Regression and Algorithmic Information Theory

Samuel Epstein*

April 13, 2023

Abstract

In this paper we prove a theorem about regression, in that the shortest description of a function consistent with a finite sample of data is less than the combined conditional Kolmogorov complexities over the data in the sample.

1 Introduction

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 [Eps21] shows that the samples can be compressed to their count.

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

The $\mathbf{K}$ term is Kolmogorov complexity and the $\mathbf{I}$ term is defined in Section 2. Another area of machine learning is regression, in which one is given a set of pairs $\{(x_i,y_i)\}$, $i=1\dots n$, 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 purposes 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. This paper presents the following bounds on the simplest total computable function completely consistent with the data.

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

2 Conventions

For positive real functions f, by $<^+f$, $>^+f$, $=^+f$, and $<^{\log}f$, $>^{\log}f$, $\sim f$ we denote $\leq f+O(1)$, $\geq f-O(1)$, $=f\pm O(1)$ and $\leq f+O(\log(f+1))$, $\geq f-O(\log(f+1))$, $=f\pm O(\log(f+1))$. $\mathbf{K}(x|y)$ is the conditional prefix Kolmogorov complexity. The chain rule states $\mathbf{K}(x,y)=^+\mathbf{K}(x)+\mathbf{K}(y|\mathbf{K}(x),x)$.

*JP Theory Group. samepst@jptheorygroup.org

$[A] = 1$ if the mathematical statement A is true, otherwise it is 0. Let $\mathbf{K}_t(x|y) = \inf\{\|p\| : U_y(p) = x \text{ in } t \text{ steps}\}$. The information the halting sequence $\mathcal{H}$ has about x is $\mathbf{I}(x;\mathcal{H}|y) = \mathbf{K}(x|y) - \mathbf{K}(x|y,\mathcal{H})$. $\mathbf{I}(x;\mathcal{H}) = \mathbf{I}(x;\mathcal{H}|\emptyset)$. A probability measure is elementary if its support is finite and it has rational values. The deficiency of randomness of $x \in \{0,1\}^*$ with respect to elementary probability measure Q is $\mathbf{d}(X|Q) = \lceil -\log Q(X) - \mathbf{K}(x|\langle Q\rangle)\rceil$. The stochasticity of x is $\mathbf{Ks}(x) = \min_Q \mathbf{K}(Q) + 3\log\max\{\mathbf{d}(X|Q),1\}$.

Lemma 1 ([Eps21, Lev16]) $\mathbf{Ks}(x) <^{\log} \mathbf{I}(x;\mathcal{H})$.

Lemma 2 ([Eps22]) *For partial computable f, $\mathbf{I}(f(x):\mathcal{H}) <^+ \mathbf{I}(x;\mathcal{H}) + \mathbf{K}(f)$.*

3 Results

Let $\Omega = \sum\{2^{-\|p\|} : U(p) \text{ halts}\}$ be Chaitin's Omega, $\Omega_n \in \mathbb{Q}_{\geq 0}$ be be the rational formed from the first n bits of Ω, and $\Omega^t = \sum\{2^{-\|p\|} : U(p) \text{ halts in time } t\}$. For $n \in \mathbb{N}$, let $\mathbf{bb}(n) = \min\{t : \Omega_n < \Omega^t\}$. $\mathbf{bb}^{-1}(m) = \arg\min_n\{\mathbf{bb}(n-1) < m \leq \mathbf{bb}(n)\}$. Let $\Omega[n] \in \{0,1\}^*$ be the first n bits of Ω.

Lemma 3 *For $n = \mathbf{bb}^{-1}(m)$, $\mathbf{K}(\Omega[n]|m,n) = O(1)$.*

Proof. For a string x, let $BB(x) = \inf\{t : \Omega^t > 0.x\}$. Enumerate strings of length n, starting with 0^n, and return the first string x such that $BB(x) \geq m$. This string x is equal to $\Omega[n]$, otherwise let y be the largest common prefix of x and $\Omega[n]$. Thus $BB(y) = \mathbf{bb}(\|y\|) \geq BB(x) \geq m$, which means $\mathbf{bb}^{-1}(m) \leq \|y\| < n$, causing a contradiction. □

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

Proof. Let $S = \{(x_i,y_i)\}$. Let $K = \sum_{i=1}^n \mathbf{K}(y_i|x_i)$. We have $T = \arg\min_t \sum_{i=1}^n \mathbf{K}_t(y_i|x_i) = K$. Let $N = \mathbf{bb}^{-1}(T)$ and $M = \mathbf{bb}(N)$ and we define $m(x|y) = 2^{-\mathbf{K}_M(x|y)}$, setting $m(\emptyset|y) = 1 - m(\mathbb{N}|y)$.

We condition all terms on M and K, and later in the proof, we'll make this condition explicit. Let Q be an elementary probability that realizes the stochasticity of S, where $d = \max\{\mathbf{d}(S|Q),1\}$. Without loss of generality, we can assume the support of Q consists entirely of samples $R = \{(x_j,y_j)\}_{j=1}^{n_R}$ (of potentially different sizes) such that $\prod_{j=1}^{n_R} m(y_j|x_j) = 2^{-M}$. Let $z = \arg\max_y(x,y) \in R \in \text{Support}(Q)$. We define a probability measure κ over $d2^M$ lists $\mathcal{L}$ of size z over $\mathbb{N}$, where each ℓ is chosen independently, and for each $\ell \in \mathcal{L}$, $\ell(i)$ is chosen independently according to $m(\cdot|i)$. We say a sample $R = \{(x_j,y_j)\}$ is inconsistent with a list ℓ, $R \ltimes \ell$, if there exists j, where $\ell(x_j) \neq y_j$. $\eta(R,\mathcal{L}) = [\forall \ell \in \mathcal{L}, E \ltimes \ell]$.

$$\mathbf{E}_{\mathcal{L}\sim\kappa}\mathbf{E}_{R\sim Q}[\eta(R,\mathcal{L})] = \mathbf{E}_{R\sim Q}\mathbf{E}_{\mathcal{L}\sim\kappa}\left(1 - \prod_j m(y_j|x_j)\right)^{d2^M} < \mathbf{E}_{R\sim Q}e^{-d} = e^{-d}.$$

Thus there exists a set of $d2^M$ lists $\mathcal{L}$, where $\mathbf{E}_{R\sim Q}[\eta(R,\mathcal{L})] < e^{-d}$. Thus let $t(R) = \eta(R,\mathcal{L})e^d$ be a Q-test, where $\mathbf{E}_{R\sim Q}[t(R)] \leq 1$. It must be $t(S) = 0$, otherwise, up to $O(1)$ constants we have

$$1.44d \leq \log t(S) \leq \mathbf{d}(S|Q) \leq d,$$

which is contradiction for large d, which we can assume without loss of generality. So there exists a list ℓ such that $\ell(x_i) = xy_i$, for all $(x_i,y_i) \in S$. Thus one can construct a total computable

function $f : \mathbb{N} \to \mathbb{N}$ from ℓ that is consistent with S, for example $f(x) = \ell(x)$ if $x \leq z$ and $f(x) = 0$ otherwise. Making the condition term M explicit and keeping the condition term K implicit we have,

$$\begin{aligned}\mathbf{K}(f|M) &<^+ \mathbf{K}(\ell|M)\\ &<^+ \log|\mathcal{L}| + \mathbf{K}(\mathcal{L}|M)\\ &<^+ K + \log d + \mathbf{K}(Q,d|M)\\ &<^+ K + \mathbf{Ks}(S|M).\end{aligned}$$

Using Lemma 1, we get, noting $M = \mathbf{bb}(N)$.

$$\begin{aligned}\mathbf{K}(f|M) &<^{\log} K + \mathbf{I}(S;\mathcal{H} \mid M).\\ \mathbf{K}(f) &<^{\log} K + \mathbf{K}(S|M) + \mathbf{K}(M) - \mathbf{K}(S|\mathcal{H}) + \mathbf{K}(N).\end{aligned}$$

By noting the chain rule, K is conditional, and $\mathbf{bb}(N) = M$, we have

$$\begin{aligned}&\mathbf{K}(S|M) + \mathbf{K}(M)\\ <^+&\mathbf{K}(S|M,\mathbf{K}(M)) + \mathbf{K}(\mathbf{K}(M)|M) + \mathbf{K}(M)\\ <&\mathbf{K}(S,M) + O(\log N)\\ <&\mathbf{K}(S) + O(\log N).\\ \mathbf{K}(f) &<^{\log} K + \mathbf{K}(S) - \mathbf{K}(S|\mathcal{H}) + O(\log N). \qquad (1)\end{aligned}$$

From K, and S, one can compute T, where $\mathbf{bb}^{-1}(T) = N$. Therefore by Lemma 3, $\mathbf{K}(\Omega[N]|S) <^+ \mathbf{K}(N)$, so by Lemma 2,

$$N <^{\log} \mathbf{I}(\Omega[N];\mathcal{H}) <^{\log} \mathbf{I}(S;\mathcal{H}) + \mathbf{K}(N) <^{\log} \mathbf{I}(S;\mathcal{H}). \qquad (2)$$

The above equation used the common fact that the first n bits of Ω had $n - O(\log n)$ bits of mutual information with $\mathcal{H}$. So combining Equations 1 and 2, we get

$$\mathbf{K}(f) <^{\log} K + \mathbf{I}(S;\mathcal{H}).$$

The proof is completed by noting the log precision, and the K term in the equation removes the implicit conditioning of K. □

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.

[Eps21] Samuel Epstein. All sampling methods produce outliers. *IEEE Transactions on Information Theory*, 67(11):7568–7578, 2021.

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

[Lev16] L. A. Levin. Occam bound on lowest complexity of elements. *Annals of Pure and Applied Logic*, 167(10), 2016.