8803 Machine Learning Theory

Maria-Florina Balcan

Lecture 5: January 26, 2010

1 Learning Models for the Realizable Case

So far we have focused on the *realizable case* – the algorithm gets a sample that is consistent with a function in a fixed concept class C. The table below summarizes some of the classes that are learnable or not learnable in the learning models we have discussed so far.

Table 1: Learnable/not learnable in polynomial time; k is assumed to be constant.

	Consistency model	Mistake Bound Model	PAC model
Disjunctions	Yes	Yes	Yes
Conjunctions	Yes	Yes	Yes
k-CNF	Yes	Yes	Yes
k-term DNF	No ($NP \neq RP$ assumption)	Yes	Yes
k-Decision Lists	Yes	Yes	Yes
Linear separators	Yes	Yes	Yes
Blum'91	Yes	No (crypto assumption)	Yes
poly size circuits	Yes	No (crypto assumption)	No (crypto assumption)
Decision Trees	Yes	Not known	Not known
DNF	Yes	Not known	Not known

We have also seen some general relationships between these models. For example, if C is learnable in the consistency model and $\ln(C)$ is $poly(n)$, then C is learnable in the PAC model. If C is learnable in the mistake bound model, then C is learnable in the PAC model.

If C is the class of linear separators of "large" L_2 margin (i.e., L_2 margin is at least $1/poly(n)$), then C is learnable in the mistake bound model via the Perceptron algorithm. If C is the class of linear separators of "large" L_1 margin, then C is learnable in the mistake bound model via the Winnow algorithm. The general class of linear separators is learnable via an ellipsoid style algorithm in the mistake bound model.

Under crypto assumptions, there exist classes that are learnable in the PAC model, but not in the mistake bound model.

Winnow versus Perceptron One can generalize the basic analysis we did for Winnow to the case of learning linear separators; the guarantee depends on the L_1, L_∞ margin of the target. If "n" is large but most features are irrelevant (i.e. target is sparse but examples are dense), then Winnow is better because adding irrelevant features increases $L_2(X)$ but not

$L_\infty(X)$. On the other hand, if the target is dense and examples are sparse, then perceptron is better.

2 The Non-realizable case

In the general case, the target function might not be in the class of functions we consider. Formally, in the non-realizable or agnostic passive supervised learning setting, we assume that the input to a learning algorithm is a set S of labeled examples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. We assume that these examples are drawn i.i.d. from some fixed but unknown distribution D over the the instance space X and that they are labeled by some target concept c^*. So $y_i = c^*(x_i)$. The goal is just as in the realizable case to do optimization over the given sample S in order to find a hypothesis $h : X \to \{0, 1\}$ of small error over whole distribution D. The error of h is defined as

$$err(h) = \Pr_{x \sim D}(h(x) \neq c^*(x)).$$

We denote by

$$err_S(h) = \Pr_{x \sim S}(h(x) \neq c^*(x))$$

the empirical error over the sample. Technically, our goal is to compete with the best function (the function of smallest true error rate) in some concept class C.

A natural hope is that picking a concept c with a small observed error rate gives us small true error rate. It is therefore useful to find a relationship between *observed* error rate for a sample and the *true* error rate.

2.1 Concentration Inequalities

Consider a hypothesis with true error rate p (or a coin of bias p) observed on m examples (the coin is flipped m times). Let S be the number of observed errors (the number of heads seen) so S/m is the observed error rate.

Hoeffding bounds state that for any $\epsilon \in [0, 1]$,

1. $\Pr[\frac{S}{m} > p + \epsilon] \leq e^{-2m\epsilon^2}$, and
2. $\Pr[\frac{S}{m} < p - \epsilon] \leq e^{-2m\epsilon^2}$.

Chernoff bounds state that under the same conditions,

1. $\Pr[\frac{S}{m} > p(1 + \epsilon)] \leq e^{-mp\epsilon^2/3}$, and
2. $\Pr[\frac{S}{m} < p(1 - \epsilon)] \leq e^{-mp\epsilon^2/2}$.

2.2 Simple sample complexity results for finite hypotheses spaces

We can use the Hoeffding bounds to show the following:

Theorem 1 *Let C be a finite hypothesis space. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size*

$$m \geq \frac{1}{2\epsilon^2}\left(\ln(|C|) + \ln\left(\frac{2}{\delta}\right)\right),$$

then probability at least $(1-\delta)$, all hypotheses h in C have

$$|err(h) - err_S(h)| \leq \epsilon. \tag{1}$$

Proof: Let us fix a hypothesis h. By Hoeffding, we get that the probability that its observed error is not within ϵ of its true error is at most $2e^{-2m\epsilon^2}$. By union bound over all h in C, we get that the probability that there exists a hypothesis $h \in C$ with $|err(h) - err_S(h)| > \epsilon$ is at most $2|C|e^{-2m\epsilon^2}$. By setting this to δ, we get the desired result. ■

Note 1 *A statement of type (1) is called a* uniform convergence *result. It implies that the hypothesis that minimizes the empirical error rate will be very close in generalization error to the best hypothesis in the class. In particular if $\widehat{h} = \text{a}rgmin_{h\in C} err_S(h)$ we have $err(\widehat{h}) \leq err(h^*) + 2\epsilon$, where $h^* \in C$ is a hypothesis in C of smallest true error rate.*

Note 2 *The sample size grows quadratically with $1/\epsilon$. Recall that the learning sample size in the realizable (PAC) case grew only linearly with $1/\epsilon$.*

Note 3 *Another way to write the bound in Theorem 1 is as follows:*

For any ϵ, $\delta > 0$, if we draw a sample from D of size m then with probability at least $1-\delta$, all hypotheses h in C have

$$err(h) \leq err_S(h) + \sqrt{\frac{\ln(|C|) + \ln\left(\frac{2}{\delta}\right)}{2m}}$$

This is the more "statistical learning theory style" way of writing the same bound.

If we believe that the best hypothesis $h^* \in C$ has a low error rate, then we can get rid of that pesky ϵ^2 by relaxing our goal, to say that for hypotheses whose true error is greater than ϵ, we are satisfied if their observed error comes just within a *factor of 2*. Specifically:

Theorem 2 *Let C be a finite hypothesis space. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size*

$$m \geq \frac{6}{\epsilon}\left(\ln(|C|) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in C$ with $err(h) > 2\epsilon$ have $err_S(h) > \epsilon$, and all $h \in C$ with $err(h) \leq \epsilon/2$ have $err_S(h) \leq \epsilon$.

Thus, if the hypothesis h^* of minimum true error has $err(h^*) \leq \epsilon/2$ then the hypothesis $\hat{h}$ of minimum empirical *error has $err(\hat{h}) \leq 2\epsilon$.*

Proof: Using Chernoff bounds, we calculate as follows. Let $\delta' = \delta/|C|$. Fix h with $err(h) = p \geq 2\epsilon$; it is enough to ensure that the empirical error $err_S(h)$ is at least $p/2 \geq \epsilon$ with confidence $1 - \delta'$. By Chernoff it is enough to ensure that $e^{-mp/8} \leq \delta'$. To get this it is enough to ensure that $e^{-m\epsilon/4} \leq \delta'$, which is true as long as $m \geq \frac{6}{\epsilon}\left(\ln(|C|) + \ln\left(\frac{1}{\delta}\right)\right)$.

On the other hand, if $err(h) = p \leq \epsilon/2$, we we want to ensure that with confidence $1 - \delta'$ the observed error $err_S(h)$ is no more than $\frac{\epsilon}{2}(1 + 1)$. The worst case occurs for $\tilde{h}$ such that $err(\tilde{h}) = \epsilon/2$; by Chernoff, we have the probability that $err_S(\tilde{h}) \geq 2err(\tilde{h})$ is at most $e^{-m \cdot err(\tilde{h})/3} = e^{-m\epsilon/6}$ which is at most δ' for $m \geq \frac{6}{\epsilon}\ln(1/\delta')$, as desired . ■

Or, to be analogous to Note 3, given m examples, with probability at least $1-\delta$, all $h \in C$ with $err(h) > \frac{12}{m}\ln(2|C|/\delta)$ satisfy $err_S(h) \geq err(h)/2$ and all $h \in C$ with $err(h) < \frac{3}{m}\ln(2|C|/\delta)$ satisfy $err_S(h) < \frac{6}{m}\ln(2|C|/\delta)$.

3 Sample complexity results for infinite hypothesis spaces

Let C be a concept class over an instance space X, i.e. a set of functions functions from X to $\{0, 1\}$ (where both C and X may be infinite). For any $S \subseteq X$, let's denote by $C(S)$ the set of all behaviors or dichotomies on S that are induced or realized by C, i.e. if $S = \{x_1, \cdots, x_m\}$, then $C(S) \subseteq \{0, 1\}^m$ and

$$C(S) = \{(c(x_1), \cdots, c(x_m)); c \in \mathcal{C}\}.$$

Also, for any natural number m, we consider $C[m]$ to be the maximum number of ways to split m points using concepts in C, that is

$$C[m] = \max\{|C(S)|; |S| = m, S \subseteq X\}.$$

With these conventions we have the following two results:

Theorem 3 *Let C be an arbitrary hypothesis space. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size*

$$m > \frac{2}{\epsilon} \cdot \left[\log_2\left(2 \cdot \mathcal{C}[2m]\right) + \log_2\left(\frac{1}{\delta}\right)\right] \tag{2}$$

then with probability $(1-\delta)$, all bad hypothesis in $\mathcal{C}$ (with error $> \epsilon$ with respect to c^* and D) are inconsistent with the data.

Theorem 4 *Let C be an arbitrary hypothesis space. Let D be an arbitrary, fixed unknown probability distribution over X and let c^* be an arbitrary unknown target function. For any ϵ, $\delta > 0$, if we draw a sample S from D of size $m > (8/\epsilon^2)[\ln(2C[2m]) + \ln(1/\delta)]$ then with probability $1-\delta$, all h in C have $|err_D(h) - err_S(h)| < \epsilon$.*

If C is the class of thresholds, then $C[m] = m+1$; if C is the class of intervals, then $C[m] = O(m^2)$,

Definition 1 *If $|C\,(S)\,| = 2^{|S|}$ then S is* **shattered** *by C.*

Definition 2 *The* **Vapnik-Chervonenkis dimension** *of C, denoted $VC_{DIM}(C)$, is the largest cardinality d such that there exists a sample set of that cardinality $|S| = d$ that is shattered by C. If no largest cardinality exists then $VC_{DIM}(C) = \infty$.*

In general if C has VC-dimension d, then $C[m] = O(m^d)$, so if we solve for example Equation 2 we get that

$$m > \frac{2}{\epsilon} \cdot \left[d \log_2\left(\frac{1}{\epsilon}\right) + \log_2\left(\frac{1}{\delta}\right)\right]$$

is sufficient to show that then with probability $1-\delta$, all hypotheses/functions in C with error $\geq \epsilon$ are inconsistent with the data.