Vapnik-Chervonenkis Dimension

A collection $\mathcal{C}$ of subsets of a set X *shatters* a finite subset F if $\{F \cap C \,|\, C \in \mathcal{C}\} = \mathcal{P}(F)$, where $\mathcal{P}(F)$ is the set of all subsets of F. The collection $\mathcal{C}$ is a *VC-class* if there is some $n \in \mathbb{N}$ such that no set F containing n elements is shattered by $\mathcal{C}$, and the least such n is the VC-dimension, $\mathcal{V}(\mathcal{C})$, of $\mathcal{C}$.

Let $\mathcal{C} \cap F := \{C \cap F \mid C \in \mathcal{C}\}$ and for $n = 1, 2, \ldots$ let

$$f_{\mathcal{C}}(n) := \max\left\{|\mathcal{C} \cap F| \,\middle|\, F \subset X \text{ and } |F| = n\right\}.$$

Also, let $p_d(n) = \sum_{i<d} \binom{n}{i}$.

Theorem (Sauer). *Suppose that $f_{\mathcal{C}}(d) < 2^d$ for some d. Then $f_{\mathcal{C}}(n) \leq p_d(n)$ for all n.*

An $\mathcal{L}$-formula $\varphi(x_1, \ldots, x_k; y_1, \ldots, y_m)$ has the *independence property* with respect to the $\mathcal{L}$-structure $\mathfrak{R}$ if for every $n = 1, 2, \ldots$ there are $\bar{b}_1, \ldots, \bar{b}_n \in \mathbb{R}^m$ such that for every $X \subseteq \{1, \ldots, n\}$, there is some $\bar{a}_X \in \mathbb{R}^k$ satisfying

$$\varphi(\bar{a}_X; \bar{b}_i) \text{ is true in } \mathfrak{R} \iff i \in X.$$

If φ does not have the independence property with respect to $\mathfrak{R}$, we let $\mathcal{I}(\varphi)$ be the least n for which the property above fails.

For an $\mathcal{L}$-formula $\varphi(\bar{x}; \bar{y})$ and a structure $\mathfrak{R}$, let $S \subseteq \mathbb{R}^{k+m}$ be the set defined by φ. We let

$$\mathcal{C}_\varphi := \{S_{\bar{b}} \mid \bar{b} \in \mathbb{R}^m\}$$

denote the family of subsets of $\mathbb{R}^k$ determined by S.

Theorem (Laskowski). *The definable family $\mathcal{C}_\varphi$ is a VC-class if and only if φ does not have the independence property. Moreover, if $\mathcal{V}(\mathcal{C}_\varphi) = d$ and $\mathcal{I}(\varphi) = n$, then $n \leq 2^d$ and $d \leq 2^n$ (and these bounds are sharp).*

Let $\psi(\bar{y}; \bar{x}) := \varphi(\bar{x}; \bar{y})$ be the *dual formula* of φ. That is, ψ and φ are the same formula (and so define the same set) with the roles of $\bar{x}$ and $\bar{y}$ reversed.

The theorem follows from the next two lemmas.

Lemma 1. *With the notation as above, $\mathcal{V}(\mathcal{C}_\varphi) \geq d$ if and only if $\mathcal{I}(\psi) \geq d$.*

Lemma 2. *Let the notation be as above. Then $\mathcal{I}(\varphi) \leq n$ implies $\mathcal{I}(\psi) \leq 2^n$.*

We say that the $\mathcal{L}$-structure $\mathfrak{N}$ has the *independence property* if there is a formula $\varphi(x;\bar{y})$ *with just the single variable* x that has the independence property with respect to $\mathfrak{N}$.

Applying model theoretic methods, Laskowski gives a clear combinatorial proof of

Theorem (Shelah 1971). *An $\mathcal{L}$-structure $\mathfrak{N}$ has the independence property if and only if there is a formula $\varphi(\bar{x};\bar{y})$ (in any number of x variables) that has the independence property with respect to $\mathfrak{N}$.*

Again, using model-theoretic methods

Proposition (Pillay-CS 1986). *O-minimal structures do not have the independence property.*

Theorem (Laskowski '92). *Let* $\mathfrak{R} = (\mathbb{R}, <, \ldots)$ *be o-minimal and let* $S \subset \mathbb{R}^{k+m}$ *be definable. Then the collection* $\mathcal{C} = \{S_{\bar{x}} | \bar{x} \in \mathbb{R}^m\}$ *is a VC-class.*

Remark. Many structures are known not to have the independence property (by work of Shelah), and thus Laskowski's theorem provides significantly more examples of VC-classes.

To illustrate, the field of complex numbers, $(\mathbb{C}, +, \cdot)$ does not have the independence property, and thus any definable family of sets in this structure is a VC-class.

Probably Approximately Correct (PAC) Learning

Idea: Begin with an *instance space* X that is supposed to represent all instances (or objects) in a learner's world. A *concept* c is a subset of X, which we can identify with a function $c\colon X \to \{0,1\}$. A *concept class* $\mathcal{C}$ is a collection of concepts.

A *learning algorithm* for the concept class $\mathcal{C}$ is a function L which takes as input m-tuples $((x_1, c(x_1)), \ldots, (x_m, c(x_m))$ for $m = 1, 2, \ldots$ and outputs hypothesis concepts $h \in \mathcal{C}$ that are consistent with the input.

If X comes equipped with a probability distribution, then we can define the *error* of h to be $\operatorname{err}(h) = P(h \triangle c)$.

The learning algorithm L is said to be PAC if for every $\epsilon, \delta \in (0,1)$ there is $m_L(\epsilon, \delta)$ so that for *any* probability distribution P on X and any concept $c \in \mathcal{C}$, we have for all $m \geq m_{L(\epsilon,\delta)}$ that

$$P\big(\{\bar{x} \in X^m \mid \mathrm{err}(L((x_i, c(x_i))_{i \leq m}) \leq \epsilon\}\big) \geq 1 - \delta.$$

It can be shown that an algorithm that outputs a hypothesis concept h consistent with the sample data is PAC provided that $\mathcal{C}$ is a VC-class. Moreover, for given ϵ and δ, the number of sample points needed is, roughly speaking, proportional to the VC-dimension $\mathcal{V}(\mathcal{C})$.

Neural Networks

Macintyre-Sontag 1993 and Karpinski-Macintyre 1994 apply Laskowski's result and the uniform bounds available in o-minimal structures to answer questions about neural networks.

The output in a sigmoidal neural network is the result of computing a quantifier-free formula whose atomic formulas have the form $\tau(\bar{x}, \bar{w}) > 0$ or $\tau(\bar{x}, \bar{w}) = 0$, where τ is built from polynomials and exp, $\bar{x}$ are input values, and $\bar{w}$ represent a tuple of programmable parameters. Varying the parameters gives rise to a definable family in an o-minimal structure and hence Laskowski's theorem applies, which tells us that it is possible to PAC learn the architecture of such a network.

The first results of Macintyre and Sontag applied Laskowski's theorem to prove finite VC-dimension. Using quantitative results of Khovanskii, Karpinski and Macintyre give an upper bound for the VC-dimension that is $O(m^4)$, where m is the number of weights.

Koiran and Sontag 1997 have established a quadratic lower bound (in the number of weights) for the VC-dimension.

Some References

M. Anthony, Probabilistic 'Generalization of Functions and Dimension-based Uniform Convergence Results, *Statistics and Computing*, **8** (1998), 5–14. (available—with much more—on his website:
`www.maths.lse.ac.uk/Personal/martin/`.

L. van den Dries, *Tame Topology and O-minimal Structures* (London Mathematical Society Lecture Note Series, vol. 248), Cambridge: Cambridge University Press, 1998.

M. Karpinski and A. Macintyre, Polynomial Bounds for VC Dimension of Sigmoidal and General Pfaffian Neural Networks, *J. Computing and System Sciences*, **54** (1997), 169–176.

References (cont'd)

M. Kearns and U. Vazirani, *An Introduction to Computational Learning Theory*, Cambridge, MA and London: The MIT Press, 1994.

P. Koiran and E. Sontag, Neural Networks with Quadratic VC Dimension, *J.C.S.S.*, **54** (1997), 190–198.

M. C. Laskowski, Vapnik-Chervonenkis Classes of Definable Sets, *J. London Math. Soc.*, **245** (1992), 377–384.