8803 Connections between Learning, Game Theory, and Optimization

Maria-Florina Balcan Lecture 10: September 23, 2010

General Sample Complexity Results

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

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

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 result:

The Realizable Case

Theorem 1 *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*

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

then with probability at least $1-\delta$, all the hypotheses in C with $err_D(h) > \epsilon$ are inconsistent with the data, i.e., $err_S(h) \neq 0$.

Proof: It suffices to bound the probability of the following "bad" event:

B : $\exists h \in C$ with $err_S(h) = 0$ but $err_D(h) > \epsilon$.

Let us denote the training sample by $S = \{x_1, x_2, ..., x_m\}$. Now suppose $S' = \{x'_1, x'_2, ..., x'_m\}$ is another sample drawn i.i.d. from D (a "ghost sample").

Let us consider the following event:

B' : $\exists h \in C$ with $err_S(h) = 0$ but $err_{S'}(h) > \epsilon/2$.

Claim 1 *If $m > \frac{8}{\epsilon}$, then* $\mathbf{Pr}[B'|B] \geq 1/2$.

Proof: Suppose h is a consistent with S but $err_D(h) > \epsilon$. Let $M(h, S')$ denote the number of mistakes made by h on S'. Since S' is drawn i.i.d. from D, $E[M(h, S')] \geq \epsilon m$. Further, by Chernoff, we have $\mathbf{Pr}[M(h, S') \leq \epsilon m/2] < e^{-m\epsilon/8} \leq 1/2$, for $m > \frac{8}{\epsilon}$. This then implies the desired result. ■

We have

$$\frac{\mathbf{Pr}[B']}{\mathbf{Pr}[B]} \geq \frac{\mathbf{Pr}[B' \wedge B]}{\mathbf{Pr}[B]} = \mathbf{Pr}[B'|B] \geq \frac{1}{2},$$

so $\mathbf{Pr}[B] \leq 2\mathbf{Pr}[B']$. Thus it suffices to bound $\mathbf{Pr}[B']$ (this probability is over choices of S and S').

Given two samples S and S', consider the following random process SwapR.

> For i from 1 to m, do the following:
>
> > Flip a fair coin. If you get heads, swap x_i and x'_i, else do nothing.
>
> Let us denote the new collections by T and T'.

We clearly have:

Claim 2 *Suppose we pick S and S' according to D and then perform SwapR. Then the sets T and T' are identically distributed to S and S'.*

Let us now define the event:

B'' : $\quad \exists h \in C$ with $err_T(h) = 0$ but $err_{T'}(h) \geq \epsilon/2$.

Claim 2 implies that $\mathbf{Pr}[B''] = \mathbf{Pr}[B']$. The first probability is over the choice of S, S' and the random bits of SwapR while the second probability is over choice of S, S'.

Claim 3 *Fix $h \in C$. We have*

$$\mathbf{Pr}[err_T(h) = 0 \wedge err_{T'}(h) > \epsilon/2|S, S'] \leq 2^{-\epsilon m/2}.$$

Proof: Consider

$$\begin{array}{cccc} h(x_1), & h(x_2), & \ldots, & h(x_m) \\ h(x'_1), & h(x'_2), & \ldots, & h(x'_m) \end{array}$$

First, note that if there is a column with both predictions wrong then $M(h, T) = 0$ can never happen and so we are done (the desired probability is 0). Similarly, if more than $(1 - \epsilon/2)m$ of the columns have both predictions right, we are done since $M(h, T') > \epsilon m/2$

cannot happen. Thus at least $r \geq \epsilon m/2$ columns have one right and one wrong prediction. If we need $M(h,T) = 0$, it must happen that in all such columns, SwapR must ensure that the right prediction goes to the top and the wrong one goes to the bottom row. Thus the probability is $2^{-r} \leq 2^{-\epsilon m/2}$. ■

We are now ready to bound $\mathbf{Pr}[B'']$.

Claim 4 $\mathbf{Pr}[B''] \leq C[2m]2^{-\epsilon m/2}$

Proof: By definition we have:

$$\begin{aligned}\mathbf{Pr}[B''] &= \mathbf{E}_{S,S'}[\mathbf{Pr}_{swapR}[\exists h \in C, M(h,T) = 0 \wedge M(h,T') \geq m\epsilon/2 \mid S, S']] \\ &= \mathbf{E}_{S,S'}[\mathbf{Pr}_{swapR}[\exists h \in C[S \cup S'], M(h,T) = 0 \wedge M(h,T') \geq m\epsilon/2 \mid S, S']].\end{aligned}$$

By union bound:

$$\begin{aligned}\mathbf{Pr}[B''] &\leq \mathbf{E}_{S,S'}\left[\sum_{h \in C[S \cup S']} \mathbf{Pr}_{swapR}[M(h,T) = 0 \wedge M(h,T') \geq m\epsilon/2 \mid S, S']\right] \\ &\leq C[2m]2^{-\epsilon m/2},\end{aligned}$$

as desired. ■

Combining all these we get that $\mathbf{Pr}[B] \leq \delta$ whenever $2C[2m]2^{-\epsilon m/2} \leq \delta$ which proves that if

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

then with probability at least $1-\delta$, all the hypotheses in C with $err_D(h) > \epsilon$ are inconsistent with the data, i.e., $err_S(h) \neq 0$.

■

Intuition: For a fixed h it is clear that

$$\mathbf{Pr}_{S,S'}[M(h,S) = 0 \wedge M(h,S') > \epsilon m/2] \leq 2^{-\epsilon m/2}.$$

However, there are potentially infinitely many hypotheses, and we would want to somehow do union bound as in the proof of the corresponding theorem in the finite case. Once we draw S, there are finitely many hypotheses left, but no randomness left; so we cannot bound the probability of bad events happening. However if we do this symmetrization trick, somewhere in the middle we manage to get to a finite class and do union bound, but still have some randomeness saved to bound the probability of a bad event happening.

The Non-realizable Case

Theorem 1 presents a sample complexity statement for the realizable case (the case when the target function is in our class of functions). We can get similar statements in the non-realizable case:

Theorem 2 *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*

$$m > \frac{8}{\epsilon^2} \cdot \left[\log_2 (2 \cdot C[2m]) + \log_2 \left(\frac{1}{\delta}\right)\right],$$

then with probability at least $1 - \delta$, all h in C have

$$|err_D(h) - err_S(h)| < \epsilon.$$

Proof Sketch: Just need to redo the proof using Hoeffding.

Draw $2m$ examples. Let B be the event that on first m, there exists hypothesis in C with empirical and true error that differ by at least ϵ. (This is what we want to bound for the theorem). Let B' be the event that there exists a concept in C whose empirical error on 1st half differs from empirical error on 2nd half by at least $\epsilon/2$.

For large enough m, we have $\mathbf{Pr}[B'|B] \geq 1/2$ so $\mathbf{Pr}[B] \leq 2 \cdot \mathbf{Pr}[B']$. Now we have to show that $\mathbf{Pr}[B']$ is low.

As before, let's first pick S, S', then we do the symmetrization (or swapping). Once $S \cup S'$ is determined, there are only $C[2m]$ hypotheses we need to worry about. Using the fact

$$\mathbf{Pr}_{Swap}[|M(h,T) - M(h,T')| > \epsilon m/2 | S, S'] \leq 2^{-\epsilon^2 m/8},$$

and a similar reasoning as in Theorem 1, we get the desired result. ■