GC and Donsker classes

Lab 9, BIO251

04/13/2014

1 A Donsker Theorem

We now provide a sufficient condition for a class being Donsker. This condition will be defined in terms of a "uniform entropy" of the covering numbers, but there are other sufficient conditions using the entropy of the bracketing numbers which we won't consider. We have the following result:

Theorem. Let $\mathcal{F}$ be a class of measurable functions, with envelope F, that satisfies the following uniform entropy bound:

$$\int_0^\infty \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon < \infty$$

The supremum is taken over all finitely discrete probability measures Q on $(\mathcal{X}, \mathcal{A})$, such that $||F||_{Q,2}^2 = \int F^2 dQ > 0$. Let the classes $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, ||f - g||_{P,2} < \delta\}$ and $\mathcal{F}_\infty^2 = \{f^2 : f \in \mathcal{F}_\infty\}$ be P-measurable for any $\delta > 0$. If $P^* F^2 < \infty$, then F is P-Donsker.

Proof. Let $\delta_n \downarrow 0$ be arbitrary. By Markov's inequality we have:

$$P^*(||\mathbb{G}_n||_{\mathcal{F}_{\delta_n}} \geq x) \leq \frac{\mathrm{E}^* ||\mathbb{G}_n||_{\mathcal{F}_{\delta_n}}}{x}$$

We then use the symmetrization lemma we proved last time:

$$\frac{\mathrm{E}^* ||\mathbb{G}_n||_{\mathcal{F}_{\delta_n}}}{x} \leq \frac{2}{x} \mathrm{E}^* \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right|\right|_{\mathcal{F}_{\delta_n}}$$

Now since we are assuming the class $\mathcal{F}_{\delta_n}$ is P-measurable the E^* can be replaced by iterative expectation $\mathrm{E}^* = \mathrm{E}_X \mathrm{E}_\varepsilon$. We now fix the values of $X_1, \ldots, X_n$. We next apply Hoeffding's inequality to the Rademacher process $f \mapsto \{n^{-1/2} \sum_{i=1}^n \varepsilon_i f(X_i)\}$ to conclude that this process is sub-Gaussian for the $L_2(\mathbb{P}_n)$-seminorm:

$$||f||_n = \sqrt{\frac{1}{n} \sum_{i=1}^n f^2(X_i)}$$

Now we use the second part of the corollary we derived in the end of Lab 6 (with $f_0 = 0$), to conclude that:

$$\mathrm{E}_\varepsilon \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right|\right|_{\mathcal{F}_{\delta_n}} \lesssim \int_0^\infty \sqrt{\log D(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))} d\varepsilon \lesssim \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))} d\varepsilon$$

Where the last inequality follows upon noting that we can change the variable $\varepsilon \to 1/2\varepsilon$, and increase the inequality constant a bit.

Note here that when ε is large enough the space $\mathcal{F}_{\delta_n}$ can be contained in only 1 ball. This certainly happens when $\varepsilon > \theta_n$, where:

$$\theta_n^2 = \sup_{f \in \mathcal{F}_{\delta_n}} ||f||_n^2 = \left|\left| \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right|\right|_{\mathcal{F}_{\delta_n}}$$

This is true because then we can center a ball at 0 and radius ε to cover the whole class $\mathcal{F}_{\delta_n}$. Thus we have:

$$\mathrm{E}_\varepsilon \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right|\right|_{\mathcal{F}_{\delta_n}} \lesssim \int_0^{\theta_n} \sqrt{\log N(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))} d\varepsilon$$

Furthermore obviously the covering numbers of the class $\mathcal{F}_\delta \subset \mathcal{F}_\infty$ are bounded by the covering numbers of $\mathcal{F}_\infty$. The latter numbers satisfy $N(\varepsilon, \mathcal{F}_\infty, L_2(Q)) \leq N^2(\varepsilon/2, \mathcal{F}, L_2(Q))$ for all measures Q (see why in the appendix). Therefore upon a change of variables, and bounding the integrand we consequently get:

$$\begin{aligned}
\mathrm{E}_\varepsilon \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right|\right|_{\mathcal{F}_{\delta_n}} &\lesssim \int_0^{\theta_n/||F||_n} \sqrt{\log N(\varepsilon ||F||_n, \mathcal{F}, L_2(\mathbb{P}_n))} d\varepsilon ||F||_n \\
&\lesssim \int_0^{\theta_n/||F||_n} \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon ||F||_n \\
&\leq \int_0^{\infty} \mathbb{1}(\varepsilon \leq \theta_n^*/||F_*||_n) ||F^*||_n \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon
\end{aligned}$$

Where the supremum is taken over all finitely discrete measures Q (this set trivially incudes $\mathbb{P}_n$ so the bound is trivial in that sense). Note that by doing this trivial bounding we got rid of the dependence of the integrand on the dataset. The only thing that depend on the dataset still is $\mathbb{1}(\varepsilon \leq \theta_n^*/||F_*||_n)||F^*||_n$. Everything so far was conditional on the data $X_1, \ldots, X_n$.

Note that we can always add the constant 1 to the envelope function F without changing the second moment condition. We still need to get the expectation E_X.This will ensure that $F_* \geq 1$. Taking it results in:

$$\begin{aligned}
&\int_0^{\infty} \mathrm{E}_X \, \mathbb{1}(\varepsilon \leq \theta_n^*/||F_*||_n) ||F^*||_n \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon \\
&\overset{\text{CS}}{\leq} \underbrace{\sqrt{\mathrm{E}_X \, ||F^*||_n^2}}_{O(1)} \int_0^{\infty} \sqrt{P(\varepsilon \leq \theta_n^*)} \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon
\end{aligned}$$

Now since the integrand $\sqrt{P(\varepsilon \leq \theta_n^*)} \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} \leq \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))}$ which is integrable by our assumption, the Dominated Convergence Theorem would ensure that the above integral converges to 0 provided that $\theta_n \overset{P_*}{\to} 0$, which will finish the proof of the asymptotic equicontinuity because then $P(||\mathbb{G}_n||_{\mathcal{F}_{\delta_n}} \geq x)$ will converge to 0, as $\delta_n \to 0$.

Note that $\theta_n = \left|\left|\mathbb{P}_n f^2\right|\right|_{\mathcal{F}_{\delta_n}}$. Note then that since $\sup\{Pf^2 : f \in \mathcal{F}_{\delta_n}\} \to 0$, and $\mathcal{F}_{\delta_n} \subset \mathcal{F}_\infty$, it is enough to show that:

$$\left|\left|\mathbb{P}_n f^2 - Pf^2\right|\right|_{\mathcal{F}_\infty} \to 0$$

This is of course a ULLN for the class $\mathcal{F}_\infty^2$. This class has an integrable envelope $(2F)^2$, and is P-measurable by assumption. For any pair of functions $f, g \in \mathcal{F}_\infty$ we have:

$$\mathbb{P}_n|f^2 - g^2| \leq \mathbb{P}_n|f-g|4F \leq ||f-g||_n \, ||4F||_n$$

Therefore if $||f-g||_n \leq \varepsilon ||F||_n$ it follows that $\mathbb{P}_n|f^2 - g^2| \leq \varepsilon(2||F||_n)^2$. This statement translated to covering numbers is $N(\varepsilon||2F||_n^2, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n)) \leq N(\varepsilon||F||_n, \mathcal{F}_\infty, L_2(\mathbb{P}_n))$. As we argued earlier we have:

$$N(\varepsilon||F||_n, \mathcal{F}_\infty, L_2(\mathbb{P}_n)) \leq N^2(\varepsilon||F||_n/2, \mathcal{F}, L_2(\mathbb{P}_n))$$

and the latter must be a finite number in order for us to have the uniform entropy bounded. It is shown in the appendix that the condition $N(\varepsilon||2F||_n^2, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$ is a finite number implies the GC condition in the GC theorem above (or you can argue from the remark in the end of section 1). This concludes the asymptotic equicontinuity part of the proof.

The final step of the proof is to show that, the space $\mathcal{F}$ is totally bounded in $L_2(P)$. From the result that we just showed we know that there exists a sequence of finitely discrete measures P_n such that $||(P_n - P)f^2||_{\mathcal{F}_\infty}$ converges to 0. Take n sufficiently large so that the supremum is bounded by ε^2. We know that $N(\varepsilon, \mathcal{F}, L_2(P_n))$ is finite (this can be shown along the lines of the fact shown in the appendix). Any ε-net for $\mathcal{F}$ in $L_2(P_n)$ is a $\sqrt{2}\varepsilon$-net in $L_2(P)$, since $P(f-g)^2 \leq \varepsilon^2 + P_n(f-g)^2 \leq 2\varepsilon^2$. This concludes the proof.

Example. The set $\mathcal{F}$ of all indicator functions $\mathbb{1}_{(-\infty,t]}$ of cells in $\mathbb{R}$ satisfies:

$$N(\varepsilon, \mathcal{F}, L_2(Q)) \leq N_{[]}(\varepsilon^2, \mathcal{F}, L_1(Q)) \leq \frac{2}{\varepsilon^2}$$

for any probability measure and $\varepsilon \leq 1$, for any probability measure Q. The first inequality follows from the fact that, $\sqrt{Q(f-g)^2} = \sqrt{Q|f-g|}$ so that if $Q|f-g| \leq \varepsilon^2$ we would have $\sqrt{Q(f-g)^2} \leq \varepsilon$. Therefore $N(\varepsilon, \mathcal{F}, L_2(Q)) \leq N(\varepsilon^2, \mathcal{F}, L_1(Q))$. Furthermore if we have a bracket on the set $\mathcal{F} - [l, u]$ we can put a ball with a center at the function $\frac{l+u}{2}$ and radius ε^2, this ball will obviously cover all the functions within the bracket. And thus $N(\varepsilon^2, \mathcal{F}, L_1(Q)) \leq N_{[]}(\varepsilon^2, \mathcal{F}, L_1(Q))$. The right inequality follows because the total probability mass is 1, and we are splitting it in intervals of ε^2 we have at most $1/\varepsilon^2 + 1 \leq 2/\varepsilon^2$ brackets. Therefore for the uniform entropy in $[0, 1]$ we have $\lesssim \int_0^1 \log(1/\varepsilon)d\varepsilon < \infty$. Of course when $\varepsilon > 1$ the number of brackets required is only 1 so that it's 0 in the integral. Thus the class $\mathcal{F}$ would be Donsker if we can show that $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2$ are P-measurable. This is not hard however, since $\mathbb{Q}$ is dense in $\mathbb{R}$ so we can get to the supremums by countable number of operations.

Compare this result to Slide 55, noteset 2!

We consider much more general classes that satisfy the uniform entropy condition next time.

2 Uniform Entropy Numbers

The uniform entropy integral will be convergent if:

$$\sup_Q \log N(\varepsilon||F||_{Q,2}, \mathcal{F}, L_2(Q)) \leq K\left(\frac{1}{\varepsilon}\right)^{2-\delta}$$

For some $\delta > 0$. Here we consider classes that satisfy a much stronger condition, namely:

$$\sup_Q N(\varepsilon||F||_{Q,2}, \mathcal{F}, L_2(Q)) \leq K\left(\frac{1}{\varepsilon}\right)^{V}, 0 < \varepsilon < 1$$

2.1 VC Classes of Sets

VC stands for Vapnik and Červonenkis, who were the first to study these sets.

Let $\mathcal{C}$ be a collection of subsets of a set $\mathcal{X}$. An arbitrary set of n points $\{x_1, \ldots, x_n\}$ possesses 2^n subsets. Say that $\mathcal{C}$, *picks out* a certain subset from $\{x_1, \ldots, x_n\}$ if this subset takes the form $C \cap \{x_1, \ldots, x_n\}$ for some $C \in \mathcal{C}$. $\mathcal{C}$ is said to *shatter* $\{x_1, \ldots, x_n\}$ if all possible 2^n subsets can be picked out by $\mathcal{C}$.

The *VC-index* $V(\mathcal{C})$ of the collection $\mathcal{C}$ is the smallest n for which there is no set of size n, which is shattered by $\mathcal{C}$. More formally we can defined the VC-index by:

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = \#\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}$$
$$V(\mathcal{C}) = \inf\{n : \max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) < 2^n\}$$

Note here that it is possible to have a set with $V(\mathcal{C}) = \infty$. Here we will be focusing only on sets with finite $V(\mathcal{C})$ index. These collections of sets are called *VC-classes*.

Example (Cells in $\mathbb{R}^d$). The collections of all cells of the form $(-\infty, c]$ in $\mathbb{R}$ shatters no two-point set $\{x_1, x_2\}$. This is because we can't pick out the larger of the two points only. Thus the $V(\mathcal{C})$ index of this collection is 2. The collection of sets $(a, b]$ for $a, b \in \mathbb{R}$ shatters every two point set, but it cannot shatter any set consisting of three points $\{x_1, x_2, x_3\}$, because it can't pick out the set $\{x_1, x_3\}$ (assuming $x_1 < x_2 < x_3$). Thus the VC-index of this collection is 3. Similarly it can be shown that the VC indexes of cells in $\mathbb{R}^d$ of the first type is $d + 1$ and of the second type is $2d + 1$. In the appendix we sketch a quick proof of the first of these facts, the other is left as an exercise.

VC classes are important because of the following important combinatorial result: the number of subsets shattered by a class $\mathcal{C}$ is at least the number of subsets picked out by $\mathcal{C}$. Formally we express this statement as:

Lemma. Let $\{x_1, \ldots, x_n\}$ be arbitrary points. Then the total number of subsets $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ picked out by $\mathcal{C}$ is bounded above by the number of subsets of $\{x_1, \ldots, x_n\}$ shattered by $\mathcal{C}$.

This result is known as Sauer-Shelah lemma, even though it was first proved by Vapnik and Červonenkis.

Proof. Assume WLOG that every C is a subset of the given set of points, such that $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ is the cardinality of $\mathcal{C}$.

Call a collection of sets in $\mathcal{C}$ *hereditary* if it has the property that $B \in \mathcal{C}$ whenever there exists C such that $B \subset C \in \mathcal{C}$. Each set in a hereditary collection is shattered (each of its subsets subsets are part of $\mathcal{C}$). This means that in a hereditary collection of sets the total number of shattered sets is at least $|\mathcal{C}|$ which of course bounds the number of sets that $\mathcal{C}$ can pick out. The proof proceeds

to show that any collection of sets $\mathcal{C}$ can be transformed in to hereditary collection of sets, without changing its cardinality and without increasing the number of shattered sets.

For a fixed $1 \leq i \leq n$, consider the following operation on the collection. Define:

$$T_i(C) = \begin{cases} C - \{x_i\}, & \text{if } C - \{x_i\} \notin \mathcal{C} \\ C, & \text{if } C - \{x_i\} \in \mathcal{C} \end{cases}$$

Or in words, T_i deletes the i^{th} element of the set C if this creates a new set in $\mathcal{C}$. Therefore if a set doesn't contain x_i it will be left untouched by this operation, and if a set did contain x_i it will be deleted only if this creates a new set.

Note several facts about this operation on the whole collection of sets $\mathcal{C}$. First, $T_i(\mathcal{C})$ is of the same cardinality as $\mathcal{C}$ ($|T_i(\mathcal{C})| = |\mathcal{C}|$) because the map T_i is a bijection.

Second, note that if a subset of $\{x_1, \ldots, x_n\}$ is shattered by $T_i(\mathcal{C})$ it is shattered by $\mathcal{C}$. To see this take a subset $A \subset \{x_1, \ldots, x_n\}$, which is shattered by $T_i(\mathcal{C})$. If $x_i \notin A$ we have that $C \cap A = T_i(C) \cap A$ for $C \in \mathcal{C}$, and therefore $T_i(\mathcal{C})$ shatteres A if and only if $\mathcal{C}$ shatters A. Now consider the case when $x_i \in A$. If $T_i(\mathcal{C})$ shatters A it follows for that each subset $B \subset A$ then since $B \cup \{x_i\} \subset A$ we have $B \cup \{x_i\} = A \cap T_i(C)$ for some $C \in \mathcal{C}$. IIt follows that $x_i \in T_i(C)$ and therefore $T_i(C) = C$. This means that both $C, C - \{x_i\} \in \mathcal{C}$.Therefore we have the following representations of the sets $B \cup \{x_i\} = A \cap C$ and $B - \{x_i\} = A \cap (C - \{x_i\})$. Finally note that exactly one of these two sets is B.

The last two facts showed that applying T_i to the collection $\mathcal{C}$ preserves the cardinality and doesn't increase the number of shattering sets. Therefore the same is valid for the transformation $T_1 \circ T_2 \circ \cdots \circ T_n$. We can apply this operator until the collection of sets stops to change. This will happen until at most $\sum_{C \in \mathcal{C}} |C|$ number of steps, because $\sum_C |T_i(C)| < \sum_C |C|$ when the two collections are different (when they are different at least one set has lost an element). Finally note that the stable collection $\mathcal{D}$ we end up with is hereditary. This is the case since for any element $D \in \mathcal{D}$, the sets $D - \{x_i\} \in \mathcal{D}$ for all i. Finally this finishes the proof.

Corollary. For a VC-class of sets of index $V(\mathcal{C})$, one has:

$$\max_{x_1, \ldots, x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq \sum_{j=0}^{V(\mathcal{C})-1} \binom{n}{j}$$

And further, for $n \geq V(\mathcal{C}) - 1$ we have:

$$\sum_{j=0}^{V(\mathcal{C})-1} \binom{n}{j} \leq \left(\frac{ne}{V(\mathcal{C}) - 1} \right)^{V(\mathcal{C})-1}$$

Proof. Indeed for a VC-class of index $V(\mathcal{C})$ it doesn't shatter any set of size at least $V(\mathcal{C})$. Therefore from the previous lemma we directly obtain the bound of the first inequality. The second inequality is easily verified through a Taylor expansion of $e^{V(\mathcal{C})-1}$

A Some Details

We now show why $N(\varepsilon, \mathcal{F}_\infty, L_2(Q)) \leq N^2(\varepsilon/2, \mathcal{F}, L_2(Q))$ for all measures Q. Take an $\epsilon/2$ covering of $\mathcal{F}$ consisting of $N = N(\varepsilon/2, \mathcal{F}, L_2(Q))$. Denote with $S = \{f_1, f_2, \ldots, f_N\}$ the covering set. We show that the set $\{f - g : f, g \in S\}$, which are N^2 points, is an ε-cover of $\mathcal{F}_\infty$. Take any point $h \in \mathcal{F}_\infty$. We know that $h = s - t$ for some functions $s, t \in \mathcal{F}$. Take f and g such that $\sqrt{Q(s-f)^2} < \varepsilon/2$ and $\sqrt{Q(t-g)^2} < \varepsilon/2$. Use triangle inequality to conclude that $\sqrt{Q((s-t)-(f-g))^2} < \varepsilon$, which concludes the proof.

Here we show that if $N(\varepsilon||2F||_n^2, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$ is finite for all ε, it must be the case that $\log N(\varepsilon, \mathcal{F}_{\infty,M}^2, L_1(\mathbb{P}_n)) = o_P^*(n)$, in fact this number turns out to also be finite. First note that $N(\varepsilon, \mathcal{F}_{\infty,M}^2, L_1(\mathbb{P}_n)) \leq N(\varepsilon, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$. Fix $\varepsilon > 0$. Since $P^*F^2 < \infty$, then there exists S such that $||2F||_n^2 \leq S$ with probability 1. Thus $N(\varepsilon S, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n)) \leq N(\varepsilon||2F||_n^2, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$ with probability 1, and thus since $\varepsilon > 0$ was arbitrary $N(\varepsilon, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$ is $O_P^*(1)$, and therefore $\log N(\varepsilon, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n)) = o_P^*(n)$, which implies that $\log N(\varepsilon, \mathcal{F}_{\infty,M}^2, L_1(\mathbb{P}_n)) = o_P^*(n)$.

Why is the VC-index of $(-\infty, x], x \in \mathbb{R}^d$ – d+1? First note that the set of d points $\{[0, \ldots, \underbrace{1}_{i}, \ldots, 0]\}_{i=1}^d$ can be shattered. For the other part we show that no set of size $d+1$ in $\mathbb{R}^d$ can be shattered. Take the union of points such that for each index they have the largest number. These points are at most d. There is no way we can shatter this set without including all points, which would be a contradiction. This finishes the proof.