INTRO TO VAPNIK-CHERVONENKIS THEORY

LIAM HODGKINSON

So far in our chosen text, we have seen that all finite hypothesis classes are PAC-learnable, and that the trivial hypothesis class is not PAC-learnable. Firstly, Chapter 6 highlights that finiteness of the hypothesis class is not a necessary condition for PAC-learnability, so the cardinality of the hypothesis class is not an effective measure of the sample complexity function. The objective of Chapter 6 is to find middle ground and ascertain exactly what makes a hypothesis class (agnostic) PAC-learnable. While this is primarily of theoretical interest, there are other motivations: the discretisation trick may produce a gross over-estimate of the required number of samples to train, and provides little in the way of helping choose one hypothesis class over another. The characterisation to be discussed is *equivalent* to PAC-learnability, and provides an excellent description for the size of an hypothesis class. The idea was first introduced by Vapnik and Chervonenkis, and stems from the well-studied theory of uniform approximation of empirical distributions (the connection between this area and machine learning has already been previously discussed and utilised). Indeed, the idea of the VC dimension reportedly led to the development of support vector machines, which are widely utilised even today.

Fortunately, I am quite familiar with the concept of the VC dimension (that I'm speaking today is not a coincidence...), as it arises a lot in my area of expertise as well: the development of concentration inequalities for uniform convergence of measure. We saw in Chapter 4 that this is intimately connected with machine learning. Recall that

Definition (Uniform Convergence)**.** A hypothesis class $\mathcal{H}$ has the *uniform convergence property* if for every $\epsilon, \delta > 0$, and every probability distribution $\mathcal{D}$ on the sample data, there is an integer $m_{\mathcal{H}}(\epsilon, \delta)$ such that if S is a sample of size $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ or greater (with iid samples according to $\mathcal{D}$), then

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |L_S(h) - \mathbb{E}L_S(h)| > \epsilon\right) < \delta,$$

recalling that $\mathbb{E}L_S(h) = L_{\mathcal{D}}(h)$.

In a very quick proof, it was shown that

$$\textit{uniform convergence property} \implies \textit{agnostic PAC learnable}$$

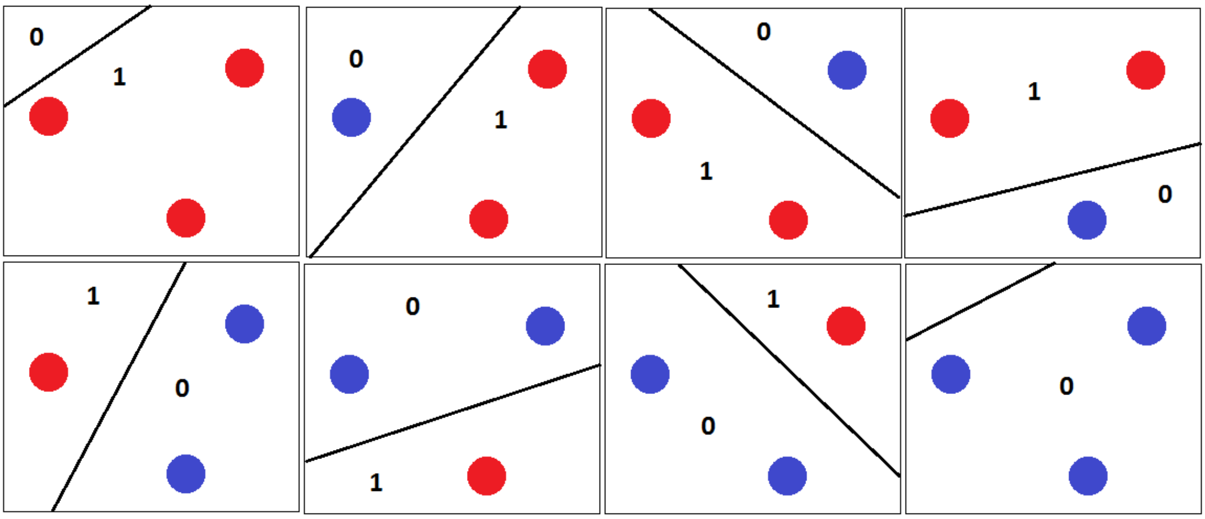
In the previous chapter, we found that if our hypothesis class $\mathcal{H}$ is too rich, the estimation error will not decrease in the number of samples. The VC dimension provides a measure of the complexity of $\mathcal{H}$, and how close it can get to describing a pathological distribution. It is the opposite of the idea of "density of functions" in analysis (e.g. Stone-Weierstrass Theorem). And much like the Stone-Weierstrass Theorem in analysis, the definition of VC dimension relies on a "separation of points" concept.

Definition. A class $\mathcal{H}$ *shatters* the set C if for any subset $C' \subset C$, there is a function $h \in \mathcal{H}$ such that $h(c) = 1$ for $c \in C'$ and $h(c) = 0$ for $c \notin C'$. The VC-dimension of $\mathcal{H}$ is the size of the largest set C which can be shattered by $\mathcal{H}$.

To show that the VC dimension of $\mathcal{H}$ is d, you need to show that

(1) There is a set of size d that can be shattered by $\mathcal{H}$
(2) No set of size $d+1$ can be shattered by $\mathcal{H}$, that is, there is some colouring of the points in C that $\mathcal{H}$ cannot possibly assign.

Example (Halfplanes)**.** Consider a plane in d-dimensional space. The case $d = 1$ has VC dimension $V(\mathcal{H}) = 2$ for the same reasons as the interval case in the book (in fact, the VC dimension of balls corresponds exactly to the VC dimension of halfplanes!). The following image shows that if $d = 2$, then the VC-dimension of the set of halfplanes is $V(\mathcal{H}) \geq 3$.



However, no set of four points is shattered by the set of halfplanes: choosing any three of the points, you can colour them in such a way that by constructing any line passing between them (according to the above picture), the remaining point will lie on the side of the pair of coloured points. The remaining point need now only be coloured in the opposing colour. Thus, $V(\mathcal{H}) = 3$. Indeed, the VC dimension of the space of d-dimensional halfplanes is $d+1$. To see this, observe that the plane forming the boundary of the halfplane can be represented as a linear function (or classifier...)

$$f\left(x_1, \ldots, x_d\right) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i,$$

which has $d+1$ free parameters and can fit exactly to $k \leq d$ points. Taking $C' \subsetneq C \subset \mathbb{R}^d$ with $|C| = d+1$, you can fit a plane to match every point in C' and then perturb slightly to obtain a plane separating points in C' from those in $C \backslash C'$. In this way, the VC dimension can be seen as a similar idea to the "degrees of freedom" in statistics, however, the number of parameters *does not* imply the VC dimension in general.

Other examples include:

- The class of axis-aligned rectangles in d dimensions has VC-dimension $2d$.
- The class of rectangles in d dimensions has infinite VC dimension.
- The class of closed balls (or spheres) in d dimensions has VC dimension $d+1$.
- A finite class of functions $\mathcal{H}$ has VC dimension bounded above by $\log_2 |\mathcal{H}|$.
- The class of polynomials has infinite VC dimension (polynomial fitting).

As a consequence of the proof of the No Free Lunch Theorem, if $V(\mathcal{H}) \geq d$, then for $\epsilon < \frac{1}{8}$ and $\delta < \frac{1}{7}$, $m_{\mathcal{H}}(\epsilon, \delta) \geq \frac{1}{2} d$ (probably; there is a more accurate refinement of this in the back of the book). Thus,

$$\textit{agnostic PAC learnable} \implies \textit{finite VC dimension}$$

It now only remains to show that finite VC dimension implies PAC learnability, for which we introduce the idea of shatter coefficients. The shatter coefficients $S_{\mathcal{H}}(m)$ are defined by

$$S_{\mathcal{H}}(m) := \max_{c_1, \ldots, c_m \in \mathcal{X}} \left|\left\{(h(c_1), \ldots, h(c_m)) \,:\, h \in \mathcal{H}\right\}\right|.$$

For brevity, let $\mathcal{H}(C)$ denote the set inside the cardinality, for $C = (c_1, \ldots, c_m)$. Recall that in the earlier proofs for uniform convergence, we applied the trivial union bound over functions $h \in \mathcal{H}$ by introducing the factor $|\mathcal{H}|$. As shown in Theorem 6.11, now we do the same thing, but with the shatter coefficients instead. That's it. The reason this works is that for a sample S of size m, there exists a representative set $C = (c_1, \ldots, c_m)$ such that $\mathcal{H}(S) \subset \mathcal{H}(C)$ and $|\mathcal{H}(C)| = S_{\mathcal{H}}(m)$. Thus,

$$\sup_{h \in \mathcal{H}} \text{ may be replaced by } \max_{h \in \mathcal{H}(C)},$$

and applying the union bound gives $|\mathcal{H}(C)| = S_{\mathcal{H}}(m)$ instead of $|\mathcal{H}|$. For m less than the VC dimension, the shatter coefficients grow exponentially. But fortunately, beyond the VC dimension, they grow only at polynomial rate, and this is what makes the VC dimension powerful.

Lemma (Sauer's Lemma). *Let $\mathcal{H}$ be a hypothesis class with VC dimension $V(\mathcal{H}) \leq d < \infty$. Then for all m,*

$$S_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq (m+1)^d .$$

The traditional proof of Sauer's lemma is inductive in nature, and is evidently combinatorial. Thus, $\log |S_{\mathcal{H}}(m)| \leq d \cdot \log(m+1)$, implying that

finite VC dimension $\implies$ *uniform convergence property*

Thus, the following fundamental theorem is obtained.

Theorem (Fundamental Theorem of Statistical Learning). *The following statements are equivalent:*

(1) $\mathcal{H}$ has finite VC dimension
(2) $\mathcal{H}$ has the uniform convergence property
(3) $\mathcal{H}$ is agnostic PAC learnable
(4) $\mathcal{H}$ is PAC learnable

Moreover, if $\mathcal{H}$ has finite VC dimension, then the sample complexity function satisfies

$$C_1 \cdot \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \cdot \frac{d + \log(1/\delta)}{\epsilon^2}.$$

Proofs of the upper inequality seem to follow a specific procedure: first, using McDiarmid's inequality,

$$m_{\mathcal{H}}(\epsilon, \delta) \leq C \left[\mathrm{Rad}(\mathcal{H}) + \frac{\log(1/\delta)}{\epsilon^2} \right],$$

where $\mathrm{Rad}(\mathcal{H})$ is the Rademacher complexity. The Rademacher complexity can be bounded by covering numbers (chaining), and then by the Vapnik-Chervonenkis dimension, or by more sophisticated techniques (e.g. entropy bounds). Indeed,

$$\mathrm{Rad}(\mathcal{H}) \leq \frac{c_1}{\sqrt{n}} \int_0^H \sqrt{\log N_r(\mathcal{H})}\mathrm{d}r \leq c_2 H \sqrt{\frac{\mathscr{V}(\mathcal{H})}{n}} \leq c_3 H \sqrt{\frac{\log |\mathcal{H}|}{n}}.$$

A full proof of the Vapnik-Chervonenkis inequality, which implies this theorem, can be found in "Combinatorial Methods in Density Estimation" by Devroye and Lugosi. The lower bound essentially follows from ideas for the No Free Lunch Theorem.