WHAT IS...

the VC-Dimension?

Samuel C. Tenka

Wetzel's Cake Problem

Combinatorists and bakers alike know the sequence $1, 2, 4, 8, 16, \cdots$ by heart. It continues, of course, with 31, for its nth element $c(n)$ 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.

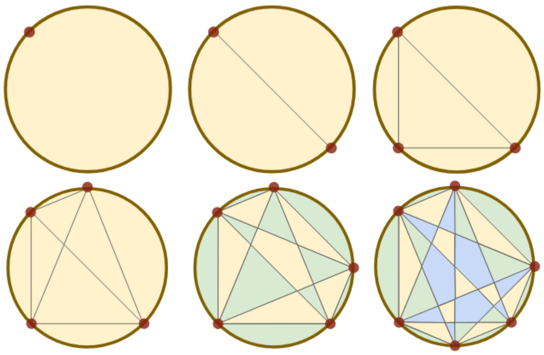

Figure 1: Cakes for $n = 1$ through 6. The $n = 4$ cake (bottom left) has $c(4) = 8$ pieces. Using colors to ease counting, we see that $c(6)$ is odd: the pieces besides the central yellow triangle partition into sets of six.

In fact, $c(n)$ is a polynomial [3]. We may compute $c(n)$ by regarding a sliced cake as a planar graph, noting that two cuts (and hence four boundary vertices) determine each interior vertex, and applying Euler's formula. One finds that $c(n)$ is $\binom{n-1}{0} + \cdots + \binom{n-1}{4}$, which explains why $c(n)$ initially coincides with 2^{n-1}.

Patterns do not always generalize. But then, how is learning from data possible at all? That is, if from a collection $\mathcal{H}$ of possible patterns we find some $f \in \mathcal{H}$ that matches or performs well on N observed data points, when should we expect that f matches unseen data? This question motivates machine learning's theory and guides its practice.

Learning and Generalization

Let us explicitly frame a simple case of the problem. Say that $\mathcal{X}$ is a space of images, $\{\pm 1\} = \{\mathrm{Cow}, \mathrm{Dog}\}$ is a set of labels, and we seek a classifier $f : \mathcal{X} \to \{\pm 1\}$ that accords with nature. More precisely, we model nature as a probability distribution $\mathcal{D}$ over the space $\{\pm 1\} \times \mathcal{X}$ of labeled images and we fix a set $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ of candidate classifiers. With $\mathcal{S} \sim \mathcal{D}^N$ denoting N samples drawn i.i.d. from $\mathcal{D}$, the **training error** of $f \in \mathcal{H}$ averages over the empirical distribution $\mathcal{S}$:

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

and the **testing error** averages over $\mathcal{D}$:

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

To learn well is to map an $\mathcal{S}$ to an f with low testing error. We call a map $\mathcal{L} : (\{\pm 1\} \times \mathcal{X})^N \to \mathcal{H}$ a **learning rule**. $\mathcal{L}$ might be defined by an approximate minimization of the training error over $\mathcal{H}$. Then the testing error decomposes into the failures of $\mathrm{trn}_{\mathcal{S}}$ to estimate tst (generalization), of $\mathcal{L}$ to minimize $\mathrm{trn}_{\mathcal{S}}$ (optimization), and of $\mathcal{H}$ to contain nature's truth (approximation):

$$\begin{aligned}\mathrm{tst}(\mathcal{L}(\mathcal{S})) = {} & \mathrm{tst}(\mathcal{L}(\mathcal{S})) && - \mathrm{trn}_{\mathcal{S}}(\mathcal{L}(\mathcal{S})) && \text{gener.}\\ & + \mathrm{trn}_{\mathcal{S}}(\mathcal{L}(\mathcal{S})) && - \inf_{\mathcal{H}}(\mathrm{trn}_{\mathcal{S}}(f)) && \text{optim.}\\ & + \inf_{\mathcal{H}}(\mathrm{trn}_{\mathcal{S}}(f)) && && \text{approx.}\end{aligned}$$

Focusing on the top term, we wonder when we may bound the **generalization gap**

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

When $\mathcal{L}(\mathcal{S})$ and $\mathcal{S}$ are independent (e.g. if $|\mathcal{H}| = 1$), $\mathrm{trn}_{\mathcal{S}}(\mathcal{L}(\mathcal{S}))$ is an unbiased estimate of $\mathrm{tst}(\mathcal{L}(\mathcal{S}))$, so the Weak Law controls $\mathrm{gap}_{\mathcal{S}}$. However, to reduce the approximation error we typically choose $|\mathcal{H}|$ large; can we bound the gap in this case, too? We will use the techniques of concentration and symmetrization to show that the answer is YES for "finite-dimensional" $\mathcal{H}$.

Concentration

Lemma (Chernoff)**.** *The fraction of heads among* N *i.i.d. flips of a biased coin exceeds its mean* p *by* g *with chance at most* $\exp(-Ng^2)$*, for* $0 \le p < p+g \le 1$*.*

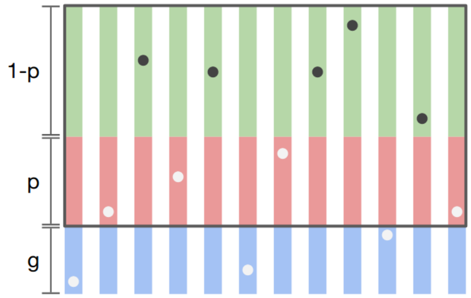

Figure 2: We sample points uniformly at random on N sticks, each with three parts: **green** with length $1-p$, **red** with length p, and **blue** with length g. We call non-blue points **boxed** and non-green points **hollow**.

Proof. Let our coin flips arise from sampling points on sticks (Figure 2), where green means tails, red means heads, and we condition on the event that blues do not occur. To show that less than $(p+g)N = p'N$ flips are heads is to show — given that all points are **boxed** — that less than $p'N$ points are red. For any M:

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

We sum over $M \ge p'N$, bound $\mathbb{P}[\cdots p'N \cdots] \le 1$, then invoke $(x \mapsto x^{p'})$'s concavity and exp's convexity:

$$\begin{aligned} &\mathbb{P}[\text{at least } p'N \text{ are red} \mid \text{all are boxed}] \\ &\le (1-g/p')^{p'N} \cdot (1+g)^N \cdot \mathbb{P}[\text{at least } p'N \text{ are hollow}] \\ &\le (1-g)^N \cdot (1+g)^N = (1-g^2)^N \le \exp(-Ng^2) \qquad \square \end{aligned}$$

The Chernoff bound gives us the control over tails we would 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 training error will **concentrate** near the testing error.

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

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

For example, if $\mathcal{H}$ is parameterized by P numbers, each represented on a computer by 32 bits, then $|\mathcal{H}| \le 2^{32P}$ and, with probability $1-\delta$, the gap is less than

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

This bound's sensitivity to the description length $32P$ may seem artificial. Indeed, the various $\mathcal{H}$ used in practice — e.g. linear models or neural networks — depend smoothly on their parameters, so the parameters' least significant bits barely affect the classifier. In other words, $\mathcal{H}$'s cardinality is not an apt measure of its size. The VC-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 $\mathrm{tst}(f)$, an expectation over all of $\mathcal{D}$, by $\mathrm{trn}_{\check{\mathcal{S}}}(f)$, an expectation over fresh samples $\check{\mathcal{S}} \sim \mathcal{D}^N$ independent of the $\mathcal{S}$ from which we learn.

To show that $\mathrm{trn}_{\mathcal{S}} + g \ge \mathrm{tst}$ when evaluated at $\mathcal{L}(\mathcal{S})$, we simply show that $\mathrm{trn}_{\mathcal{S}} + g/2 \ge \mathrm{trn}_{\check{\mathcal{S}}}$ and that $\mathrm{trn}_{\check{\mathcal{S}}} + g/2 \ge \mathrm{tst}$. 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, one finds that $\mathrm{gap}_{\mathcal{S}}(\mathcal{L})$ exceeds g (provided $g \ge 2/\sqrt{N}$) with chance at most

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

Thus we may control the gap by bounding $H(n) := \max_{|S|=n} |\mathcal{H}_S|$. We have achieved this reduction by putting the training and testing data on equal footing, hence the name **symmetrization** [8].

To progress, we bound $H(n)$. It is clear that $H(n) \le 2^n$. In fact, when this bound is loose, it is very loose:

Lemma (Sauer and Shelah)**.** *Unless* $H(n) = 2^n$ *for all* n, $H(n)$ *is bounded by a polynomial.*

Proof. Consider $\mathcal{H}_S$ for $|S| = n$. Ordering S, let us write each $f \in \mathcal{H}_S$ as a string of plus ($+$) and minus ($-$) signs. We will count these strings by translating them from the alphabet $\{+, -\}$ to the alphabet $\{■, □\}$. Intuitively, ■ represents "surprisingly $+$". More precisely, we work from left to right; whenever two (partially translated) strings differ **only** in their leftmost untranslated coordinate we overwrite the $+$ in that coordinate with ■. Otherwise, we overwrite with □. See Figure 3.

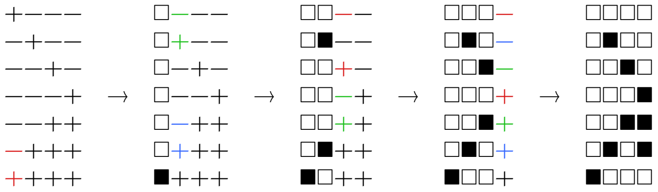

Figure 3: Translating elements of H_S (left) to strings of choice points (right). Each row corresponds to one of 7 classifiers and each column corresponds to one of 4 data points. We color pairs of strings that differ in-and-only-in their leftmost untranslated coordinate.

Each step of translation keeps distinct strings distinct. Moreover, whenever some k indices $T \subseteq S$ of a translated string are ■s, $|\mathcal{H}_T| = 2^k$. This is because ■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 ■s. So $\mathcal{H}_S$ contains no more strings than there are subsets in S of size $\leq k$. In turn, we may encode each nonempty subset as the image of a function with a size-k domain:

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

As with Wetzel's Cake, what might have grown like 2^n grows only polynomially. □

Since $\mathcal{H}_S$ grows no faster than the volume of a k-dimensional solid of diameter $|S|$, we call k a dimension.

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

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

$$\begin{aligned}&(\text{polynomial in } N \text{ of degree } \dim(\mathcal{H})) \;\cdot \\ &(\text{exponential decay in } Ng^2)\end{aligned}$$

When $N \gg \dim(\mathcal{H}) \log(N)$, the gap is probably small and generalizing from data is possible.

For example, if $\mathcal{X}$ is a d-dimensional real vector space and $\mathcal{H}$ is a set of "linear classifiers"

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

then $\dim(\mathcal{H})$ is at most d, because any $d+1$ points $x_0, x_1, \cdots x_d \in \mathcal{X}$ engage in a linear relation $\sum_{i \in I} c_i x_i = \sum_{j \in J} c_j x_j$ with each c positive, so no $f \in \mathcal{H}$ classifies every x_i as $+$ and every x_j as $-$. A learned linear classifier will hence generalize when $N \gg d \log(N)$.

Statistical Learning Theory

We've seen that finite VC-dimension implies that learning will generalize. A converse alse holds:

Theorem (Vapnik and Chervonenkis)**.** *The VC-dimension of* $\mathcal{H}$ *is finite if and only if for all 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 *grows.*

The VC theorem is but one result in **statistical learning theory**, which 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 improvements given information-theoretic hypotheses. Recent progress [6] uses the *mutual information* between $\mathcal{S}$ and $\mathcal{L}(\mathcal{S})$. Or, absent prior knowledge of $\mathcal{D}$, one may seek post hoc bounds by estimating properties of $\mathcal{D}$ from $\mathcal{S}$. For example, *margin bounds* detect when $\mathcal{D}$'s two classes are geometrically well-separated and hence generalization is probable [7].

Other work specifically analyzes deep neural networks (nets). The VC bound is empirically very loose for nets: though nets have nearly exponential $H(n)$ for typical data sets [5], they achieve state-of-the-art testing errors on many real-world tasks [4]. A large $H(n)$ allows nets to model complex patterns yet — in a phenomenon invisible to VC theory — seems not to hinder generalization. Thus, the mystery of modern learning: with deep neural networks, may we continually halve our cake and eat it, too?

References

We thank Arthur Ozga and Joe Trate for testing this exposition in a high school class. We believe the use of three-segment sticks and {■, □}-encoding to present the VC bound in elementary terms is new. For simplicity, we stated suboptimal constants and neglected to distinguish sampling with and without replacement.

The textbooks [7] and [8] both further explore concentration and symmetrization. As symmetrization enables restriction to finite data sets, we may study it via finite-dimensional geometry, whether through [7]'s *Rademacher complexities* or [8]'s *Gaussian widths.*

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

[2] N. Sauer. On the density of families of sets (Theorem 2). *J. Combinatorial Theory*, 1972.

[3] J. E. Wetzel. On the division of the plane by lines (§6). *The American Mathematical Monthly*, October 1978.

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

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017.

[6] A. R. Asadi, E. Abbe, S. Verdú. Chaining mutual information and tightening generalization bounds. *Neural Information Processing Systems*, 2018.

[7] M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of machine learning (§3.1, §5.4). *MIT Press*, 2018.

[8] R. Vershynin. High-dimensional probability (§8.3). *Cambridge University Press*, 2018.

Sam Tenka is a graduate student in computer science at MIT. Their email address is coli@mit.edu. This article's graphics are all provided courtesy of the author.