10-806 Foundations of Machine Learning and Data Science

Maria-Florina Balcan Lecture 4-5: September 21st and September 23rd, 2015

Sample Complexity Results for Infinite Hypothesis Spaces

The Shattering Coefficient

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 labelings 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

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

$C[m]$ is called the *shatter coefficient* or *growth function* of class C.

We will soon prove an important theorem which roughly says that we can replace $\ln(|C|)$ with $C[2m]$ in our basic sample complexity bound, allowing us to address infinite concept classes such as linear separators. First, however, we define the notion of *VC-dimension* and state and prove *Sauer's lemma* which relates the shatter coefficient to VC-dimension.

VC Dimension

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

Definition 2 *The* **Vapnik-Chervonenkis dimension** *of C, denoted as $VCdim(C)$, is the cardinality of the largest set S shattered by C. If arbitrarily large finite sets can be shattered by C, then $VCdim(C) = \infty$.*

Note 1 *In order to show that the VC dimension of a class is at least d we must simply find some shattered set of size d. In order to show that the VC dimension is at most d we must show that no set of size $d+1$ is shattered.*

Examples

1. Let C be the concept class of thresholds on the real number line. Clearly samples of size 1 can be shattered by this class. However, no sample of size 2 can be shattered since it is impossible to choose threshold such that x_1 is labeled positive and x_2 is labeled negative for $x_1 < x_2$. Hence the $VCdim(C) = 1$.

2. Let C be the concept class intervals on the real line. Here a sample of size 2 is shattered, but no sample of size 3 is shattered, since no concept can satisfy a sample whose middle point is negative and outer points are positive. Hence, $VCdim(C) = 2$.

3. Let C be the concept class of k non-intersecting intervals on the real line. A sample of size $2k$ shatters (just treat each pair of points as a separate case of example 2) but no sample of size $2k + 1$ shatters, since if the sample points are alternated positive/negative, starting with a positive point, the positive points can't be covered by only k intervals. Hence $VCdim(C) = 2k$.

4. Let C the class of linear separators in $\mathbf{R^2}$. Three points can be shattered, but four cannot; hence $VCdim(C) = 3$. To see why four points can never be shattered, consider two cases. The trivial case is when one point can be placed within a triangle formed by the other three; then if the middle point is positive and the others are negative, no half space can contain only the positive points. If however the points cannot be arranged in that pattern, then label two points diagonally across from each other as positive, and the other two as negative In general, one can show that the VCdimension of the class of linear separators in $\mathbf{R^n}$ is $n + 1$.

5. The class of axis-aligned rectangles in the plane has $VC_{DIM} = 4$. The trick here is to note that for any collection of five points, at least one of them must be interior to or on the boundary of any rectangle bounded by the other four; hence if the bounding points are positive, the interior point cannot be made negative.

Sauer's Lemma

Lemma 1 *If $d = VCdim(C)$, then for all m, $C[m] \leq \Phi_d(m)$, where $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$.*

Proof: The proof proceeds by induction on both d and m. We have two base cases: when $m = 0$ and d is arbitrary, and when $d = 0$ and m is arbitrary. When $m = 0$, there can only be one subset, hence $C[0] \leq 1 = \Phi_d(0)$. When $d = VCdim(C) = 0$, no set of points can be shattered, hence all points can be labelled only one way. From this we conclude that $C[m] = 1 \leq \Phi_0(m)$. So the lemma holds for the base case.

We assume for induction that for all m', d' such that $m' \leq m$ and $d' \leq d$ and at least one of these inequalities is strict, we have $C[m'] \leq \Phi_{d'}(m')$.

Now suppose we have a set $S = \{x_1, x_2, \ldots, x_m\}$ of cardinality m. Let H be a class of functions defined only over $\{x_1, x_2, \ldots, x_m\}$ such that $C(S) = H(S) = H$. Since any $\tilde{S} \subseteq S$ that is shattered by H is also shattered by C, we have $VCdim(H) \leq VCdim(C)$.

We now construct H_1 and H_2 on which we apply our induction hypothesis as follows: for each possible labeling of $\{x_1, x_2, \ldots, x_{m-1}\}$ induced by a function in H, we add a representative function from H to H_1; we let $H_2 = H \setminus H_1$. So for each $h \in H_2$, $\exists \tilde{h} \in H_1$ such that $h(x_i) = \tilde{h}(x_i)$ for $i \in \{1, \ldots, m-1\}$ and $h(x_m) \neq \tilde{h}(x_m)$. For convenience, let's choose the representatives such that $h(x_m) = 1$ and $\tilde{h}(x_m) = 0$, so all $h \in H_2$ label x_m as positive.

By construction we have

$$|C(S)| = |H(S)| = |H_1(S)| + |H_2(S)|.$$

Since $H_1 \subseteq H$ we have $VCdim(H_1) \leq VCdim(H) \leq d$. Moreover, we can show

$$|H_1(S)| = |H_1(S \setminus \{x_m\})|.$$

In one direction it is clear that $|H_1(S)| \geq |H_1(S \setminus \{x_m\})|$. In the other direction we have $|H_1(S)| \leq |H_1(S \setminus \{x_m\})|$ since there is no labeling h of $\{x_1, x_2, \ldots, x_{m-1}\}$ such that both $(h(x_1), h(x_2), \ldots, h(x_{m-1}), 0)$ and $(h(x_1), h(x_2), \ldots, h(x_{m-1}), 1)$ are in H_1.

By induction we have:

$$|H_1(S)| \leq \Phi_d(m-1).$$

Now note that if T is shattered by H_2, then $T \cup \{x_m\}$ is shattered by H. If T is shattered by H_2 then (a) $x_m \notin T$ (because all $h \in H_2$ label x_m as positive), and (b) $T \cup \{x_m\}$ is shattered by H (because each $h \in H_2$ has a twin $\tilde{h} \in H_1$ that is identical except on x_m). So,

$$VCdim(H_2) \leq VCdim(H) - 1 \leq d - 1.$$

We can also show

$$|H_2(S)| = |H_2(\{x_1, x_2, \ldots, x_{m-1}\})|,$$

and by induction we get:

$$|H_2(S)| \leq \Phi_{d-1}(m-1).$$

Combining all these we get

$$|C(S)| \leq \Phi_d(m-1) + \Phi_{d-1}(m-1).$$

Since

$$\sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} = \binom{m}{0} + \sum_{i=1}^{d} \binom{m-1}{i} + \sum_{i=1}^{d} \binom{m-1}{i-1} = \sum_{i=0}^{d} \binom{m}{i},$$

we get $|C(S)| \leq \Phi_d(m)$, as desired. ■

Note that for C the class of intervals we achieve $C[m] = \Phi_d(m)$, where $d = VCdim(C)$.

Lemma 2 *For $m > d$ we have:*

$$\Phi_d(m) \leq \left(\frac{em}{d}\right)^d.$$

Proof: Since $m > d$, we have $0 \leq \frac{d}{m} < 1$. Therefore:

$$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \leq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} = \left(1 + \frac{d}{m}\right)^m \leq e^d.$$

■

Shattering Coefficient Based Sample Complexity Results

We now prove an important sample complexity result using the shatter coefficient. We focus on the realizable case (where the target function belongs to class C). It can be easily changed to handle the non-realizable case (and will cover it in a future lecture).

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*

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

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 : $\quad \exists h \in C$ 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' : $\quad \exists h \in C$ 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:

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

By union bound:

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

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

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

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, in 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.

VC-dimension Based Sample Complexity Results

We can now combine our sample complexity statement based on the shatter coefficient with Sauer's lemma to get a nice closed form expression on sample complexity (an upper bound on the number of samples needed to learn concepts from the class) based on the VC-dimension of a concept class. For convenience, we focus on the realizable case.

Theorem 2 *Let C be an arbitrary hypothesis space of VC-dimension d. Let D be an arbitrary unknown probability distribution over the instance space 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 satisfying*

$$m \geq \frac{8}{\epsilon}\left[d\ln\left(\frac{16}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right].$$

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: Note that it suffices to set $2^{\epsilon m/2} \geq \frac{2C[2m]}{\delta}$. To do so it suffices to set $m \geq \frac{4}{\epsilon}\ln\left(\frac{2C[2m]}{\delta}\right)$. We can now use Sauer's lemma to show that it is enough to set

$$m \geq \frac{4}{\epsilon}\left[d\ln\left(\frac{2me}{d}\right) + \ln\left(\frac{2}{\delta}\right)\right]$$

or

$$m \geq \frac{4}{\epsilon}\left[d\ln m + d\ln\left(\frac{2e}{d}\right) + \ln\left(\frac{2}{\delta}\right)\right]$$

We now use the inequality $\ln x \leq \alpha x - \ln\alpha - 1$ for $\alpha, x > 0$ to show

$$\frac{4d}{\epsilon}\ln m \leq \frac{4d}{\epsilon}\left[\frac{\epsilon}{8d}m + \ln\left(\frac{8d}{\epsilon}\right) - 1\right] = \frac{m}{2} + \frac{4d}{\epsilon}\ln\left(\frac{8d}{e\epsilon}\right).$$

So it suffices to set

$$m \geq \frac{m}{2} + \frac{4d}{\epsilon} \ln\left(\frac{8d}{e\epsilon}\right) + \frac{4d}{\epsilon} \ln\left(\frac{2e}{d}\right) + \frac{4}{\epsilon} \ln\left(\frac{2}{\delta}\right).$$

Simplifying we get:

$$m \geq \frac{8}{\epsilon}\left[d \ln\left(\frac{16}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right].$$

■