What is...

Vapnik-Chervonenkis Dimension?

Samuel C. Tenka

The Cake Problem

Mathematicians and bakers alike know the sequence $1, 2, 4, 8, 16, \cdots$ by heart. It continues, of course, with 31, for its nth element $p(n)$ counts the pieces obtained from a disk-shaped cake by cutting along all $\binom{n}{2}$ lines determined by n points placed generically on the cake's perimeter.

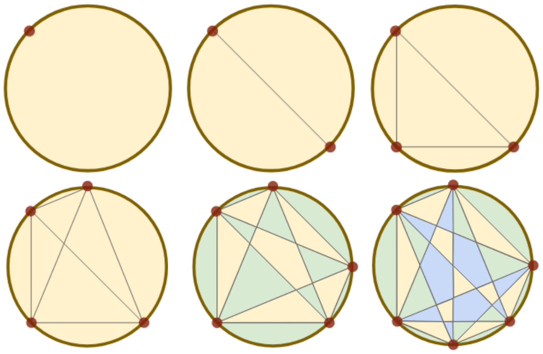

Figure 1: *Cakes for $n = 1, \cdots, 6$. The $n = 4$ cake (bottom left) has $p(4) = 8$ pieces. We color some pieces to make them easier to see and to count. $p(6)$ is clearly odd: the pieces besides the central yellow triangle group into sets of six.*

Rather than growing exponentially, $p(n)$ is a polynomial. We may compute $p(n)$ by regarding each sliced cake as a planar graph, observing that each interior point is determined by two cuts and hence by one of $\binom{n}{4}$ many sets of 4 perimeter points, and then applying Euler's polyhedron formula. One finds that $p(n)$ is $\binom{n-1}{0} + \cdots + \binom{n-1}{4}$, which explains why $p(n)$ initially coincides with 2^{n-1}.

This example, like many others in mathematics and in science, serves as a warning and a mystery: patterns do not always generalize. But then — *how is generalizing from data possible at all*?

Learning and Generalization

In general, we wonder: *if from a collection $\mathcal{H}$ of possible patterns we find some $f \in \mathcal{H}$ that matches N observed data points, when should we expect that f matches unseen data*? This question motivates statistical learning theory and the foundations of machine learning.

We may frame the problem in the setting of image classification, where $\mathcal{X}$ is a space of images, $\{\pm 1\} = \{\mathrm{Cow}, \mathrm{Dog}\}$ is a set of (for simplicity, two) labels, and we seek a classifier $f : \mathcal{X} \to \{\pm 1\}$ that accords with nature. More precisely, we posit a probability distribution $\mathcal{D}$ over the space $\mathcal{X} \times \{\pm 1\}$ of input-output pairs and we let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a set of (measurable) functions. If $\mathcal{S} \sim \mathcal{D}^N$ denotes a sequence of N observations drawn independently from $\mathcal{D}$, the **in-sample error** of $f \in \mathcal{H}$ is

$$\mathrm{trn}_{\mathcal{S}}(f) = \mathbb{P}_{(x,y)\sim\mathcal{S}}[f(x) \neq y]$$

and the **out-of-sample error** is

$$\mathrm{tst}(f) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$$

A **learning rule** $\mathcal{L} : (\mathcal{X} \times \{\pm 1\})^N \to \mathcal{H}$ maps $\mathcal{S}$s to fs. We wonder when a small in-sample error implies a small out-of-sample error, that is, when we may bound the **generalization gap**

$$\mathrm{gap}_{\mathcal{S}}(\mathcal{L}) = \mathrm{tst}(\mathcal{L}(\mathcal{S})) - \mathrm{trn}_{\mathcal{S}}(\mathcal{L}(\mathcal{S}))$$

In degenerate cases where $\mathcal{L}(\mathcal{S})$ and $\mathcal{S}$ are independent, $\mathrm{trn}_{\mathcal{S}}(\mathcal{L}(\mathcal{S}))$ is an unbiased estimator for $\mathrm{tst}(\mathcal{L}(\mathcal{S}))$, and by laws of large numbers, $\mathrm{gap}_{\mathcal{S}}$ is small for large N. We wonder: *can we still control the gap when $\mathcal{L}(\mathcal{S})$ depends on $\mathcal{S}$*? **Vapnik-Chervonenkis theory** answers this question affirmatively for sufficiently nice $\mathcal{H}$.

The two ingredients are *concentration* and *symmetrization.*

Concentration of Measure

Lemma 1 (Chernoff)**.** *The fraction of heads among N i.i.d. flips of a biased coin exceeds its mean p by more than g with probability at most* $\exp(-Ng^2)$*, for all $p, g \in [0, 1]$.*

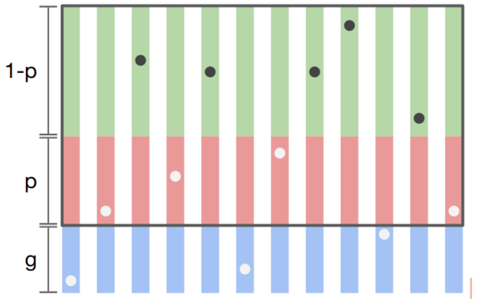

Figure 2: *We randomly select points on N vertical sticks. Each stick has three parts:* ***green*** *with length $1 - p$,* ***red*** *with length p, and* ***blue*** *with length g. We call non-blue points* ***boxed*** *and non-green points* ***hollow****.*

Proof. We'll switch viewpoints: flipping a coin is like choosing a boxed point on a stick where green means tails and red means heads (Figure 2). We'll show that with high probability less than $M_0 = (p + g)N$ flips heads. That is — given that all points are boxed — less than M_0 points are red. For any $M \geq M_0$:

$$\begin{aligned} &\mathbb{P}[M \text{ are red} \mid \text{all are boxed}] \\ = \;&\mathbb{P}[M \text{ red and all are boxed}] \;/\; \mathbb{P}[\text{all are boxed}] \\ = \;&\mathbb{P}[M \text{ hollow}] \cdot \frac{\mathbb{P}[\text{all hollows are red} \mid M \text{ hollow}]}{\mathbb{P}[\text{all are boxed}]} \\ = \;&\mathbb{P}[M \text{ hollow}] \cdot (1 + g/p)^{-M} \;/\; (1 + g)^{-N} \end{aligned}$$

Since the above holds for all $M \geq M_0$, the chance of too many heads is:

$$\begin{aligned} &\mathbb{P}[\text{at least } M_0 \text{ are red} \mid \text{all are boxed}] \\ \leq\; &(1 + g/p)^{-M_0} / (1 + g)^{-N} \end{aligned}$$

We finish by plugging in $M_0 = (p+g)N$, bounding $1 + x \leq \exp(x)$, and simplifying. □

The Chernoff bound, proved above with suboptimal constants, is one of the most basic **concentration inequalities**.

For any $f \in \mathcal{H}$, $\text{trn}_{\mathcal{S}}(f)$ is the average of N independent Bernoullis of mean $\text{tst}(f)$. So for $\mathcal{H}$ finite and N large, the gap is probably small:

$$\begin{aligned} &\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}[\text{gap}_{\mathcal{S}}(\mathcal{L}) \geq g] \\ \leq\; &\sum_{f\in\mathcal{H}} \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}[\text{trn}_{\mathcal{S}}(f) \geq \text{tst}(f) + g] \\ \leq\; &|\mathcal{H}| \cdot \exp(-Ng^2) \end{aligned}$$

For example, if $\mathcal{H}$ is parameterized by P numbers, each represented on a computer by 32 bits, then $|\mathcal{H}| \leq 2^P$ and, with probability $1 - \delta$, the gap is no more than

$$\sqrt{\log(2/\delta) \cdot 32P/N}$$

But shouldn't 32 bits or 64 bits or infinitely many bits yield similar behavior? Intuitively, the $\mathcal{H}$s used in practice — for instance, linear models or neural networks — depend smoothly on their parameters; tiny changes in the parameters yield practically the same candidate, so $\mathcal{H}$'s cardinality is not an apt measure of its size. As we will see, the V-C dimension measures $\mathcal{H}$ more subtly.

Symmetrization

The key observation is that, even though $\mathcal{H}$ may be infinite, the restriction $\mathcal{H}_{\mathcal{S}} = \{f|_{\mathcal{S}} : f \in \mathcal{H}\}$ is finite for finite $\mathcal{S}$. So let us fix $f \in \mathcal{H}$ and estimate $\text{tst}(f)$ by $\text{trn}_{\check{\mathcal{S}}}(f)$ for $(\mathcal{S}, \check{\mathcal{S}}) \sim \mathcal{D}^{2N}$ drawn independently. By Chernoff, tst and $\text{trn}_{\check{\mathcal{S}}}$ are probably close: provided $g \geq 2/\sqrt{N}$,

$$\begin{aligned} &\mathbb{P}[\text{trn}_{\mathcal{S}} + g \leq \text{tst}] \\ = \;&\mathbb{P}[\text{trn}_{\mathcal{S}} + g \leq \text{tst} \mid \text{tst} \leq \text{trn}_{\check{\mathcal{S}}} + g/2] \\ \leq \;&\mathbb{P}[\text{trn}_{\mathcal{S}} + g/2 \leq \text{trn}_{\check{\mathcal{S}}} \mid \text{tst} \leq \text{trn}_{\check{\mathcal{S}}} + g/2] \\ \leq \;&\mathbb{P}[\text{trn}_{\mathcal{S}} + g/2 \leq \text{trn}_{\check{\mathcal{S}}}] \cdot 2 \end{aligned}$$

Imagining $\mathcal{S}$ as sampled *without replacement* from $\mathcal{S} \sqcup \check{\mathcal{S}}$ and applying a variant of Chernoff, we find that $g \leq \text{gap}_{\mathcal{S}}(\mathcal{L})$ with chance at most

$$\max_{|\mathcal{S}|=|\check{\mathcal{S}}|=N} |\mathcal{H}_{\mathcal{S}\sqcup\check{\mathcal{S}}}| \;\cdot\; \exp(-Ng^2/16) \cdot 2$$

Finally, to show that the gap is usually small, we need only bound $H(n) = \max_{|\mathcal{S}|=n} |\mathcal{H}_{\mathcal{S}}|$.

Clearly, $H(n) \leq 2^n$. In fact, this bound is never somewhat tight: depending on $\mathcal{H}$, it either is an equality or extremely loose!

Indeed, consider $\mathcal{H}_S$ for $|S| = n$. Ordering S, let us write each $f \in \mathcal{H}_S$ as a string of +s and −s. To count these strings, we will first translate them from the alphabet $\{+, -\}$ to the alphabet $\{\blacksquare, \square\}$. Intuitively, ■ represents "surprisingly +". More precisely, working from left to right, whenever two (partially translated) strings differ **only** in their leftmost untranslated coordinate we overwrite the + version's + by ■. Otherwise, we overwrite by □.

+−−−		□−−−		□□−−		□□□−		□□□□
−+−−		□+−−		□■−−		□■□−		□■□□
−−+−		□−+−		□□+−		□□■−		□□■□
−−−+	→	□−−+	→	□□−+	→	□□□+	→	□□□■
−−++		□−++		□□++		□□■+		□□■■
−+++		□+++		□■++		□■□+		□■□■
++++		■+++		■□++		■□□+		■□□□

Figure 3: Translating elements of H_S (left) to strings of choice points (right). Each row corresponds to one of 7 candidates and each column corresponds to one of of 4 datapoints. We color pairs of strings that differ at-and-only-at their leftmost untranslated coordinate.

Each step of translation keeps distinct strings distinct. Moreover, whenever some k indices $T \subseteq S$ of a translated string are ■s, $|\mathcal{H}_T| = 2^k$. This is because ■s mark choice points where the candidates attain both + and −. Now, **either** $H(n) = 2^n$ for all n, **or** $H(k+1) \neq 2^{k+1}$ for some k. In the latter case, no translated string may have k or more ■s. Thus $\mathcal{H}_S$ contains no more strings than there are subsets in S of size $\leq k$. Therefore,

$$H(n) \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k} \leq (n+2)^k$$

As with Cake, what might grown as 2^n grows only polynomially.

In short, if some $H(k+1) \neq 2^{k+1}$, then $\text{gap}_{\mathcal{S}}(\mathcal{L}) \geq g$ with chance at most

$$(2N+2)^k \cdot \exp(-Ng^2/16) \cdot 2$$

Because this tends to 0 as N grows, generalizing from data is possible.

Since k appears as an exponent, it is morally a measure of dimension:

Definition 1. The **Vapnik-Chervonenkis dimension** of a set $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ is the supremal k for which $H(k) = \max_{|S|=k} |\mathcal{H}_S| = 2^k$.

The following theorem, whose "only if" direction we have sketched above, highlights the V-C dimension's importance to learning theory:

Theorem 1 (Vapnik and Chervonenkis, 1971)**.** *The V-C dimension of $\mathcal{H}$ is finite if and only if for all data distributions $\mathcal{D}$, learning rules $\mathcal{L}$, and gap bounds $g > 0$, the chance that gap$_{\mathcal{S}}(\mathcal{L})$ exceeds g tends to 0 as $N = |\mathcal{S}|$ grows.*

Statistical Learning Theory

References

The use of three-segment sticks and $\{\blacksquare, \square\}$-encoding to give short, elementary proofs of Chernoff's bound and the VC Theorem are, as far as the author knows, new. That said, they correspond directly to classical proofs involving moment generating functions and induction on sums of binomial coefficients, respectively.