearly have (3), so suppose $\text{radius}(V_m) > r_n$. Likewise, if $\mathbb{P}(\text{DIS}(V_m)) < \frac{8}{m} \ln \frac{3}{\delta} \leq \frac{24}{n} \ln \frac{3}{\delta}$, then (3) is valid (as is (4) below) since $\text{radius}(V_n) \leq \text{radius}(V_m) \leq \mathbb{P}(\text{DIS}(V_m))$. Otherwise, by a Chernoff bound, with probability $\geq 1 - \delta/3$, we have

$$|\{x_{m+1}, x_{m+2}, \dots, x_n\} \cap \text{DIS}(V_m)| \geq \mathbb{P}(\text{DIS}(V_m)) \lceil n/2 \rceil / 2 =: N.$$

¹ Here and below we use *outer* probabilities (van der Vaart and Wellner, 1996), so that quantities such as $\mathbb{P}(\text{DIS}(B(f, r)))$ are always well-defined, even when $\text{DIS}(B(f, r))$ is not measurable.

Thus, the first N samples in this set that are in $DIS(V_m)$ represent an iid sample from the conditional given $DIS(V_m)$. (1) tells us that given this event, with probability $\geq 1 - \delta/3$,

$$\begin{aligned} radius(V_n) &= \mathbb{P}(DIS(V_m))radius(V_n|DIS(V_m)) \\ &\leq \mathbb{P}(DIS(V_m))\frac{4}{N} \left(d \ln \frac{2eN}{d} + \ln \frac{12}{\delta} \right) \leq \frac{16}{n} \left(d \ln \frac{2e\mathbb{P}(DIS(V_m))n}{4d} + \ln \frac{12}{\delta} \right) \\ &\leq \frac{16}{n} \left(d \ln \frac{e\theta_f radius(V_m)n}{2d} + \ln \frac{12}{\delta} \right). \end{aligned}$$

Applying the inductive hypothesis for $radius(V_m)$ combined with a union bound over these 3 failure events (each of probability $\delta/3$), we have that with probability $\geq 1 - \delta$,

$$radius(V_n) \leq \frac{16}{n} \left(d \ln \left(48e\theta_f \left(\ln(880\theta_f) + \frac{1}{d} \ln \frac{36}{\delta} \right) \right) + \ln \frac{12}{\delta} \right). \quad (4)$$

If $d \geq \frac{1}{e} \ln \frac{12}{\delta}$, then the right side of (4) is at most

$$\begin{aligned} &\frac{16}{n} \left(d \ln (\theta_f 48e \ln(880 \cdot 3 \cdot e^e \theta_f)) + \ln \frac{12}{\delta} \right) \\ &\leq \frac{16}{n} \left(d \ln (\theta_f 48e \ln(40008\theta_f)) + \ln \frac{12}{\delta} \right) \\ &\leq \frac{16}{n} \left(d \ln (26099\theta_f^{3/2}) + \ln \frac{12}{\delta} \right) \leq \frac{24}{n} \left(d \ln(880\theta_f) + \ln \frac{12}{\delta} \right). \end{aligned}$$

Otherwise $d < \frac{1}{e} \ln \frac{12}{\delta}$, so that the right side of (4) is at most

$$\begin{aligned} &\frac{16}{n} \left(d \ln \left(\theta_f 48e \ln(880 \cdot 3\theta_f) \frac{1}{d} \ln \frac{12}{\delta} \right) + \ln \frac{12}{\delta} \right) \\ &\leq \frac{16}{n} \left(d \ln (6705\theta_f^{3/2}) + d \ln \left(\frac{1}{d} \ln \frac{12}{\delta} \right) + \ln \frac{12}{\delta} \right) \\ &\leq \frac{24}{n} \left(d \ln(356\theta_f) + \frac{2}{3} \left(\frac{1}{e} + 1 \right) \ln \frac{12}{\delta} \right) \leq \frac{24}{n} \left(d \ln(880\theta_f) + \ln \frac{12}{\delta} \right). \end{aligned}$$

The theorem now follows by the principle of induction.

With this result in hand, we can immediately get some interesting results, such as the following corollary.

Corollary 1. *Suppose \mathbb{C} is the space of linear separators in d dimensions that pass through the origin, and suppose the distribution is uniform on the surface of the origin-centered unit sphere. Then with probability $\geq 1 - \delta$, any $h \in \mathbb{C}$ consistent with the n i.i.d. training examples has (for some finite universal c)*

$$er(h) \leq c \frac{d \log d + \log \frac{1}{\delta}}{n}.$$

Proof. (Hanneke, 2007) proves that $\sup_{f \in \mathbb{C}} \theta_f \leq \pi\sqrt{d}$ for this problem.

This improves over the best previously known bound for consistent classifiers for this problem in its dependence on n , which was $\propto \frac{d\sqrt{\log(n/d)+\log(1/\delta)}}{n}$ (Li and Long, 2007) (though we picked up an extra $\log d$ factor in the process).

3 Specializing to Particular Algorithms

The above analysis is for arbitrary algorithms that select a classifier consistent with the training data. However, we can modify the disagreement coefficient to be more interesting for more specific algorithms. Specifically, suppose there are sets \mathbb{C}_f such that with high probability algorithm \mathcal{A} will output a classifier in \mathbb{C}_f when f is the target function. Then we only need to worry about the regions of disagreement within these \mathbb{C}_f sets, which may be significantly smaller than within the full space \mathbb{C} .

To give a concrete example, consider the Closure algorithm: output the $h \in \mathbb{C}$ with smallest $\mathbb{P}(h(X) = +1)$ that is consistent with the data. For intersection-closed \mathbb{C} , the sets are $\mathbb{C}_f = \{h \in \mathbb{C} : h(x) = +1 \Rightarrow f(x) = +1\}$. So effectively, this becomes our concept space, and the disagreement coefficient of f with respect to \mathbb{C}_f and \mathcal{D} can be significantly smaller than it is with respect to the full space \mathbb{C} . For instance, if \mathbb{C} is axis-aligned rectangles, then the disagreement coefficient of any $f \in \mathbb{C}$ with respect to \mathbb{C}_f and \mathcal{D} is at most d . This implies a bound

$$\propto \frac{d \log d + \log(1/\delta)}{n}.$$

We already have better bounds than this for using Closure with this concept space. However, if the d upper bound on the disagreement coefficient with respect to \mathbb{C}_f is true for *general* intersection-closed spaces \mathbb{C} , this would match the best known bounds for general intersection-closed spaces (Auer and Ortner, 2004).

Note that a similar trick can be applied to create an algorithmic-specific version of the doubling dimension analysis of Li and Long (2007), which gives a similar result for the above example of learning axis-aligned rectangles with the Closure algorithm.

Bibliography

- Auer, P. and Ortner, R. (2004). A new PAC bound for intersection-closed concept classes. In *17th Annual Conference on Learning Theory (COLT)*.
- Balcan, M.-F., Hanneke, S., and Wortman, J. (2008). The true sample complexity of active learning. In *Proceedings of the 21st Conference on Learning Theory*.
- Dasgupta, S., Hsu, D., and Monteleoni, C. (2007). A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*.
- Hanneke, S. (2007). A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*.
- Hanneke, S. (2009). Rates of convergence in active learning. (*Forthcoming*).
- Hausler, D., Littlestone, N., and Warmuth, M. (1994). Predicting $\{0, 1\}$ -functions on randomly drawn points. *Information and Computation*, **115**, 248–292.
- Li, Y. and Long, P. M. (2007). Learnability and the doubling dimension. In *Advances in Neural Information Processing*.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.
- Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York.