On Sequence Prediction for Arbitrary Measures

Daniil Ryabko and Marcus Hutter

ISIT, June 2007

Sequence prediction

Given a sequence $x_1, \dots, x_n$ generated by the *environment* predict x_{n+1}, where x_i are from a finite set X. *Environment* here is just a probability measure μ over X^∞.

The task can be formulated as forecasting probabilities for x_{n+1}.

In this case the predictor also defines a probability measure over X^∞.

Sequence prediction

Laplace: weather forecasting. The Sun has risen every day for 5000 years, what is the probability that it will not rise tomorrow? X is binary: the Sun rises vs. it does not.

Laplace suggested that x_i — the Sun rising on different days — are independent and identically distributed.
His predictor:

$$\rho_L(x_{n+1} = 0 | x_1, \dots, x_n) = \frac{k+1}{n+2} \approx \frac{1}{1830000}$$

where k is $\#1$ in $x_1 \dots x_n$ (derived as a Bayesian w. uniform prior).

Markov processes and Stationary

The same idea generalizes to Markov and k-order Markov measures. For each k, a predictor ρ_k can be constructed that predicts any k-order Markov process.

A predictor ρ_R (B. Ryabko, 1988) for the class of all *stationary* process is constructed as a sum of predictors for k-order Markov measures:

$$\rho_R(x_1, \dots, x_n) = \sum_{k=0}^{\infty} w_i \rho_k(x_1, \dots, x_n),$$

Side question: what else does it predict?

Solomonoff: computable probability measures

Another assumption: μ is computable.

The class of all computable measures is countable: $(\nu_i)_{i\in\boldsymbol{N}}$.

A Bayesian predictor: $\xi(A) = \sum_{i=1}^{n} w_i \nu_i(A)$ for any measurable set A, where the weights w_i are positive and sum to one.

A measure μ is the best predictor for itself; for a countable class of measures we can just sum all the predictors for individual measures.

Dominance by a constant and absolute continuity

For the Bayes mixture ξ over a countable class ν_i, $i \in \mathbb{N}$ we have

$$\xi(A) \geq c\nu_i(A)$$

for every ν_i and every (measurable) set A, where c is a constant $c = w_i$. ξ dominates each ν_i with a constant $c = w_i$. In particular, each ν_i is absolutely continuous with respect to ξ.

Absolute continuity is sufficient for prediction (Blackwell and Dubins, 1962).

General open questions

- For which classes of measures is prediction possible? So far we have only some interesting examples.
- Given two probability measures, under which conditions does one of them predict the other? So far we only have absolute continuity — which is too strong, and some examples.

New stuff: dominance with decreasing coefficients

For Bayes mixture ξ over (computable) measures ν_i, $i \in \boldsymbol{N}$ we have $\xi(A) \geq c\nu_i(A)$ for every ν_i and every (measurable) set A.

For Laplace measure ρ_L we have

$$\rho_L(x_1, \ldots, x_n) \geq \frac{1}{n+1}\mu_\delta(x_1, \ldots, x_n)$$

for each Bernoulli μ_δ.

Is any such property in itself sufficient for prediction?

$$\rho(x_1, \ldots, x_n) \geq c_n\mu(x_1, \ldots, x_n) \tag{1}$$

for any $x_1, \ldots, x_n$, where $c_n \to 0$ not too fast.

Divergence characteristics

(d) Kullblack-Leibler (KL) divergence
$d_t(\mu, \rho, x_{<n}) = \sum_{x \in X} \mu(x_n = x | x_{<n}) \log \frac{\mu(x_n = x | x_{<n})}{\rho(x_n = x | x_{<n})}$,

($\bar{d}$) average KL divergence $\bar{d}_n(\mu, \rho) = \frac{1}{n} \sum_{i=1}^{n} d_i(\mu, \rho, x_{<i})$,

(a) absolute distance

$$a_t(\mu, \rho, x_{<n}) = \sum_{x \in X} |\mu(x_n = x | x_{<n}) - \rho(x_n = x | x_{<n})|,$$

($\bar{a}$) average absolute distance $\bar{a}_n(\mu, \rho) = \frac{1}{n} \sum_{i=1}^{n} a_i(\mu, \rho, x_{<n})$.

Thus we say that ρ predicts μ

(d) in KL divergence if $d_n(\mu, \rho, x_{<n}) \to 0$ μ-a.s.,

($\bar{d}$) in average KL divergence if $\bar{d}_n(\mu, \rho, x_{1..n}) \to 0$ μ-a.s.

($\mathbf{E}\,\bar{d}$) in expected average KL divergence if $\mathbf{E}_\mu \, \bar{d}_t(\mu, \rho, x_{1..t}) \to 0$

(a) in absolute distance if $a_n(\mu, \rho, x_{<n}) \to 0$ μ-a.s.,

($\bar{a}$) in average absolute distance if $\bar{a}_n(\mu, \rho, x_{1..n}) \to 0$ μ-a.s.

($\mathbf{E}\,\bar{a}$) in expected average absolute distance if $\mathbf{E}_\mu \, \bar{a}_n(\mu, \rho, x_{1..n}) \to 0$

Results about dominance with decreasing coefficients

	$\mathbf{E}\,\bar{d}_n$	$\bar{d}_n$	d_n	$\mathbf{E}\,\bar{a}_n$	$\bar{a}_n$	a_n
$\log c_n^{-1} = o(n)$	$+$	?	$-$	$+$	?	$-$
$\sum_{n=1}^{\infty} \frac{\log c_n^{-1}}{n^2} < \infty$	$+$	$+$	$-$	$+$	$+$	$-$
$c_n \geq c > 0$	$+$	$+$	$+$	$+$	$+$	$+$

Theorem

Let μ and ρ be two measures on X^∞ and suppose that $\rho(x_{1..n}) \geq c_n \mu(x_{1..n})$ for any $x_{1..n}$, where c_n are positive constants satisfying

$$\sum_{n=1}^{\infty} \frac{(log c_n^{-1})^2}{n^2} < \infty.$$

Then ρ predicts μ in average KL divergence μ-a.s.

	$\mathbf{E}\,\bar{d}_n$	$\bar{d}_n$	d_n	$\mathbf{E}\,\bar{a}_n$	$\bar{a}_n$	a_n
$\log c_n^{-1} = o(n)$	$+$	?	$-$	$+$	?	$-$
$\sum_{n=1}^{\infty} \frac{\log c_n^{-1}}{n^2} < \infty$	$+$	$+$	$-$	$+$	$+$	$-$
$c_n \geq c > 0$	$+$	$+$	$+$	$+$	$+$	$+$

Theorem

For each sequence of positive numbers c_n that goes to 0 there exist measures μ and ρ and a number $\varepsilon > 0$ such that $\rho(x_{1:n}) \geq c_n \mu(x_{1:n})$ for all $x_{1:n}$, yet $a_n(\mu, \rho | x_{1:n}) > \varepsilon$ and $d_n(\mu, \rho | x_{1:n}) > \varepsilon$ infinitely often μ-a.s.

How to combine predictors?

If a measure ρ predicts a measure μ does $\rho + \chi$ also predict μ, for an arbitrary measure χ?
In particular, if we have two predictors, can we just sum them to obtain a predictor that combines predictive powers?

$\mathbf{E}\,\bar{d}_n$	$\bar{d}_n$	d_n	$\mathbf{E}\,\bar{a}_n$	$\bar{a}_n$	a_n
$+$	?	$-$	$-$	$-$	$-$

Open questions

- Which classes of measures admit a predicting measure (that predicts all of them)?
- Under which conditions on two process measures does one measure predict the other?
- How to combine predictors, saving there predictive abilities?