Lecure Notes 8
36-705

We continue our discussion of uniform convergence. Recall that

$$P_n(A) = \frac{1}{n}\sum_i I_A(X_i)$$

and

$$\Delta_n(\mathcal{A}) = \sup_{A\in\mathcal{A}} |P_n(A) - P(A)|.$$

1 Finite Collections

The first case to consider is when the collection of sets $\mathcal{A}$ has finite cardinality $|\mathcal{A}|$. In other words

$$\mathcal{A} = \{A_1, \ldots, A_N\}.$$

In this case, for a fixed A we know by Hoeffding's inequality that,

$$\mathbb{P}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t) \leq 2\exp(-2nt^2).$$

However, we want something stronger we want that this convergence happens uniformly for all sets in $\mathcal{A}$, so we can use the union bound, i.e.

$$\begin{aligned}\mathbb{P}(\Delta_n(\mathcal{A}) \geq t) &= \mathbb{P}(\cup_{A\in\mathcal{A}}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t)) \\ &\leq \sum_{A\in\mathcal{A}} \mathbb{P}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t)) \\ &\leq 2|\mathcal{A}|\exp(-2nt^2).\end{aligned}$$

If we set the right hand side equal to δ we get

$$t = \sqrt{\frac{\ln(2|\mathcal{A}|/\delta)}{2n}}.$$

So

$$\mathbb{P}\left(\Delta_n(\mathcal{A}) \geq \sqrt{\frac{\ln(2|\mathcal{A}|/\delta)}{2n}}\right) \leq \delta.$$

In other words we have that with probability at least $1-\delta$,

$$\Delta(\mathcal{A}) \leq \sqrt{\frac{\ln(2|\mathcal{A}|/\delta)}{2n}}.$$

This is already quite a nice result and once again highlights one of the main reasons why Hoeffding type exponential concentration inequalities are much more useful than Chebyshev type concentration inequalities: to obtain uniform convergence over $\mathcal{A}$ we pay a price which is logarithmic in the size of the collection.

2 VC dimension

Often we are interested in controlling $\Delta(\mathcal{A})$ for infinite classes of sets. The example from last lecture of uniform convergence of the empirical CDF is a canonical example. One way to do this is to use the notion of VC dimension. First, we need to understand the concept of shattering.

Shattering: Let $\{z_1, \ldots, z_n\}$ be a finite set of n points. We let $N_{\mathcal{A}}(z_1, \ldots, z_n)$ be the number of distinct sets in the collection of sets

$$\Big\{\{z_1, \ldots, z_n\} \cap A : A \in \mathcal{A}\Big\}.$$

$N_{\mathcal{A}}(z_1, \ldots, z_n)$ is counting the *number of subsets* of $\{z_1, \ldots, z_n\}$ that the collection of sets $\mathcal{A}$ picks out. Note that, $N_{\mathcal{A}}(z_1, \ldots, z_n) \leq 2^n$.

We now define the n-th shatter coefficient of $\mathcal{A}$ as:

$$s(\mathcal{A}, n) = \max_{\{z_1, \ldots, z_n\}} N_{\mathcal{A}}(z_1, \ldots, z_n).$$

The shatter coefficient is the maximal number of different subsets of n points that can be picked out by the collection $\mathcal{A}$.

Example: Consider points on the real line and let $\mathcal{A}$ be the collection of left intervals $\mathbb{I}(-\infty, t]$ for all t. If we have n points on the line then we can pick out any left subset of the points, i.e. $s(\mathcal{A}, n) = n + 1$. For example, let $n = 3$ and consider the points $\{0, 1, 2\}$. We can pick out the following subsets:

$$\{\emptyset\}, \{0\}, \{0, 1\}, \{0, 1, 2\},$$

and no others. This is true for any set of three points. So $s(\mathcal{A}, 3) = 4$. We will see more examples soon.

VC Theorem: For *any distribution* $\mathbb{P}$, and class of sets $\mathcal{A}$ we have that,

$$\mathbb{P}(\Delta_n(\mathcal{A}) \geq t) \leq 8 s(\mathcal{A}, n) \exp(-nt^2/32).$$

Notes: There are two noteworthy aspects of this theorem.

1. The result is very general and it applies to any distribution on the samples, and such results are often called *distribution free*.

2. The VC theorem essentially reduces the question of uniform convergence to a combinatorial question about the collection of sets, i.e. we now need only to understand the shatter coefficients which are completely independent from probability/statistics.

3. The proof of this result is quite straightforward using some of the machinery (introducing a ghost sample, symmetrization) that we will see in the next lecture.

Glivenko-Cantelli: This theorem immediately implies the Glivenko-Cantelli theorem we studied in the last lecture, i.e. that the empirical CDF converges in probability to the true CDF. To see this we note that the shatter coefficients of the left intervals are bounded by $n+1$ so the VC theorem tells us that,

$$\mathbb{P}\left(\sup_x |\widehat{F}_n(x) - F_X(x)| \geq t\right) \leq 8(n+1)\exp(-nt^2/32).$$

Now verifying convergence in probability is straightforward by noting that for any $t > 0$, $\lim_{n\to\infty} 8(n+1)\exp(-nt^2/32) = 0$.

VC dimension: We now that $s(\mathcal{A}, n) \leq 2^n$ for each n. The *VC dimension* d is the largest integer d for which $s(\mathcal{A}, d) = 2^d$.

It follows that, for any $n > d$, we have that $s(\mathcal{A}, n) < 2^n$. The surprising combinatorial result of Vapnik and Chervonenkis (sometimes called Sauer's lemma) is that there is a phase transition of shattering coefficients: once it is no longer exponential (i.e. once $n > d$) the shattering coefficients become polynomial in n, i.e.

Sauer's Lemma: If $\mathcal{A}$ has finite VC dimension d, then for $n > d$ we have that,

$$s(\mathcal{A}, n) \leq (n+1)^d.$$

We can use Sauer's lemma to conclude that for a system $\mathcal{A}$ of VC dimension d.

$$\mathbb{P}(\Delta_n(\mathcal{A}) \geq t) \leq 8(n+1)^d \exp(-nt^2/32).$$

Doing the usual thing we see that with probability $1-\delta$,

$$\Delta_n(\mathcal{A}) \leq \sqrt{\frac{32}{n}\left[d\log(n+1) + \log(8/\delta)\right]}.$$

There are some important notes:

1. If $d < \infty$ then $\Delta(\mathcal{A}) \xrightarrow{p} 0$, and so we have a uniform LLN for the collection of sets $\mathcal{A}$.
2. There are converses to the VC theorem that say roughly that if the VC dimension is infinite then there exists a distribution over the samples for which we do not have a uniform LLN.
3. Roughly, one should think of the VC result as saying for a class with VC dimension d,

$$\Delta(\mathcal{A}) \approx \sqrt{\frac{d\log n}{n}}.$$

3 More examples

There are many examples of collections of sets for which the VC dimension is known. A few popular ones are in Table 1.

Class $\mathcal{A}$	VC dimension $V_{\mathcal{A}}$
$\mathcal{A} = \{A_1, \ldots, A_N\}$	$\leq \log_2 N$
Intervals $[a, b]$ on the real line	2
Discs in $\mathbb{R}^2$	3
Closed balls in $\mathbb{R}^d$	$\leq d + 2$
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d + 1$
Convex polygons in $\mathcal{R}^2$	∞
Convex polygons with d vertices	$2d + 1$

Table 1: The VC dimension of some classes $\mathcal{A}$.

4 Back to Binary Classification

In binary classification, we have a collection of classifiers $\mathcal{F}$. This collection induces a set system:

$$\mathcal{A} = \left\{ \Big\{ \{x : f(x) = 1\} \times \{0\} \Big\} \bigcup \Big\{ \{x : f(x) = 0\} \times \{1\} \Big\}, f \in \mathcal{F} \right\}.$$

If $\mathcal{A}$ has VC dimension d then we can use the VC theorem in a straightforward way to conclude that with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - \mathbb{P}(f(X) \neq y)| = \Delta(\mathcal{A}) \leq \sqrt{\frac{32}{n} \left[d \log(n + 1) + \log(8/\delta) \right]}.$$

It is not too hard to verify that the VC dimension is essentially driven by the complexity of the sets $\mathbb{I}(f(x) = 1)$ and their complements for the classifiers in $\mathcal{F}$. This in a straightforward way, for instance, leads to a uniform convergence guarantee for empirical risk minimization over linear classifiers since they induce relatively simple sets (half-spaces) whose VC dimension is well-understood.

5 Rademacher Complexity

Now we discuss a different notion of the complexity of a class of functions. Suppose we have a collection of functions $\mathcal{F}$, we observe samples $X_1, \ldots, X_n \sim P$ for some distribution P and we are interested in (upper bounding) the quantity:

$$\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f] \right|.$$

Fix a set of points $\{x_1, \ldots, x_n\}$. Let $\epsilon = \{\epsilon_1, \ldots, \epsilon_n\}$ denote a collection of n Rademacher random variables, i.e. they take the values $\{+1, -1\}$ with equal probabilities. We define the *empirical* Rademacher complexity by

$$\mathcal{R}(x_1, \ldots, x_n) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right].$$

When $\{x_1, \ldots, x_n\}$ is a random sample then the empirical Rademacher complexity is a random variable. We define the Rademacher complexity of the class $\mathcal{F}$ as the expectation of this quantity, i.e.

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\epsilon \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right].$$

The Rademacher complexity measures the maximum absolute covariance between $\{f(X_1), \ldots, f(X_n)\}$ and a vector of random signs $\{\epsilon_1, \ldots, \epsilon_n\}$. Intuitively, we think of a class $\mathcal{F}$ as too large if for many random sign vectors we can find a function in $\mathcal{F}$ that is strongly correlated with the random sign vectors. The main utility of the Rademacher complexity is that it upper bounds the quantity $\Delta_n(\mathcal{F})$.

Rademacher Theorem:

$$\mathbb{E}[\Delta_n(\mathcal{F})] \leq 2\mathcal{R}_n(\mathcal{F}).$$

This theorem again might not appear to be so useful since we still need to understand the Rademacher complexity. It turns out that the Rademacher complexity is relatively easy to upper bound in terms of more geometric measures of the function class $\mathcal{F}$ (these are things like covering numbers or bracketing numbers of $\mathcal{F}$). This is analogous to how VC theory gave us a way to go from the uniform convergence question to a combinatorial property of the collection of sets.

Proof: The proof will resemble what we did when we proved Hoeffding's inequality. We will introduce a ghost sample, and symmetrize the empirical process. Let $\{Y_1, \ldots, Y_n\}$ be

another independent identically distributed sample. Then,

$$\begin{aligned}
\mathbb{E}[\Delta_n(\mathcal{F})] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f]\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{Y_i} f(Y_i)\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y \frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)]\right|\right] \\
&\leq \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\mathbb{E}_Y\left|\frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)]\right|\right] \\
&\leq \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)]\right|\right].
\end{aligned}$$

The distribution of the difference $f(X_i) - f(Y_i)$ is the same as the distribution of $\epsilon_i(f(X_i) - f(Y_i))$ so we obtain,

$$\begin{aligned}
\mathbb{E}[\Delta_n(\mathcal{F})] &\leq \mathbb{E}_{X,Y,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i[f(X_i) - f(Y_i)]\right|\right] \\
&\leq 2\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \\
&= 2\mathcal{R}_n(\mathcal{F}),
\end{aligned}$$

which gives us the Rademacher theorem. □

If the function class is bounded, i.e. for every $f \in \mathcal{F}$ we have that $\|f\|_\infty \leq b$, then by McDiarmid's inequality, $\Delta_n(\mathcal{F})$ is sharply concentrated around its mean, i.e.

$$\mathbb{P}(|\Delta(\mathcal{F}) - \mathbb{E}[\Delta(\mathcal{F})]| \geq t) \leq 2\exp(-nt^2/(2b^2)).$$

Putting this inequality together with the upper bound on the mean we obtain that for a bounded class $\mathcal{F}$ with probability at least $1-\delta$,

$$\Delta(\mathcal{F}) \leq 2\mathcal{R}(\mathcal{F}) + b\sqrt{\frac{2\ln(2/\delta)}{n}}.$$