What is...

the Vapnik-Chervonenkis Dimension?

Samuel C. Tenka

Wetzel's 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 generic points on the cake's boundary.

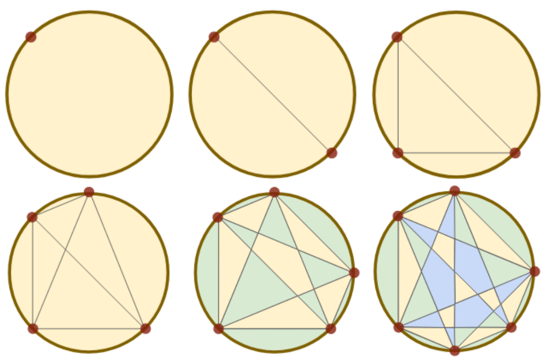

Figure 1: *Cakes for $n = 1, \cdots, 6$. The $n = 4$ cake (bottom left) has $p(4) = 8$ pieces. We color some pieces to make them easier to see and to count. $p(6)$ is clearly odd: the pieces besides the central yellow triangle group into sets of six.*

Rather than growing exponentially, $p(n)$ is a polynomial [2]. 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 boundary 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 learning from finite data possible at all*?

Learning and Generalization

We thus 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 machine learning theory and guides machine learning practice.

We may frame the problem in the setting of image classification, where $\mathcal{X}$ is a space of images, $\{\pm 1\} = \{\text{Cow}, \text{Dog}\}$ 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 $\{\pm 1\} \times \mathcal{X}$ of labeled images 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

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

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

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

A **learning rule** $\mathcal{L} : (\{\pm 1\} \times \mathcal{X})^N \to \mathcal{H}$ maps $\mathcal{S}$s to fs. Often, $\mathcal{L}$ is induced by an approximate minimization of the in-sample error. However, as the goal of machine learning is typically to achieve out-of-sample error, we wonder when a small in-sample error implies a small out-of-sample error, that is, when we may bound the **generalization gap**

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

In degenerate cases where $\mathcal{L}(\mathcal{S})$ and $\mathcal{S}$ are independent, $\text{trn}_{\mathcal{S}}(\mathcal{L}(\mathcal{S}))$ is an unbiased estimator for $\text{tst}(\mathcal{L}(\mathcal{S}))$; by laws of large numbers, $\text{gap}_{\mathcal{S}}$ is small for large N. The key question is: *can we control the gap when $\mathcal{L}(\mathcal{S})$ depends on $\mathcal{S}$*?

The answer is affirmative when $\mathcal{H}$ is "finite-dimensional" for a certain notion of dimension. The two ingredients in the story are *concentration* and *symmetrization*.

Concentration

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* $\exp(-Ng^2)$*, whenever $p, g, p+g \in [0,1]$.*

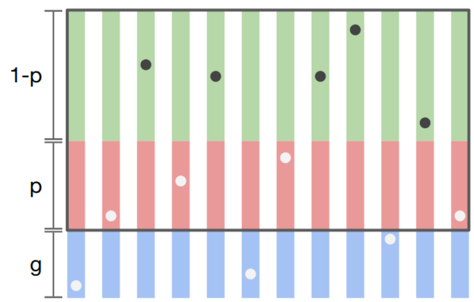

Figure 2: *We uniformly randomly sample points on N sticks, each with three parts:* ***green*** *with length $1-p$,* ***red*** *with length p, and* ***blue*** *with length g. We call non-blue points* ***boxed*** *and non-green points* ***hollow****.*

Proof. Let our coin flips arise from sampling points on sticks (Figure 2), where green is tails and red is heads. We'll show that probably less than $(p+g)N = p'N$ flips are heads. That is — given that all points are **boxed** — probably less than $p'N$ points are red. For any $M \geq p'N$:

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

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

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

We finish using difference of squares and the convexity of exp. □

The Chernoff bound gives us the control over tails we'd expect from the Central Limit Theorem, but for finite instead of asymptotically large N. In particular, when we learn from much but finite data, the in-sample error will concentrate near the out-of-sample error.

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:

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

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

$$\sqrt{(\log(1/\delta) + 32P)/N}$$

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 classifier, 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

Though $\mathcal{H}$ may be infinite, the restriction $\mathcal{H}_S = \{f|_S : f \in \mathcal{H}\}$ is finite for finite S. If we train and test on finitely many points total, we may treat $\mathcal{H}$ as finite. Thus, let us estimate $\text{tst}(f)$, which is an expectation over all of $\mathcal{D}$, by $\text{trn}_{\check{\mathcal{S}}}(f)$, an expectation over fresh samples $\check{\mathcal{S}} \sim \mathcal{D}^N$ independent from the samples $\mathcal{S}$ on which we learn.

To show that $\text{trn}_{\mathcal{S}} + g \leq \text{tst}$ when evaluated at $\mathcal{L}(\mathcal{S})$, we simply show that $\text{trn}_{\mathcal{S}} + g/2 \leq \text{trn}_{\check{\mathcal{S}}}$ and that $\text{tst} \leq \text{trn}_{\check{\mathcal{S}}} + g/2$. The former usually holds, since $|\mathcal{H}_{\mathcal{S}\sqcup\check{\mathcal{S}}}|$ is finite; the latter usually holds, since $\mathcal{S}$ and $\check{\mathcal{S}}$ are independent. Quantifying with Chernoff, we find that $\text{gap}_{\mathcal{S}}(\mathcal{L})$ exceeds g with chance at most

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

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

Claim 1 (Sauer)**.** Clearly, $H(n) \leq 2^n$. In fact, this bound is never somewhat tight: depending on $\mathcal{H}$, it either is an equality or very loose!

Proof. 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. We will count these strings by translating them from the alphabet $\{+, -\}$ to the alphabet $\{\blacksquare, \square\}$. Intuitively, $\blacksquare$ 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 $\blacksquare$. Otherwise, we overwrite by $\square$.

$+---$		$\square---$		$\square\square--$		$\square\square\square-$		$\square\square\square\square$
$-+--$		$\square+--$		$\square\blacksquare--$		$\square\blacksquare\square-$		$\square\blacksquare\square\square$
$--+-$		$\square-+-$		$\square\square+-$		$\square\square\blacksquare-$		$\square\square\blacksquare\square$
$---+$	$\to$	$\square--+$	$\to$	$\square\square-+$	$\to$	$\square\square\square+$	$\to$	$\square\square\square\blacksquare$
$--++$		$\square-++$		$\square\square++$		$\square\square\blacksquare+$		$\square\square\blacksquare\blacksquare$
$-+++$		$\square+++$		$\square\blacksquare++$		$\square\blacksquare\square+$		$\square\blacksquare\square\blacksquare$
$++++$		$\blacksquare+++$		$\blacksquare\square++$		$\blacksquare\square\square+$		$\blacksquare\square\square\square$

Figure 3: Translating elements of H_S (left) to strings of choice points (right). Each row corresponds to one of 7 classifiers and each column corresponds to one of 4 data points. We color pairs of strings that differ in-and-only-in 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 $\blacksquare$s, $|\mathcal{H}_T| = 2^k$. This is because $\blacksquare$s mark choice points where the classifiers attain both $+$ and $-$. Now, **either** $H(n) = 2^n$ for all n, **or** there is a greatest k for which $H(k) = 2^k$. In the latter case, no translated string may have more than k $\blacksquare$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 have grown like 2^n grows only polynomially. $\square$

Intuitively, the exponent k is a dimension:

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

We conclude that $\text{gap}_{\mathcal{S}}(\mathcal{L})$ exceeds g with chance at most

$$(2N+2)^{\dim(\mathcal{H})} \cdot \exp(-Ng^2/16) \cdot 2 \qquad (1)$$

For sufficiently large but finite N, the gap is small, so generalizing from data is possible.

Statistical Learning Theory

The following theorem, whose "only if" direction we have sketched above, summarizes 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 $\text{gap}_{\mathcal{S}}(\mathcal{L})$ exceeds g tends to 0 as $N = |\mathcal{S}|$ grows.*

For example, if $\mathcal{X}$ is a d-dimensional real vector space, $\mathcal{X}^*$ is its dual, and

$$\mathcal{H} = \{\text{sign} \circ \theta : \theta \in \mathcal{X}^*\}$$

is the set of "linear classifiers", then $\mathcal{H}$'s V-C dimension is at most d, because any $d+1$ points $x_0, x_1, \cdots x_d \in \mathcal{X}$ must participate in a linear relation $\sum_{i \in I} c_i x_i = \sum_{j \in J} c_j x_j$ for some I, J disjoint and each c positive, so no $f \in \mathcal{H}$ classifies each x_i as positive and each x_j as negative. By bound 1, a learned linear classifier will generalize when $N \gg d\log(N)$.

Beyond the V-C theorem, **statistical learning theory** abounds with variations on the theme that $\text{gap}_{\mathcal{S}} \leq \sqrt{\log(|\mathcal{H}|/\delta)/N}$.

For instance, viewing $\log(|\mathcal{H}|)$ as the maximum entropy of $\mathcal{L}(\mathcal{S}) \in \mathcal{H}$, one may seek tighter bounds given information-theoretic data. Recent progress [6] uses the mutual information between the random variables $\mathcal{S}$ and $\mathcal{L}(\mathcal{S})$.

In another direction, absent control over $\mathcal{D}$, one may seek to estimate properties of $\mathcal{D}$ from $\mathcal{S}$. For instance, *margin bounds* detect when $\mathcal{D}$'s two classes are geometrically well-separated and hence generalization is probable [5].

Other work specifically analyzes deep neural networks (nets). The V-C bound is empirically very loose for nets. Indeed, though nets seem to have nearly exponential $H(n)$s for n comparable to modern dataset sizes [4], they achieve state-of-the-art out-of-sample errors on a variety of real-world tasks [3]. A large $H(n)$ means that nets are flexible enough to fit arbitrary data. This flexibility allows nets to model complex patterns yet — in a phenomenon invisible to V-C theory — seems not to hinder generalization. Thus, the mystery of modern machine learning: with deep neural networks, may we continually halve our cake — and eat it, too?

References

The use of three-segment sticks and $\{\blacksquare, \square\}$-encoding to present the V-C bound is, to the author's knowledge, new. That said, the constant factors throughout this note are suboptimal. The textbook [5] surveys learning theory.

[1] **V. N. Vapnik, A. Y. Chervonenkis**. On uniform convergence of the frequencies of events to their probabilities. *Theory of Probability and its Applications*, 1971.

[2] **J. E. Wetzel**. On the Division of the Plane by Lines. *The American Mathematical Monthly*, October 1978.

[3] **Y. LeCun, Y. Bengio, G. Hinton**. Deep Learning. *Nature*, 2015.

[4] **C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals**. Understanding Deep Learning Requires Rethinking Generalization. *International Conference on Learning Representations*, 2017.

[5] **M. Mohri, A. Rostamizadeh, A. Talwalkar**. Foundations of Machine Learning. *MIT Press*, 2018.

[6] **A. R. Asadi, E. Abbe, S. Verdú**. Chaining Mutual Information and Tightening Generalization Bounds. *Neural Information Processing Systems*, 2018.