arXiv:2304.07825v1 [cs.CC] 16 Apr 2023

回归与算法信息论

Samuel Epstein*

2023年4月18日

摘要

本文证明了一个关于回归的定理：与有限样本数据相一致的函数的最短描述，短于该样本中各数据的条件柯尔莫哥洛夫复杂度之和。

1 引言

分类是学习一个二元函数的任务 c 从 $\mathbb{N}$ 到比特 $\{0,1\}$。学习者会得到一个由以下配对组成的样本（x，b）：字符串 x 和比特 b，并输出一个二元分类器 h：$\mathbb{N}\to\{0,1\}$它应尽可能匹配 c。奥卡姆剃刀认为："最简单的解释通常是最好的。"简单的假设能够抵御对样本数据的过拟合。在某些概率假设下，产生低柯尔莫哥洛夫复杂度假设的学习算法很可能正确预测目标函数 [BEHW89]。以下定理[Eps21] 表明，样本可以压缩为其数量。

定理。给定一组样本 $\{(x_i,b_i)\}_{i=1}^n$，存在一个函数 f：$\mathbb{N}\to\{0,1\}$ 使得$f(x_i)=b_i$，对于 $i=1,\dots,n$，且 $\mathrm{K}(f) <^{\log} n+\mathrm{I}(\{(x_i,b_i)\};\mathcal{H})$。

该**K** 项为柯尔莫哥洛夫复杂度，而 **I** 项的定义见第 2节。机器学习的另一个领域是回归，其中给定一组数对 $\{(x_i,y_i)\}$，$i=1\dots n$，目标是找到一个函数 f，使得 $f(x_i)=y_i$。通常，每个 x_i 和 y_i都表示欧几里得空间中的一个点，但就本文而言，它们是自然数。与分类一样，目标是运用奥卡姆剃刀找到最简单的函数，以防止对样本数据中固有的随机噪声过拟合。本文给出了与数据完全一致的最简单全可计算函数的以下界限。

定理。对于 $\{(x_i,y_i)\}_{i=1}^n$，存在 f：$\mathbb{N}\to\mathbb{N}$，使得 $f(x_i)=y_i$对于 $i\in\{1,\dots,n\}$，且 $\mathrm{K}(f) <^{\log}\sum_{i=1}^n \mathrm{K}(y_i|x_i)+\mathrm{I}(\{(x_i,y_i)\};\mathcal{H})$.

2 约定

对于正实函数 f，用 $<^+f$，$>^+f$，$=^+f$ 和 $<^{\log}f$，$>^{\log}f$，$\sim f$ 分别表示 $\le f+O(1)$、$\ge f-O(1)$、$=f\pm O(1)$ 以及 $\le f+O(\log(f+1))$、$\ge f-O(\log(f+1))$、$=f\pm O(\log(f+1))$。$\mathrm{K}(x|y)$ 表示条件前缀柯尔莫哥洛夫复杂度。链式法则指出，$\mathrm{K}(x,y)=^+\mathrm{K}(x)+\mathrm{K}(y|\mathrm{K}(x)\ x)$。

*JP 理论组。samepst@jptheorygroup.org

若数学命题 A 为真，则令 $[A]=1$；否则令 $[A]=0$。令 $\mathrm{K}_t(x|y)=\inf\{\|p\|:U_y(p)=x$ 在 t 步内停机$\}$。停机序列 $\mathcal{H}$ 关于 x 所含的信息为 $\mathrm{I}(x;\mathcal{H}|y)=\mathrm{K}(x|y)-\mathrm{K}(x|y,\mathcal{H})$。$\mathrm{I}(x;\mathcal{H})=\mathrm{I}(x;\mathcal{H}|\emptyset)$。如果一个概率测度的支撑集有限且其取值为有理数，则称该概率测度为初等概率测度。

$x\in\{0,1\}^*$ 相对于初等概率测度 Q 的随机性缺陷为 $\mathrm{d}(X|Q)=\lceil-\log Q(X)-\mathrm{K}(x|\langle Q\rangle)\rceil$。$x$ 的随机性为 $\mathrm{Ks}(x)=\min_Q\mathrm{K}(Q)+3\log\max\{\mathrm{d}(X|Q),1\}$。

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

引理 2 ([Eps22]) 对于部分可计算的 $f,\mathbf{I}(f(x):\mathcal{H})<^+\mathbf{I}(x;\mathcal{H})+\mathbf{K}(f\quad)$.

3 结果

令 $\Omega=\sum\{2^{-\|p\|}:U(p)$ 停机$\}$ 为查汀常数，$\Omega_n\in\mathbb{Q}_{\geq 0}$ 为由 Ω 的前 n 位构成的有理数，并且 $\Omega^t=\sum\{2^{-\|p\|}:U(p)$ 在时间 $t\}$ 内停机。对于 $n\in\mathbb{N}$，令 $\mathrm{bb}(n)=\min\{t:\Omega_n<\Omega^t\}$。$\mathrm{bb}^{-1}(m)=\arg\min_n\{\mathrm{bb}(n-1)<m\leq\mathrm{bb}(n)\}$。令 $\Omega[n]\in\{0,1\}^*$ 为 Ω 的前 n 位。

引理 3 对于 $n=\mathrm{bb}^{-1}(m)$，$\mathrm{K}(\Omega[n]|m,n)=O(1)$。

证明。对于字符串 x，令 $BB(x)=\inf\{t:\Omega^t>0.x\}$。枚举长度为 n 的字符串，从 0^n 开始，并返回第一个满足 $BB(x)\geq m$ 的字符串 x。该字符串 x 等于 $\Omega[n]$；否则，令 y 为 x 与 $\Omega[n]$ 的最长公共前缀。因此 $BB(y)=\mathrm{bb}(\|y\|)\geq BB(x)\geq m$，这意味着 $\mathrm{bb}^{-1}(m)\leq\|y\|<n$，从而导致矛盾。□

定理 1 对于 $\{(x_i,y_i)\}_{i=1}^n$，存在 $f\colon\mathbb{N}\to\mathbb{N}$，使得对于 $i\in\{1,\ldots,n\}$，$f(x_i)=y_i$，且 $\mathrm{K}(f)<^{\log}\sum_{i=1}^n\mathrm{K}(y_i|x_i)+\mathrm{I}(\{(x_i,y_i)\};\mathcal{H})$。

证明。令 $S=\{(x_i,y_i)\}$。令 $K=\sum_{i=1}^n\mathrm{K}(y_i|x_i)$。我们有 $T=\arg\min_t\sum_{i=1}^n\mathrm{K}_t(y_i|x_i)=K$。令 $N=\mathrm{bb}^{-1}(T)$ 和 $M=\mathrm{bb}(N)$，并定义 $m(x|y)=2^{-\mathbf{K}_M(x|y)}$，令 $m(\emptyset|y)$ 为 $=1-m(\mathbb{N}|y)$。

我们对 M and K 条件化，并将在证明的后文中明确说明这一条件。设 Q 是一个初等概率，它刻画了 S 的随机性，其中 $d=\max\{\mathrm{d}(S|Q),1\}$。不失一般性，我们可以假设 Q 的支撑集完全由样本 $R=\{(x_j,y_j)\}_{j=1}^{n_R}$（大小可能各不相同）构成，且满足 $\prod_{j=1}^{n_R}m(y_j|x_j)=2^{-M}$。令

$$z=\max\{x:(x,y)\in R\in\mathrm{Support}(Q)\}.$$

我们定义一个概率测度 κ，其定义在 $d2^K$ 个大小为 z、取自 $\mathbb{N}$ 的列表 $\mathcal{L}$ 上，其中每个 $\ell\in\mathcal{L}$ 都是独立选取的；对于每个 $\ell\in\mathcal{L}$，$\ell(i)$ 都按照 $m(\cdot|i)$ 独立选取。如果存在 j，使得 $\ell(x_j)\neq y_j$，我们就称样本 $R=\{(x_j,y_j)\}$ 与列表 ℓ，$R\ltimes\ell$ 不一致。$\eta(R,\mathcal{L})=[\forall\ell\in\mathcal{L},E\ltimes\ell]$。

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

因此，存在一个列表 $\mathcal{L}$，使得 $\mathrm{E}_{R\sim Q}[\eta(R,\mathcal{L})]<e^{-d}$。于是，令 $t(R)=\eta(R,\mathcal{L})e^d$ 为一个 Q-检验，其中 $\mathrm{E}_{R\sim Q}[t(R)]\leq 1$。它必定是 $t(S)=0$，否则我们就有

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

当 d 足够大时，这就产生了矛盾，而不失一般性，我们可以假设这一点。因此，存在一个列表 ℓ，使得对所有 $(x_i, y_i) \in S$，都有 $\ell(x_i) = y_i$。于是，可以构造一个全可计算函数$f: \mathbb{N} \to \mathbb{N}$，它从 ℓ 出发且与 S 一致，例如，当 $x \leq z$ 时，$f(x) = \ell(x)$；否则，$f(x) = 1$。将条件项 M 显式写出，并将条件项 K 隐含起来，我们得到：

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

利用引理1，注意到 $M = \mathbf{bb}(N)$，且$\mathbf{bb}$ 可相对于 $\mathcal{H}$计算，

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

因此我们有：

$$\begin{aligned}&\mathrm{K}(S|M) + \mathrm{K}(M)\\ <^+&\mathrm{K}(S|M, \mathrm{K}(M)) + \mathrm{K}(\mathrm{K}(M)|M) + \mathrm{K}(M)\\ <^+&\mathbf{K}(S, M) + \mathbf{K}(\mathbf{K}(M)|M) &(1)\\ <^+&\mathbf{K}(S, N, M) + O(\log N) &(2)\\ <^+&\mathbf{K}(S, N) + O(\log N). &(3)\\ <^+&\mathbf{K}(S) + O(\log N).\\ \mathbf{K}(f) <^{\log}&K + \mathbf{K}(S) - \mathbf{K}(S|\mathcal{H}) + O(\log N). &(4)\end{aligned}$$

等式1 来自链式法则。等式2 源于以下事实：$M = \mathbf{bb}(N)$. 等式3 由此得到$\mathbf{K}(T|S, K) = O(1)$ 和引理3，这意味着$\mathbf{K}(M|N, T) <^+ \mathbf{K}(\Omega[N]|N, T) <^+ O(1)$。

根据 K和 S，可以计算出 T，其中$\mathbf{bb}^{-1}(T) = N$。因此根据引理3，$\mathbf{K}(\Omega[N]|S) <^+ \mathbf{K}(N)$，所以根据引理2，

$$N <^{\log} \mathbf{I}(\Omega[N]; \mathcal{H}) <^{\log} \mathbf{I}(S; \mathcal{H}) + \mathbf{K}(N) <^{\log} \mathbf{I}(S; \mathcal{H}). \tag{5}$$

上述方程用到了一个常见事实：Ω 的前 n 位 $n - O(\text{对数 } n)$ 位的信息与 $\mathcal{H}$ 互信息。结合方程 4 和 5，得到

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

注意到对数精度即可完成证明，而方程中的 K 项消除了 K的隐式条件化。□

参考文献

[BEHW89] A. Blumer、A. Ehrenfeucht、D. Haussler 和 M. K. Warmuth。《可学习性与 Vapnik-Chervonenkis 维数》。《ACM 学报》，36(4):929–965，1989年。

[Eps21] Samuel Epstein。《所有抽样方法都会产生离群值》。《IEEE 信息论汇刊》，67(11):7568–7578，2021年。

[Eps22] S. Epstein. 重新审视离群值定理。CoRR，abs/2203.08733，2022 年。[Lev16] L. A. Levin. 元素最低复杂度的奥卡姆界。《纯粹与应用逻辑年鉴》，167(10)，2016 年。