Sequence Prediction based on Monotone Complexity

Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

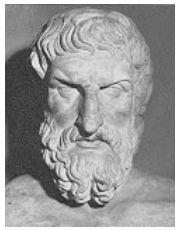
2003

Abstract

We study sequence prediction based on the monotone Kolmogorov complexity $Km = -\log m$, i.e. based on universal deterministic/one-part MDL. m is extremely close to Solomonoff's prior M, the latter being an excellent predictor in deterministic as well as probabilistic environments, where performance is measured in terms of convergence of posteriors or losses. Despite this closeness to M, it is difficult to assess the prediction quality of m, since little is known about the closeness of their posteriors, which are the important quantities for prediction. We show that for deterministic computable environments, the "posterior" and losses of m converge, but rapid convergence could only be shown on-sequence; the off-sequence behavior is unclear. In probabilistic environments, neither the posterior nor the losses converge, in general.

Table of Contents

Induction = Predicting the Future



Epicurus' principle of multiple explanations

If more than one theory is consistent with the observations, keep all theories.

Ockhams' razor (simplicity) principle

Entities should not be multiplied beyond necessity.

Hume's negation of Induction

The only form of induction possible is deduction as the conclusion is already logically contained in the start configuration.

Bayes' rule for conditional probabilities

Given the prior believe/probability one can predict all future probabilities.

Solomonoff's universal prior

Solves the question of how to choose the prior if nothing is known.

Strings and Conditional Probabilities

Strings: $x = x_1 x_2 ... x_n$ with $x_t \in \mathcal{X}$ and $x_{1:n} := x_1 x_2 ... x_{n-1} x_n$ and $x_{<n} := x_1 ... x_{n-1}$ and $\omega = x_{1:\infty}$.

Probabilities: $\rho(x_1 ... x_n)$ is the probability that an (infinite) sequence starts with $x_1 ... x_n$.

(Semi)Measures: $\rho : \mathcal{X}^* \to [0,1]$ and $\sum_{x_n \in \mathcal{X}} \rho(x_{1:n}) \stackrel{(<)}{=} \rho(x_{<n})$ and $\rho(\varepsilon) \stackrel{(<)}{=} 1$.

Deterministic environment: $\exists \omega : \rho(\omega_{1:n}) = 1 \ \forall n$. In this case we identify ρ with ω.

Conditional probability: $\rho(x_t | x_{<t}) = \rho(x_{1:t}) / \rho(x_{<t})$ is the ρ-probability that a given string $x_1 ... x_{t-1}$ is followed by (continued with) x_t.

Posterior Convergence

Assume that μ is the "true" (objective, aleatory) sequence generating probability measure, also called environment.

Usually we do not know μ, but estimate it from $x_{<t}$. Let $\rho(x_t|x_{<t})$ be an estimated (subjective, belief, epistemic) probability of x_t, given $x_{<t}$.

It is reasonable to aim for posterior convergence (consistency, self-tuningness): $\rho(x'_t|x_{<t}) \stackrel{(fast)}{\longrightarrow} \mu(x'_t|x_{<t})$ for $t \to \infty$

A sequence of random variable $z_t = z_t(\omega)$ (like $z_t = \rho(x_t|x_{<t}) - \mu(x_t|x_{<t})$ is said to converge for $t \to \infty$ to 0

$i)$ with probability 1 **(w.p.1)** $:\Leftrightarrow \mathbf{P}[\{\omega : z_t(\omega) \to 0\}] = 1$,
$ii)$ in mean sum **(i.m.s.)** $:\Leftrightarrow \sum_{t=1}^{\infty} \mathbf{E}[z_t^2] \leq c < \infty$,

Conv. i.m.s. implies conv. w.p.1 (rapid if c is of reasonable size).

Disadvantage: Neglects value/severity of correct/wrong predictions.

Self-optimizing Predictors

Let $\ell_{x_t y_t} \in [0,1]$ be the received loss when performing action/decision/prediction $y_t \in \mathcal{Y}$ and $x_t \in \mathcal{X}$ is the t^{th} symbol of the sequence, for instance

Loss ℓ_{xy}	$\mathcal{X}=\{$sunny , rainy$\}$	
$\mathcal{Y}=\{$ umbrella	0.3	0.1
$\mathcal{Y}=\{$ sunglasses $\}$	0.0	1.0

The goal is to minimize the μ-expected loss. More generally we define the Λ_ρ prediction scheme which minimizes the ρ-expected loss:

$$y_t^{\Lambda_\rho} := \arg\min_{y_t \in \mathcal{Y}} \sum_{x_t} \rho(x_t|x_{<t}) \ell_{x_t y_t}$$

The actual μ-expected loss when Λ_ρ predicts the t^{th} symbol is

$$l_t^{\Lambda_\rho}(x_{<t}) := \sum_{x_t} \mu(x_t|x_{<t}) \ell_{x_t y_t^{\Lambda_\rho}}$$

The decision theoretic counterpart of conv. is self-optimizingness

$$l_t^{\Lambda_\rho}(x_{<t}) \overset{(fast)}{\longrightarrow} l_t^{\Lambda_\mu}(x_{<t}) \quad \text{for} \quad t \to \infty$$

Predictive Properties of Universal Prior

If U is a universal prefix Turing machine, then Solomonoff's prior

$$M(x) := \sum_{p:U(p)=x*} 2^{-l(p)}, \qquad KM(x) := -\log M(x).$$

has (excellent) predictive properties (Solomonoff:78, Hutter:01):

Solomonoff's prior M is a (i) universal, (v) enumerable, (ii) monotone, (iii) semimeasure, which (vi) converges to μ i.m.s., and (vii) is self-optimizing i.m.s. More quantitatively:

$$(vi)\ \sum_{t=1}^{\infty} \mathbf{E}[\sum_{x'_t}(M(x'_t|x_{<t}) - \mu(x'_t|x_{<t}))^2] \stackrel{+}{\leq} \ln 2 \cdot K(\mu),$$
$$M(x'_t|x_{<t}) \stackrel{t\to\infty}{\longrightarrow} \mu(x'_t|x_{<t}) \text{ i.m.s. for } \mu \in \mathcal{M}^{msr}_{comp}.$$

$$(vii)\ \sum_{t=1}^{\infty} \mathbf{E}[(l_t^{\Lambda_M} - l_t^{\Lambda_\mu})^2] \stackrel{+}{\leq} 2\ln 2 \cdot K(\mu), \text{ which implies}$$
$$l_t^{\Lambda_M} \stackrel{t\to\infty}{\longrightarrow} l_t^{\Lambda_\mu} \text{ i.m.s. for } \mu \in \mathcal{M}^{msr}_{comp},$$

where $K(\mu)$ is the length of the shortest prg computing function μ.

Monotone Kolmogorov Complexity

- MDL approximation of $M(x) = \sum_{p:U(p)=x*} 2^{-l(p)}$ by the dominant contribution in the sum:
  $m(x) := 2^{-Km(x)}$ with $Km(x) := \min_p\{l(p) : U(p) = x*\}$.
  Km is called monotone complexity and is *very* close to KM (Levin:73, Gacs:83).

- A sequence $x_{1:\infty}$ is called computable if $Km(x_{1:\infty}) < \infty$.

- KM, Km, and K are ordered in the following way:
  $$0 \leq K(x|l(x)) \stackrel{+}{\leq} KM(x) \leq Km(x) \leq K(x) \stackrel{+}{\leq} l(x)\cdot\log|\mathcal{X}| + 2\log l(x)$$

- where the prefix Kolmogorov complexity is defined as
  $$K(x) := \min_p\{l(p) : U(p) = x \text{ halts}\}, \qquad k(x) := 2^{-K(x)}.$$

Minimal Description Length Principle

- Generic complexity $\tilde{K} \in \{K, KM, Km, ...\}$ and its associated Predictive functions $\tilde{k}(x) := 2^{-\tilde{K}(x)} \in \{k, M, m, ...\}$.
- $\tilde{k}$ is generally not a semimeasure, so we have to clarify what it means to predict using $\tilde{k}$.
- Popular approach: (Universal) MDL: $y_t^{MDL} := \arg\min_{y_t} \tilde{K}(x_{<t}y_t)$.
  Enumerable Posterior: $\tilde{k}_|(x|y) := 2^{-\tilde{K}_|(x|y)}$,
  Bayes Posterior: $\tilde{k}(x_n|x_{<n}) := \tilde{k}(x_{1:n})/\tilde{k}(x_{<n})$.
- MDL coincides with the $\Lambda_{\tilde{k}}$ predictor for the error loss $\ell_{xy}=1-\delta_{xy}$:
  $$y_t^{\Lambda_{\tilde{k}}} = \arg\min_{y_t} \sum_{x_t} \tilde{k}(x_t|x_{<t})\ell_{x_t y_t} = \arg\min_{y_t} \tilde{K}(x_{<t}y_t) = y_t^{MDL}$$
- Hence, self-optimizingness $l_t^{\Lambda_{\tilde{k}}} \to l_t^{\Lambda_\mu}$ tells us something about the validity of the MDL principle (what good prediction *means*).

Properties of Predictive Functions

We call functions $b, b_| : \mathcal{X}^* \to [0,\infty)$ (conditional) predictive functions. They may possess some of the following properties:

o) Proximity: $b(x_{1:n})$ is "close" to the universal prior $M(x_{1:n})$.

i) Universality: $b \stackrel{\times}{\geq} \mathcal{M}$, i.e. $\forall \nu \in \mathcal{M}\, \exists c > 0 : b(x) \geq c \cdot \nu(x) \forall x$.

ii) Monotonicity: $b(x_{1:n}) \leq b(x_{<n})\ \forall n, x_{1:n}$.

iii) Semimeasure: $\sum_{x_n} b(x_{1:n}) \leq b(x_{<n})$ and $b(\varepsilon) \leq 1$.

iv) Multiplication rule: $b(x_{1:n}) = b.(x_n|x_{<n}) b(x_{<n})$.

v) Enumerability: b is lower semi-computable.

vi) Convergence: $b.(x'_t|x_{<t}) \stackrel{t\to\infty}{\longrightarrow} \mu(x'_t|x_{<t})\ \forall \mu \in \mathcal{M}, x'_t \in \mathcal{X}$.

vii) Self-optimizingness: $l_t^{\Lambda_{b.}} \stackrel{t\to\infty}{\longrightarrow} l_t^{\Lambda_\mu}$ i.m.s. or w.p.1.

where $b.$ refers to b or $b_|$.

Predictive Relations

The importance of the properties $(i) - (iv)$ stems from the fact that they together imply convergence (vi) and self-optimizingness (vii):

$a)$ $(iii) \to (ii)$: A semimeasure is monotone.

$b)$ $(i), (iii), (iv) \to (vi)$: The posterior $b.$ as defined by the multiplication rule (iv) of a universal semimeasure b converges to μ i.m.s. for all $\mu \in \mathcal{M}$.

$c)$ $(i), (iii), (v) \to (o)$: Every w.r.t. $\mathcal{M}^{semi}_{enum}$ universal enumerable semimeasure coincides with M within a multiplicative constant.

$d)$ $(vi) \to (vii)$: Posterior convergence i.m.s./w.p.1 implies self-optimizingness i.m.s./w.p.1.

Predictive Properties of $m = 2^{-Km}$

(o) $m(x_{1:n}) \stackrel{\times}{=} M(x_{1:n})$ for every $\mu \in \mathcal{M}^{msr}_{comp}$ and μ-random $x_{1:\infty}$.

(i) m is universal w.r.t. $\mathcal{M} = \mathcal{M}^{msr}_{comp}$, but not w.r.t. $\mathcal{M} = \mathcal{M}^{semi}_{enum}$.

(ii) m is monotone.

(iii) m is not a semimeasure.

(iv) $m. = m$ respects the multiplication rule, but $m. = m_{|}$ not.

(v) m is enumerable (lower semi-computable).

(vi) For $m. = m$ converges (fast on-, somehow off-sequence) and ...

(vii) ... is self-optimizing for computable deterministic μ, but in general not for probabilistic μ.

The lesson to learn is that although m is very close to M and m dominates all computable measures μ, predictions based on m may nevertheless fail.

Detailed Properties of $m = 2^{-Km}$

(o) $\forall \mu \in \mathcal{M}^{msr}_{comp}\, \forall \mu$-random $\omega\, \exists c_\omega \,:\, Km(\omega_{1:n}) \leq KM(\omega_{1:n}) + c_\omega\, \forall n$,
$KM(x) \leq Km(x) \leq KM(x) + 2\log Km(x)\, \forall x$. Levin:70

$\neg(o)$ $\forall c \,:\, Km(x) - KM(x) \geq c$ for infinitely many x. Gacs:83

(i) $Km(x) \stackrel{+}{\leq} -\log \mu(x) + K(\mu)$ if $\mu \in \mathcal{M}^{msr}_{comp}$, Levin:73
$m \stackrel{\times}{\geq} \mathcal{M}^{msr}_{comp}$, but $m \stackrel{\times}{\not\geq} \mathcal{M}^{semi}_{enum}$ (unlike $M \stackrel{\times}{\geq} \mathcal{M}^{semi}_{enum}$).

(ii) $Km(xy) \geq Km(x) \in \mathbb{N}_0, \quad 0 < m(xy) \leq m(x) \in 2^{-\mathbb{N}_0} \leq 1.$

$\neg(iii)$ If $x_{1:n}$ is computable, then $\sum_{x_n} m(x_{1:n}) \not\leq m(x_{<n})$ for almost all n
If $Km(x_{1:n}) = o(n)$, then $\sum_{x_n} m(x_{1:n}) \not\leq m(x_{<n})$ for most n.

(iv) $0 < m(x|y) := \frac{m(yx)}{m(y)} \leq 1.$

$\neg(iv)$ $\exists x, y : m(yx) \neq m_|(x|y) \cdot m(y)$,
$Km(yx) = Km_|(x|y) + Km(y) \pm O(\log l(y))$.

Detailed Properties of $m = 2^{-Km}$

(v) m is enumerable, i.e. lower semi-computable.

(vi) $\sum_{t=1}^{n} |1-m(x_t|x_{<t})| \leq \frac{1}{2} Km(x_{1:n})$, $m(x_t|x_{<t}) \stackrel{fast}{\longrightarrow} 1$ if $x_{1:\infty}$ comp
Indeed, $m(x_t|x_{<t}) \neq 1$ at most $Km(x_{1:\infty})$ times,
$\sum_{t=1}^{n} \sum_{\bar{x}_t \neq x_t} m(\bar{x}_t|x_{<t}) \leq 2^{Km(x_{1:n})}$, $m(\bar{x}_t|x_{<t}) \stackrel{slow?}{\longrightarrow} 0$ if $x_{1:\infty}$ comp.

$\neg(vi)$ $\exists \mu \in \mathcal{M}^{msr}_{comp} \setminus \mathcal{M}_{det} \; : \; m_{(norm)}(x_t|x_{<t}) \not\to \mu(x_t|x_{<t}) \, \forall x_{1:\infty}$

(vii) $l_t^{\Lambda_m}(x_{<t}) \stackrel{slow?}{\longrightarrow} l_t^{\Lambda_\omega} := \arg\min_{y_t} \ell_{x_t y_t}$ if $\omega \equiv x_{1:\infty}$ is computable.
$\Lambda_m = \Lambda_{m_{norm}}$, i.e. $y_t^{\Lambda_m} = y_t^{\Lambda_{m_{norm}}}$ and $l_t^{\Lambda_m} = l_t^{\Lambda_{m_{norm}}}$.

$\neg(vii)$ $\forall |\mathcal{Y}| > 2 \, \exists \ell, \mu \; : \; l_t^{\Lambda_m} / l_t^{\Lambda_\mu} = c > 1 \, \forall t, n \quad (c = \frac{6}{5} - \varepsilon$ possible),
$\forall' \ell \, \forall \mathcal{X}, \mathcal{Y} \, \exists U, \mu \; : \; l_t^{\Lambda_m} / l_t^{\Lambda_\mu} = L_n^{\Lambda_m} / L_n^{\Lambda_\mu} = c > 1 \, \forall t, n,$
where $\forall' \ell$ means for all non-degenerate ℓ.

Remarks

Simple MDL: Take the shortest (non-halting) program p which outputs x, continue running p, and use the continuation y of x for prediction:

$\widetilde{m}_{|}(x_t|x_{<t}) := 1$ if shortest program for $x_{<t}*$ computes $x_{<t}x_t*$

$\widetilde{m}_{|}(\bar{x}_t|x_{<t}) := 0$.

Predictive properties are worse or at least not better than m.

Predictions based on K: $K(x)$ (and $K(x|l(x))$) are completely unsuitable for prediction, since $K(x0) \stackrel{+}{=} K(x1)$ (and $K(x0|l(x0)) \stackrel{+}{=} K(x1|l(x1))$), which implies that the predictive functions do not even converge for deterministic computable environments.

Outlook and Open Problems

- How fast does $m(\bar{x}_t|x_{<t})$ converge to zero in the deterministic case?
- Can self-optimizingness of Λ_m be violated for $every$ non-degenerate loss-function and universal Turing machine U?
- When does closeness or dominance of unconditional predictive function $\tilde{k}$ to/over M imply good prediction performance?
- What are the predictive properties of plain Kolmogorov complexity C, Schnorr's process complexity, Chaitin's complexity Kc, Cover's extension semimeasure Mc, Loveland's uniform complexity, Schmidhuber's cumulative K^E and general K^G, Vovk's predictive complexity KP, Schmidhuber's speed prior S, Levin complexity Kt, and others?

- Many properties and relations are known for the unconditional versions, but little relevant for prediction of the conditional versions is known.
- Levin's representation of M as a mixture over semi-measures leads to the universal two-part MDL approximation
  $Km_2(x) := \min_{\nu \in \mathcal{M}_{enum}^{semi}} \{-\log \nu(x) + K(\nu)\}$.
  What are the predictive properties of Km_2, similar to Km?
- More abstract proofs showing that violation of some of the criteria $(i) - (iv)$ necessarily lead to violation of (vi) or (vii) may deal with a number of complexity measures simultaneously.
- Non-convergence or non-self-optimizingness of m does not necessarily mean that m fails in practice. Characterize the class of environments for which universal MDL alias m converges and self-optimizes rapidly.

Summary

- We studied the predictive properties of complexity measures, esp. KM and Km. Performance was measured in terms of convergence of posteriors or losses.
- We enumerated and related eight important properties, which general predictive functions may posses or not: proximity to M, universality, monotonicity, being a semimeasure, the multiplication rule, enumerability, convergence, and self-optimizingness.
- The monotone complexity $Km = -\log m$ is, in a sense, closest to KM. While KM is defined via a mixture of programs, Km approximates KM by the contribution of the single shortest program.
- This captures the spirit of Occam's razor and the popular Minimal Description Length (MDL) principle.

Conclusions

- Closeness of "priors" does neither necessarily imply closeness of "posteriors", nor good performance from a decision-theoretic perspective.
- For deterministic, computable environments, the MDL posterior of $m = 2^{-Km}$ converges and is self-optimizing, but rapid convergence could only be shown on-sequence; the off-sequence behavior is unclear. In the presence of noise, m neither converges, nor is it self-optimizing, in general.
- Some complexity measures like K, fail completely for prediction.