AIT Blog

Samuel Epstein
samepst@jptheorygroup.org

September 28, 2022

This is a math blog focusing on Algorithmic Information Theory. The main focus will be on strings $x \in \{0,1\}^*$ that have low mutual information with the halting sequence H, with $\mathbf{I}(x;H) = \mathbf{K}(x) - \mathbf{K}(x|H)$, being low. $\mathbf{K}$ is the prefix free Kolmogorov complexity. There are many properties that can be proven about elementary objects that have low $\mathbf{I}(x;H)$. We say an object is (non)exotic if it has (low)high mutual information with the halting sequence. Exotic objects cannot be found in the physical world. Furthermore, $\mathbf{I}(x;H)$ enjoys conservation laws, in that deterministic and random processing cannot increase information.

- For partial computable f, $\mathbf{I}(f(a):H) < \mathbf{I}(a;H) + \mathbf{K}(f) + O(1)$.
- For program q that computes probability p over $\mathbb{N}$, $\mathbf{E}_{a\sim p}[2^{\mathbf{I}(\langle q,a\rangle;H)}] < O(1)2^{\mathbf{I}(q;H)}$.

This entry deals with the relationship between Algorithithmic Information Theory and Machine Learning. Classification is the task of learning a binary function c from $\mathbb{N}$ to bits $\{0,1\}$. The learner is given a sample consisting of pairs (x,b) for string x and bit b and outputs a binary classifier $h : \mathbb{N} \to \{0,1\}$ that should match c as much as possible. Occam's razor says that "the simplest explanation is usually the best one." Simple hypothesis are resilient against overfitting to the sample data. With certain probabilistic assumptions, learning algorithms that produce hypotheses of low Kolmogorov complexity are likely to correctly predict the target function [BEHW89]. The following theorem shows that the samples can be compressed to their count.

Theorem 1 *Given a set of samples $\{(x_i,b_i)\}$, $i = 1,\ldots,n$, there is a total function $f : \mathbb{N} \to \{0,1\}$ such that $f(x_i) = b_i$ for $i = 1,\ldots,n$ and $\mathbf{K}(f) <^{\log} n + \mathbf{I}(\{(x_i,b_i)\};H)$.*

However, usually the samples can be modeled as coming from a probabilistic model. The target concept is modeled by a random variable X with distribution p over ordered lists of natural numbers. The random variable Y models the labels, and has a distribution over lists of bits, where the distribution of $X \times Y$ is $p(x,y)$ with conditional probability requirement $p(y|x) = \prod_{i=1..|x|} p(y_i|x_i)$. Each such (x_i,y_i) is a labeled sample. A binary classifier f is consistent with labelled samples (x,y), if for all i, $f(x_i) = y_i$. Let $\Gamma(x,y)$ be the minimum Kolmogorov complexity of a classifier consistent with (x,y). *Entropy*$(Y|X)$ is the conditional entropy of Y given X.

Theorem 2 *Entropy*$(Y|X) \leq \mathbf{E}[\Gamma(X,Y)] <^{\log}$ *Entropy*$(Y|X) + \mathbf{K}(p)$.

Another area of machine learning is regression, in which one is give a set of pairs $\{(x_i,y_i)\}$, $i = 1\ldots n$, and the goal is to find a function f, such that $f(x_i) = y_i$. Usually each x_i and y_i represents a point in Euclidean space, but for our purpose they are natural numbers. As in classification, the goal is to use Occam's razor to find the simpliest function, to prevent overfitting to the random noise inherent in the sample data. The following theorem provides bounds on the simplest total computable function completely consistent with the data.

Theorem 3 *For* $\{(x_i, y_i)\}$, $i = 1, \ldots, n$, *there exists* $f : \mathbb{N} \to \mathbb{N}$ *with* $f(x_i) = y_i$ *for* $i \in \{1, \ldots, n\}$ *and* $\mathbf{K}(f) <^{\log} \sum_{i=1}^{n} \mathbf{K}(y_i|x_i) + \mathbf{I}(\{(x_i, y_i)\}; H)$.

This theorem can be proved using Theorem 8 in [Eps22]. However, this theorem is over computable probability measures, whereas the lower semi-computable $\mathbf{m}$ is needed. By using so-called left-total machines, $\mathbf{m}$ can be converted into a computable measure. In fact one of the benefits of using left-total machines and having $\mathbf{I}(,; H)$ as an error term, is that semi-computable functions can be converted into computable ones.

References

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[Eps22] S. Epstein. The outlier theorem revisited. *CoRR*, abs/2203.08733, 2022.