VC classes

Lab 10, BIO251

04/28/2014

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 $e^{V(\mathcal{C})-1}$.

Theorem. There exists a universal constant K such that for any VC-class $\mathcal{C}$ of sets, any probability measure Q, any $r \geq 1$, and $0 < \varepsilon < 1$, we have:

$$N(\varepsilon, \mathcal{C}, L_r(Q)) \leq KV(\mathcal{C})(4e)^{V(\mathcal{C})} \left(\frac{1}{\varepsilon} \right)^{r(V(\mathcal{C})-1)}$$

The proof of this theorem is quite involved. However, there exists a simple proof for the following slight weakening which we will consider. Same statement as before but for any $\delta > 0$, we can show that there exists K depending on $V(\mathcal{C})$ and δ only such that:

$$N(\varepsilon, \mathcal{C}, L_r(Q)) \leq K \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{C})-1+\delta)}$$

Proof: Note that $||\mathbb{1}_C - \mathbb{1}_C||_Q = Q^{1/r}(C \Delta D)$. Thus it suffices to show the problem for $r = 1$, and the rest is just a consequence. Take any subcollection $C_1, \ldots, C_m \in \mathcal{C}$, with $Q(C_i \Delta C_j) > \varepsilon$ for any $i \neq j$. Generate a sample $X_1, X_2, \ldots, X_n$ from Q. Two sets C_i and C_j pick out the the same subset from a realization of the sample, if and only if no X_k falls in the symmetric difference $C_i \Delta C_j$. Thus if every symmetric difference contains at least one point from the sample, then all C_i will pick out a different subset from the sample. In that case $\mathcal{C}$ picks out at least m subsets of the sample $X_1, \ldots, X_n$. The probability that this event does not occur is bounded by:

$$\begin{aligned}\sum_{i<j} Q(X_k \notin C_i \Delta C_j \text{ for every } k) &\leq \binom{m}{2}(1 - \max_{i,j} Q(C_i \Delta C_j))^n \\ &\leq \binom{m}{2}(1-\varepsilon)^n\end{aligned}$$

The last expression is less than 1 for n large enough. For such n there exists a set of n points from which $\mathcal{C}$ picks out at least m subsets. In particular for $n > -\log\binom{m}{2}/\log(1-\varepsilon)$, we have by the Corollary that:

$$m \leq \max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq K n^{V(\mathcal{C})-1}$$

With the constant K depending on $V(\mathcal{C})$ only. Note that we have the inequality $-\log(1-\varepsilon) > \varepsilon$, and since $\log\binom{m}{2} \leq 3 \log m$ we can take $n = 3(\log m)/\varepsilon$, and thus conlcude that:

$$m \leq K \left(\frac{3 \log m}{\varepsilon}\right)^{V(\mathcal{C})-1}$$

Of course for any $\delta > 0$ we know that $\log m < S m^{\kappa}$ for some S depending on κ and therefore, if we select $\kappa = \tilde{\delta} = \frac{\delta}{(V(\mathcal{C})-1+\delta)(V(\mathcal{C})-1)}$ we get, that $(\log m)^{V(\mathcal{C})-1} \leq S m^{\frac{\delta}{(V(\mathcal{C})-1+\delta)}}$. Finally, after putting everything together, we get the desired result.

A Some Details

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.