Universal Prediction without assuming either Discrete or Continuous

Joe Suzuki

Osaka University

November 13, 2012

What is the probability that the sun will rise tomorrow?

Predict $x_{n+1} \in \{0,1\}$ given $x^n := (x_1, \cdots, x_n) \in \{0,1\}^n$

Construct a computable $Q(x_{n+1}|x^n) \to P(x_{n+1}|x^n)$

such as

1. $Q(x_{n+1}|x^n) = \dfrac{c}{n}$
2. For $a, b > 0$, $Q(x_{n+1}|x^n) = \dfrac{c+a}{n+a+b}$

c: the number of x_{n+1} in x^n.

Open Problems raised by Tom Cover in 1975, Moscow

In the betting, obtain 2 dollars if you win, or lose 1 dollar otherwise.

Problem 1: Existence of a universal gambling scheme

Is there any Q^n s.t.

$$\frac{1}{n}\log[2^n Q^n(x^n)] \to \frac{1}{n}\log[2^n P^n(x^n)]$$

a.s. $n \to \infty$ for any unknown stationary ergodic P^n ?

Betting without knowledge converges to one with knowledge
(Bayesian strategy realizes the property)

Problem 2: Existence of a universal prediction scheme

Is there any Q s.t. for $x \in \{0,1\}$

$$Q(x|x_{-n}^{-1}) \to P(x|x_{-\infty}^{-1})$$

a.s. $n \to \infty$ for any unknown stationary ergodic P ?

- Ornstein 1978 (discrete, Non-Bayesian)
- Algoet 1992 (extended to the Polish spaces, Non-Bayesian)

$x_{-\infty}^{-1} \in \{0,1\}^\infty \mapsto (\{s_k\},\{t_k\})$, $s_0 < s_1 < \cdots$, $t_0 < t_1 < \cdots$ s.t.

$$Q(x|x_{-t_k}^{-1}) = \frac{\#I_k(x) + 1/2}{\#I_k(0) + \#I_k(1) + 1}$$

$$I_k(x) = \{1 \le \tau \le s_k | x = x_{-\tau}, x_{-t_k}^{-1} = x_{-\tau-t_k}^{-\tau-1}\}$$

Bayesian for binary i.i.d. sources

$$Q^n(x^n) = \int w(\theta) P(x^n|\theta) d\theta \ , \ P(x^n|\theta) = \theta^c (1-\theta)^{n-c}$$

For $a, b > 0$,

$$w(\theta) \propto \theta^{-a}(1-\theta)^{-b} \Longleftrightarrow Q(x_{n+1}|x^n) = \frac{Q^{n+1}(x^{n+1})}{Q^n(x^n)} = \frac{c+a}{n+a+b}$$

For $a = b = 1/2$ (Krichevsky-Trofimov),

$$-\frac{1}{n} \log Q^n(x^n) \to H := \sum_{x \in A} -P(x) \log P(x)$$

$$-\frac{1}{n} \log P^n(x^n) = \frac{1}{n} \sum_{i=1}^{n} -\log P(x_i) \to E[-\log P(x_i)] = H$$

Universality

There exists Q^n s.t. for any P^n

1. $$Q(x|x_{-n}^{-1}) \to P(x|x_{-\infty}^{-1}) \quad (1)$$
2. $$\frac{1}{n}\log\frac{P^n(x^n)}{Q^n(x^n)} \to 0 \quad (2)$$

- m-nary ($m \geq 2$) rather than binary
- stationary ergodic rather than i.i.d.

Ornstein 1978 (1)

Bayesian (2) as well as (1)

Problem

Construct Q^n satisfying (2) for the genaral case

X^n should be stationary ergodic but can be either

- discrete,
- continuous, or
- neither of them

- Counting how many $(X = x_{i+1}, X^i = x^i)$ occurs does not help.
- Algoet 1992 does not imply (2) for the general case.

Suppose a density function f exists for X

A: the range of X

- $A_0 := \{A\}$
- A_{j+1} is a refinement of A_j

Example 1: Quantize f over $A = [0, 1)$ to obtain histogram approximations

f_1 over $A_1 = \{[0, 1/2), [1/2, 1)\}$

f_2 over $A_2 = \{[0, 1/4), [1/4, 1/2), [1/2, 3/4), [3/4, 1)\}$

$\cdots$

f_j over $A_j = \{[0, 2^{-(j-1)}), [2^{-(j-1)}, 2 \cdot 2^{-(j-1)}), \cdots, [(2^{j-1} - 1)2^{-(j-1)}, 1)\}$

$\cdots$

$P_j^n(a^n) = \prod_{i=1}^n P_j(a_i)$, the probability of $a^n = (a_1, \cdots, a_n) \in A_j^n$

Q_j^n: a Bayesian measure $\frac{1}{n} \log \frac{P_j^n(a^n)}{Q_j^n(a^n)} \to 0$ as $n \to \infty$

$\lambda : \mathbb{R} \to \mathcal{B}$ (Lebesgue measure, $a = [b, c) \Longrightarrow \lambda(a) = c - b$)

$$(x_1, \cdots, x_n) \in (a_1, \cdots, a_n) \in A_j^n$$

$$\Longrightarrow \begin{cases} f_j^n(x^n) := f_j(x_1) \cdots f_j(x_n) = \dfrac{P_j(a_1) \cdots P_j(a_n)}{\lambda(a_1) \ldots \lambda(a_n)} \\ g_j^n(x^n) := \dfrac{Q_j^n(a_1, \cdots, a_n)}{\lambda(a_1) \cdots \lambda(a_n)} \end{cases}$$

For $\{\omega_j\}_{j=1}^{\infty}$: $\sum \omega_j = 1$, $\omega_j > 0$, $g^n(x^n) := \sum_{j=1}^{\infty} \omega_j g_j^n(x^n)$

If we choose $\{A_j\}$ such that $f_j \to f$ as $j \to \infty$, for any f, almost surely

$$\frac{1}{n} \log \frac{f^n(x^n)}{g^n(x^n)} \to 0 \tag{3}$$

B. Ryabko. *IEEE Trans. on Inform. Theory*, 55, 9, 2009.

Exactly when does density function exist?

$\mathcal{B}$: the Borel sets of $\mathbb{R}$
$\mu(D)$: the probabbility of $D \in \mathcal{B}$

When a density function exists

The following are equivalent ($\mu \ll \lambda$):

- for each $D \in \mathcal{B}$, $\lambda(D) = 0 \Longrightarrow \mu(D) = 0$
- $\exists$ $\mathcal{B}$-measurable $\dfrac{d\mu}{d\lambda} := f$ s.t. $\mu(D) = \displaystyle\int_D f(t) d\lambda(t)$

Estimating generalized density functions

Radon-Nikodym's Theorem

The following are equivalent ($\mu \ll \eta$):

- for each $D \in \mathcal{B}$, $\eta(D) = 0 \Longrightarrow \mu(D) = 0$
- $\exists$ $\mathcal{B}$-measurable $\frac{d\mu}{d\eta} := f$ s.t. $\mu(D) = \int_D f(t) d\eta(t)$

Example 2: $\mu(\{k\}) > 0$, $\eta(\{k\}) := \frac{1}{k(k+1)}$, $k \in B := \{1, 2, \cdots\}$

$$\mu(D) = \sum_{k \in D} f(k)\eta(\{k\}) \ , \ D \subseteq B$$

$$\mu \ll \eta \Longrightarrow \frac{d\mu}{d\eta}(k) = f(k) = \frac{\mu(\{k\})}{\eta(\{k\})} = k(k+1)\mu(\{k\})$$

f_1 over $B_1 := \{\{1\}, \{2, 3, \cdots\}\}$
f_2 over $B_2 := \{\{1\}, \{2\}, \{3, 4, \cdots\}\}$
$\cdots$
f_k over $B_k := \{\{1\}, \{2\}, \cdots, \{k\}, \{k+1, k+2, \cdots\}\}$
$\cdots$

$$(y_1, \cdots, y_n) \in (b_1, \cdots, b_n) \in B_k^n \Longrightarrow g_k^n(y^n) := \frac{Q_k^n(b_1, \cdots, b_n)}{\eta(b_1) \cdots \eta(b_n)}$$

$$g^n(y^n) := \sum_{k=1}^{\infty} \omega_k g_k^n(y^n)$$

If we choose $\{B_k\}$ s.t. $f_k \to f$, for any f, almost surely

$$\frac{1}{n} \log \frac{f^n(y^n)}{g^n(y^n)} \to 0 \tag{4}$$

$g^n(y^n) \prod_{i=1}^n \eta^n(\{y_i\})$ estimates $P(y^n) = f^n(y^n) \prod_{i=1}^n \eta^n(\{y_i\})$

The original case was contained as a special case

For $C = \{0, 1, \cdots, m-1\}$, if we quantize

$$C_1 = C_2 = \cdots = \{\{0\}, \{1\}, \cdots, \{m-1\}\}$$

$$\eta(\{0\}) = \cdots \eta(\{m-1\}) = 1/m$$

then $\mu \ll \eta$ and

$$z^n \in C^n \Longleftrightarrow c^n \in C_1^n = C_2^n = \cdots$$

$$\Longrightarrow \begin{cases} f^n(z^n) = \dfrac{P^n(c^n)}{(1/m)^n}, \\ g_1^n(z^n) = g_2^n(z^n) = \cdots = g^n(z^n) = \displaystyle\sum_{l=1}^{\infty} \omega_l g_l^n(z^n) = \dfrac{Q^n(c^n)}{(1/m)^n} \end{cases}$$

$$\Longrightarrow \frac{1}{n} \log \frac{f^n(z^n)}{g^n(z^n)} = \frac{1}{n} \log \frac{P^n(c^n)}{Q^n(c^n)} \to 0$$

Universality in the generalized sense

If $\mu^n \ll \eta^n$, there exists g^n without depending on f^n s.t.

$$\frac{1}{n}\log\frac{f^n(z^n)}{g^n(z^n)} \to 0$$

$$\mu^n(D^n) := \int_D f^n(z^n)d\eta^n(z^n)\ ,\ \nu^n(D^n) := \int_D g^n(z^n)d\eta^n(z^n)$$

$$\frac{f^n(z^n)}{g^n(z^n)} = \frac{d\mu^n}{d\eta^n}(z^n)/\frac{d\nu^n}{d\eta^n}(z^n) = \frac{d\mu^n}{d\nu^n}(z^n)$$

Theorem (Suzuki, 2011)

$$\frac{1}{n}\log\frac{d\mu^n}{d\nu^n}(z^n) \to 0$$

Universal Prediction in the generalized sense

The generalzed universal density function tells everything:

$$g(x_{n+1}|x^n) = \frac{g^{n+1}(x^{n+1})}{g^n(x^n)} \to f(x_{n+1}|x^n) = \frac{f^{n+1}(x^{n+1})}{f^n(x^n)}$$

For any $D \in \mathcal{B}$,

$$\nu(D|x^n) = \int_D g(x|x^n) d\eta(x)$$

Summary and Discussion

Universal Prediction

- Connection to Universal Bayesian Measures
- Generalization without assuming Discrete or Continuous
- Stronger universality in the sense of Bayes.

Many Applications except Prediction

- Bayesian network structure estimation (DCC 2012)
- The Bayesian Chow-Liu Algorithm (PGM 2012)
- Markov order estimation even when $\{X_i\}$ is continuous