Computable prediction

Kenshi Miyabe

Meiji University

6 Nov, 2023

Mini-workshop on computability in Tokyo

by ChatGPT

We compute the probability of the next bit from a finite sequence of 0 and 1.

Abstract

1. The theory of inductive inference usually considers an optimal c.e. semi-measure because no computable measure is optimal.
2. We introduce reducibility among measures by domination.
3. Domination means generality.
4. We give the convergence rate of a sufficiently general measure.

Table of Contents

Setting

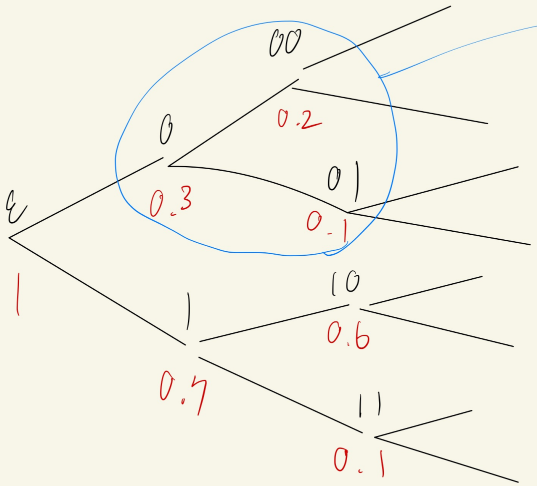
Cantor space: $\{0,1\}^{\mathbb{N}}$

μ: unknown measure on $\{0,1\}^{\mathbb{N}}$, model measure

$X = X_1X_2X_3\cdots \in \{0,1\}^{\mathbb{N}}$: random sequence w.r.t. μ

The task is to predict the conditional probability of the next bit given initial segments $X_{<n}$:

$$\mu(k|X_{<n}) = \frac{\mu(X_{<n}k)}{\mu(X_{<n})} \quad \text{for } k \in \{0,1\}.$$



$$X_{\le 1} = 0$$

$$\mu(X_{\le 1}) = 0.3$$

$$\mu(X_{\le 1}0) = \frac{0.2}{0.3}$$

$$\mu(X_{\le 1}1) = \frac{0.1}{0.3}$$

$$\mu(k \mid X_{\le 1}) = \begin{cases} \frac{0.2}{0.3} & k=0 \\ \frac{0.1}{0.3} & k=1 \end{cases}$$

Optimal prediction

$\xi : \{0,1\}^* \to [0,1]$ is called a semi-measure if

$$\xi(\epsilon) \leq 1,\ \xi(\sigma) \geq \xi(\sigma 0) + \xi(\sigma 1)$$

for all $\sigma \in \{0,1\}^{\mathbb{N}}$ where ϵ is the empty string.

$\xi : \{0,1\}^* \to [0,1]$ is called c.e.(or lower semicomputable) if if $\xi(\sigma)$ are left-c.e. uniformly in σ.

A c.e. semi-measure ξ is called optimal if it dominates all c.e. semi-measures, that is, for every c.e. semi-measure μ, there exists $c \in \mathbb{N}$ such that

$$\mu(\sigma) \leq c \cdot \xi(\sigma)$$

for all $\sigma \in \{0,1\}^*$. An optimal c.e. semi-measure exists.

Solomonoff's result

μ: model measure, ξ: prediction measure

Theorem 1 (Solomonoff 1960s-70s)

ξ: optimal c.e. semi-measure, μ: computable model measure, X: μ-random sequence

$$|\xi(k|X_{<n}) - \mu(k|X_{<n})| \to 0 \ (n \to \infty) \quad \textit{for each } k \in \{0,1\}$$

ML-randomness is not sufficient (Hutter and Muchnik 2007), but 2-randomness is sufficient.

Strengths and weaknesses

Strengths:

1. The optimal measure seems to accomplish the so-called artificial general intelligence!

Weaknesses:

1. No computable prediction is optimal. Thus, this does not make any sense in reality.
2. The convergence rate can be very slow.

Table of Contents

Domination

ξ dominates ν if

$$(\exists c \in \mathbb{N})(\forall \sigma \in \{0,1\}^{\mathbb{N}})\nu(\sigma) \leq c \cdot \xi(\sigma)$$

Optimal semi-measures perform well in prediction.

Question 2

If ξ dominates μ, does it mean that ξ performs better than μ in prediction?

If so, in what sense?

KL-divergence

μ, ξ: measures on $\{0, 1\}$

Kullback-Leibler divergence of μ w.r.t. ξ is defined by

$$d(\mu \mid\mid \xi) = \sum_{k \in \{0,1\}} \mu(k) \ln \frac{\mu(k)}{\xi(k)}$$

- $d_\sigma(\mu \mid\mid \xi) = d(\mu(\cdot \mid \sigma) \mid\mid \xi(\cdot \mid \sigma))$,
- $D_n(\mu \mid\mid \xi) = \sum_{k=1}^{n} E_{X \sim \mu}[d_{X_{<k}}(\mu \mid\mid \xi)]$,
- $D_\infty(\mu \mid\mid \xi) = \lim_{n \to \infty} D_n(\mu \mid\mid \xi)$.

KL-divergence and convergence

Suppose $D_\infty(\mu \mid\mid \xi) < \infty$. Then,

$$|\mu(k|X_{<n}) - \nu(k|X_{<n})| \to 0$$

as $n \to \infty$ almost surely. Thus, the finiteness of KL-divergence is a sufficient condition for the convergence.

Domination and convergence

Theorem 3

The following are equivalent for ξ, ν:

- ξ *dominates* ν.
- *There exists* $c \in \mathbb{N}$ *such that, for every measure* μ, *we have* $D_\infty(\mu \mid\mid \xi) \leq D_\infty(\mu \mid\mid \nu) + c$.

The sum of expected errors of ξ is smaller than that of ν up to a constant.

Remark

- $D_\infty(\mu \,||\, \xi)$ is equal to the usual KL-divergence of μ w.r.t. ξ. A finite version is called the chain-rule for the KL-divergence. I couldn't find an infinite version in the literature.
- If μ and ξ are computable, then $D_\infty(\mu \,||\, \xi)$ is left-c.e. or ∞.
- domination $\Rightarrow$ absolute continuous. Kakutani equivalence theorem is about absolute continuity. The above claim is its domination version.

Table of Contents

Sufficiently good function

Kolmogorov complexity is not computable. No computable function f satisfies

$$f(n) \le K(n) + O(1),\ \sum_n 2^{-f(n)} < 1$$

However, there exists a computable function such that

$$f(n) \le K(n) + O(1)\ i.o.,\ \sum_n 2^{-f(n)} < 1,$$

which is called a Solovay function.

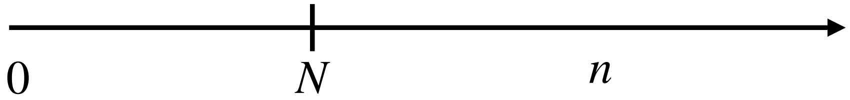
Such a function is sufficiently good in an approximation of K.

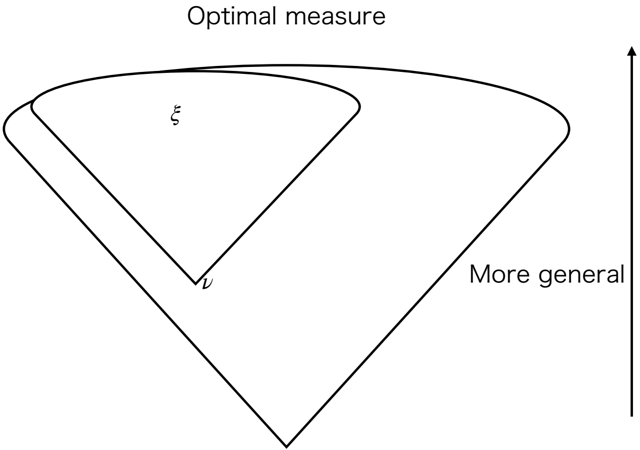
Sufficiently general prediction

$P(n)$ holds for a sufficiently large $n \in \mathbb{N}$ if there exists $N \in \mathbb{N}$ such that $P(n)$ for all $n > N$.

$P(\xi)$ holds for a sufficiently general computable measure ξ if there exists a computable measure ν such that $P(\xi)$ holds for all computable measures ξ dominating ν.

- We are talking about computable measures.
- We can prove a computable version of Solomonoff's result.


0
N
n



Optimal measure
ξ
ν
More general

Main result

Theorem 4

μ: computable model measure, ξ: sufficiently general prediction measure

Then,

$$D_\infty(\mu||\xi) < \infty$$

and is Martin-Löf left-c.e. real.

Remark: ML-randomness w.r.t. the Lebesgue measure.

Fast convergence because the sum of the expected errors converges.

Slow convergence because the sum is ML-random.

Proof sketch

Steps of the proof of ML-randomness of $D_\infty(\mu||\xi)$.

1. Construct ν which is close to μ but slightly different.
2. Show that $D_\infty(\mu||\nu)$ is ML-random.
3. If ξ dominates ν, then $D_\infty(\mu||\xi) - D_\infty(\mu||\nu)$ is left-c.e.
4. Hence, $D_\infty(\mu||\xi)$ is ML-random.

Construction of ν

z_n: computable seq. of positive rationals s.t. $s = \sum_n z_n < 1$ is ML-random

$Z^\sigma \in \{0,1\}^{\mathbb{N}}$: $\sigma \in Z^\sigma$, $\mu(Z^\sigma) = 0$

$$\mu_n(\sigma) = \begin{cases} \mu(\sigma) & \text{if } |\sigma| \le n \\ \mu(\tau)\mathbf{1}_{Z^\tau} & \text{if } |\sigma| > n,\ \tau = \sigma_{\le n} \end{cases}$$

$$\nu = \sum_n z_n \mu_n + (1-s)\mu \quad \text{(computable)}$$

μ_n divides the weights the same as μ until the n-th bit.

Afterward, μ_n puts the weight on a sequence orthogonal to μ.

ML-randomness

Consider the Radon-Nikodym derivative

$$\frac{d\mu}{d\nu} = \frac{1}{1-s}$$

Since s is ML-random, so is $D(\mu||\nu)$.

Furthermore, $\frac{d\mu}{d\nu}$ is a constant function and

$$D(\mu||\xi) = D(\mu||\nu) + \frac{1}{1-s} \cdot D(\nu||\xi),$$

which implies ML-randomness of $D(\mu||\xi)$.

Some corollaries

ML-randomness of KL-divergence implies ML-randomness for other distances of measures.

- $\ell_p(\mu, \xi) = \sum_{k \in \{0,1\}} |\mu(k) - \xi(k)|^p$
- $L_p(\mu, \xi) = \sum E_{X \sim \mu}[\ell_{p, X_{<k}}(\mu, \xi)]$

If there exists p such that $L_p(\mu, \xi)$ is ML-randomness, then such p is unique. Let $R(\mu, \xi)$ be the p.

Theorem 5

When $\mu = \mathbf{1}_A$ is a Dirac measure, then $R(\mu, \xi) = 1$ for sufficiently general ξ.

When μ is separated ($\inf_{k,\sigma} \mu(k|\sigma) > 0$), then $R(\mu, \xi) = 2$ for sufficiently general ξ.

Oscillation

In particular, when $\mu = \mathbf{1}_A$ is a Dirac measure, we have

$$-\log(1 - \xi(A_n | A_{<n})) \approx K^h(n)$$

where $K^h(n)$ is the time-bounded Kolmogorov complexity.

Thus, the error can be evaluated completely in some sense.

Nicod's criterion

The sunrise problem asks "What is the probability that the sun will rise tomorrow?"

Nicod's criterion claims that a hypothesis of the form "All A are B" is confirmed (so the probability should be greater) by further instances that are A and B.

However, the previous result violates this claim.

How should we understand this??

Bernoulli measures

Let μ be a Bernoulli measure.

We restrict ξ to be a linear combination of Bernoulli measures.

Definition 6

Let $\mathcal{B}$ be the class of prediction measures μ satisfying the following:

1. $w_n \in [0,1]$ such that $\sum_n w_n = 1$,
2. $p_n \in [0,1]$ such that $(p_n)_n$ is a sequence of uniformly computable reals,
3. $\mu = \sum_{n=1}^{\infty} w_n B_{p_n}$ is a computable measure.

For a computable real p, $\sum_n \{w_n \ : \ p_n = p\}$ is a right-c.e. real.

Bernoulli measures

Theorem 7

1. *If* $X \in \{0^{\mathbb{N}}, 1^{\mathbb{N}}\}$*, then* $\xi(X_n \mid X_{<n})$ *is monotonically increasing to* 1*.*
2. $D_\infty(B_p \mid\mid \xi)$ *is a finite left-c.e. ML-random real for sufficientlly general* ξ*.*

The constructed ν in 2 should be in $\mathcal{B}$.

Nicod's criterion may implicitly assume independence.

Summary

1. We introduced the notion of sufficient generality by domination. Domination roughly means better prediction.
2. The sum of expected errors of sufficient general prediction is left-c.e. ML-random real, which is rather different from the case of optimal prediction.
3. The rate of convergence can be studied through ML-randomness.

Related work

1. Statistical learning theory: usually parametrized model measures, independence
2. (C)PAC learning: usually studies learnability in polynomial time by topological reasons, independence
3. Algorithmic probability: computable model measures, non-independence, restricted to a space of sequences of alphabets

Thank you for listening.

明治大学
MEIJI UNIVERSITY