On-line regression competitive with reproducing kernel Hilbert spaces

Vladimir Vovk

Department of Computer Science
Royal Holloway, University of London
Egham, Surrey, England

vovk@cs.rhul.ac.uk

Beijing, May 16, 2006

On-line prediction protocol:

FOR $n = 1, 2, \ldots$:
  Reality announces $x_n \in \mathbf{X}$.
  Predictor announces $\mu_n \in \mathbb{R}$.
  Reality announces $y_n \in \mathbb{R}$.
END FOR.

y_n: label; x_n: object (all relevant information, perhaps including some of the previous labels); (x_n, y_n): example.

The quality of prediction is measured by $(y_n - \mu_n)^2$.

Predictor wants to compete with a wide range of prediction rules $D : \mathbf{X} \to \mathbb{R}$ (competitive on-line prediction).

Universal consistency

Suppose $\mathbf{X} \subseteq \mathbb{R}^m$. An on-line prediction algorithm is universally consistent if its predictions μ_n always satisfy

$$(x_n \in A \ \ \& \ \ y_n \in B, \ \ \forall n \in \{1, 2, \ldots\})$$
$$\Longrightarrow \limsup_{N \to \infty} \left(\frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 - \frac{1}{N} \sum_{n=1}^{N} (y_n - D(x_n))^2 \right) \leq 0$$

for any compact $A \subseteq \mathbf{X}, B \subseteq \mathbb{R}$ and any continuous prediction rule $D : \mathbf{X} \to \mathbb{R}$.

This is a minimal requirement for successful prediction.

Proposition 1 There exists a universally consistent on-line prediction algorithm.

Corollary: the assumption $\mathbf{X} \subseteq \mathbb{R}^m$ can be replaced by "$\mathbf{X}$ is a separable metric space" or "$\mathbf{X}$ is a compact metric space" (follows from standard topological results).

Comparison with the statistical definition

Statistical setting: the examples are generated independently from the same distribution. The definition of universal consistency (Stone 1977): similar but no requirement that D should be continuous.

If the requirement of continuity of D is dropped from our definition, universal consistency becomes impossible to achieve (Reality can choose $x_n \to c$, where c is a point of discontinuity of D).

A positive argument in favor of the requirement of continuity of D: it is natural for Predictor to compete only with computable prediction rules, and continuity is often regarded as a necessary condition for computability (Brouwer's "continuity principle").

Statistical setting: very restrictive. Compactness and continuity may be said to be satisfied automatically: every measurable prediction rule can be arbitrarily well approximated by a continuous one, and every probability measure is almost concentrated on a compact set.

From now on: Reality is obliged to produce $y_n \in [-Y, Y]$ for an *a priori* known Y.

Function spaces

Universal consistency: attractive but asymptotic and does not tell us anything about the algorithm's behavior on finite data sequences.

A function space $\mathcal{F}$ on $\mathbf{X}$ is universal if for every compact $A \subseteq \mathbf{X}$ every continuous function on A can be arbitrarily well approximated in the metric $C(A)$ by functions in $\mathcal{F}$ (compact $\mathbf{X}$: Steinwart 2001).

Constructing universally consistent algorithms

1. Choose a universal function space (benchmark class) $\mathcal{F}$.
2. Construct an algorithm satisfying

$$\frac{1}{N}\sum_{n=1}^{N}(y_n-\mu_n)^2 \le \frac{1}{N}\sum_{n=1}^{N}(y_n-D(x_n))^2 + F\left(N,\|D\|_{\mathcal{F}}\right),$$

where the regret term $F\left(N,\|D\|_{\mathcal{F}}\right)$ tends to 0 as $N\to\infty$ (and does not depend too much on $\|D\|_{\mathcal{F}}$).

The second item: a pragmatic statement of universality.

It is not enough for $\mathcal{F}$ to be universal; we also want as many prediction rules D as possible to belong to $\mathcal{F}$, or at least to be well approximated by the elements of $\mathcal{F}$; and we want $\|D\|_{\mathcal{F}}$ to be as small as possible.

Hilbert function spaces

$\mathcal{F}$: Banach space of continuous functions on $\mathbf{X}$ with the pointwise vector-space operations. Easy case: $\mathcal{F}$ is a Hilbert space imbedded in $C(\mathbf{X})$.

The imbedding constant:

$$\mathbf{c}_{\mathcal{F}} := \sup_{x \in \mathbf{X}} \sup_{f : \|f\|_{\mathcal{F}} \leq 1} |f(x)| ,$$

assumed finite. (RKHS: the internal sup is finite for all x.)

Sobolev space

If $\mathbf{X} = \mathbb{R}$,

$$\|D\|_{\mathcal{F}}^2 := \int_{-\infty}^{\infty} (D(t))^2 \, \mathrm{d}t + \int_{-\infty}^{\infty} \left(D'(t)\right)^2 \, \mathrm{d}t$$

for absolutely continuous function $D : \mathbb{R} \to \mathbb{R}$; include in $\mathcal{F}$ all D with $\|D\|_{\mathcal{F}} < \infty$.

It is universal and

$$\mathbf{c}_{\mathcal{F}} = 1/\sqrt{2} \approx 0.71.$$

Three proof techniques (there are other in competitive on-line prediction):

- defensive forecasting
- aggregating algorithm
- gradient descent and its versions

Each can be used for universally consistent prediction.

Defensive forecasting

Origin: algorithmic randomness theory (Kolmogorov 1965, Martin-Löf 1966).

Levin 1976: uniform tests of randomness $t(\omega \mid Q)$; universal test d. For each t there exists a neutral measure Q_t: $t(\omega \mid Q_t) \le 1$ for all ω.

$M = Q_d$ is universal: every sequence is random w.r. to M.

M is not computable. Possible cures: consider non-universal t; split the monolithic Q into many one-step-ahead forecasts.

Testing forecasts

$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots$:
  Reality announces $x_n \in \mathbf{X}$.
  Predictor announces $\mu_n \in [-Y, Y]$.
  Skeptic announces $s_n \in \mathbb{R}$.
  Reality announces $y_n \in [-Y, Y]$.
  $\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - \mu_n)$.
END FOR.

$\mathcal{K}_n$: Skeptic's capital.

Algorithmic theory of randomness: consider a "computable" "strategy" for Skeptic that leads to the largest $\mathcal{K}_n$ (to within a constant factor).

Proposition 2 (game-theoretic SLLN) Skeptic has a strategy which guarantees that

- $\mathcal{K}_n$ is never negative
- either

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n) = 0$$

(μ_n are unbiased) or

$$\lim_{n\to\infty} \mathcal{K}_n = \infty.$$

A less ambitious goal: fix a law of probability and produce forecasts that are ideal as far as this law is concerned.

Computable form of Levin's idea

For any continuous strategy for Skeptic there exists a strategy for Forecaster that does not allow Skeptic's capital to grow.

So: for the usual laws of probability (stated game-theoretically) one can produce ideal forecasts. Many such laws in:

Glenn Shafer and Vladimir Vovk
Probability and finance: it's only a game
New York: Springer, 2001

Modified protocol:

$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots$:
  Reality announces $x_n \in \mathbf{X}$.
  Skeptic announces continuous $S_n : [-Y, Y] \to \mathbb{R}$.
  Forecaster announces $\mu_n \in [-Y, Y]$.
  Reality announces $y_n \in [-Y, Y]$.
  $\mathcal{K}_n := \mathcal{K}_{n-1} + S_n(\mu_n)(y_n - \mu_n)$.
END FOR.

Proposition 3 (Takemura) Forecaster has a strategy (defensive forecasting strategy) that ensures $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \geq \cdots$.

Proof

- choose μ_n so that $S_n(\mu_n) = 0$
- if the equation $S_n(\mu) = 0$ has no roots (in which case S_n never changes sign),

$$\mu_n := \begin{cases} Y & \text{if } S_n > 0 \\ -Y & \text{if } S_n < 0 \end{cases}$$

QED

Rediscovery of Levin's idea (broadly understood)

Foster and Vohra 1998: there exists a randomized forecasting strategy that produces asymptotically well-calibrated forecasts with probability one. Rich literature developing it, including Kakade and Foster 2004 (how to avoid randomization).

Proposition 3: Akimichi Takemura 2004.

Theorem 1 Let $\mathcal{F}$ be a Hilbert function space on $\mathbf{X}$. There exists an on-line prediction algorithm producing $\mu_n \in [-Y, Y]$ that are guaranteed to satisfy

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - D(x_n))^2 + 2Y\sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1}\,(\|D\|_{\mathcal{F}} + Y)\sqrt{N}$$

for all $N = 1, 2, \ldots$ and all $D \in \mathcal{F}$.

In fact, the optimal leading constant is $2Y\mathbf{c}_{\mathcal{F}}\|D\|_{\mathcal{F}}$. Theorem 1 gives it when $\|D\|_{\mathcal{F}} \gg Y$ and $\mathbf{c}_{\mathcal{F}} \gg 1$.

The algorithm: Predictor plays against Skeptic who enforces suitable laws of probability.

Lower bound (after Cesa-Bianchi/Long/Warmuth 1996)

Theorem 2 Suppose $\mathbf{X} = \mathbb{R}$. For any positive constant c there exists a universal Hilbert function space $\mathcal{F}$ on $\mathbf{X}$ with $\mathbf{c}_{\mathcal{F}} = c$ and a strategy for Reality satisfying the following property. For any $N = 1, 2, \ldots$, any positive constant $d \leq (Y/\mathbf{c}_{\mathcal{F}})\sqrt{N}$, and any on-line prediction algorithm, there exists $D \in \mathcal{F}$ such that $\|D\|_{\mathcal{F}} = d$ and

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \geq \sum_{n=1}^{N} (y_n - D(x_n))^2 + 2Y\mathbf{c}_{\mathcal{F}}\|D\|_{\mathcal{F}}\sqrt{N} - \mathbf{c}_{\mathcal{F}}^2\|D\|_{\mathcal{F}}^2 .$$

Idea of the proof of Theorem 1:

$$\sum_{n=1}^{N}(y_n - D(x_n))^2 = \sum_{n=1}^{N}(y_n - \mu_n)^2 + \sum_{n=1}^{N}(D(x_n) - \mu_n)^2 + 2\sum_{n=1}^{N}(\mu_n - D(x_n))(y_n - \mu_n).$$

The last sum is small if $\|D\|_{\mathcal{F}}$ is not too large: a kind of LLN.

New phenomenon:

$$\sum_{n=1}^{N}(y_n - D(x_n))^2 \approx \sum_{n=1}^{N}(y_n - \mu_n)^2 + \sum_{n=1}^{N}(D(x_n) - \mu_n)^2$$

if $\|D\|_{\mathcal{F}}$ is not too large ($\ll \sqrt{N}$).

Aggregating algorithm

A Bayes-type algorithm: all possible linear predictors are mixed with the weight given by a mixture of Gaussian prior distributions.

Γ: gamma function

U: Kummer's U function

Theorem 3 Let $\mathcal{F}$ be separable. There exists an on-line prediction algorithm producing $\mu_n \in [-Y, Y]$ that are guaranteed to satisfy, for all $N = 1, 2, \ldots$ and all $D \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - D(x_n))^2 - 2Y^2 \ln\left(\Gamma\left(\frac{N}{2}+1\right) U\left(\frac{N}{2}+1, 0, \frac{\mathbf{c}_{\mathcal{F}}^2 \|D\|_{\mathcal{F}}^2}{2Y^2}\right)\right)$$

$$\le \sum_{n=1}^{N} (y_n - D(x_n))^2 + 2Y \max\left(\mathbf{c}_{\mathcal{F}} \|D\|_{\mathcal{F}}, Y\delta N^{-1/2+\delta}\right) \sqrt{N+2} + \frac{3}{2} Y^2 \ln N + \frac{\mathbf{c}_{\mathcal{F}}^2 \|D\|_{\mathcal{F}}^2}{4} + O(Y^2),$$

where $\delta > 0$ is an arbitrarily small constant and the constant implicit in $O(Y^2)$ depends only on δ.

The bound of Theorem 3 is close to being stronger than that of Theorem 1 as $N \to \infty$:

$$2Y\sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1}\,(\|D\|_{\mathcal{F}} + Y) \mapsto 2Y\,\mathbf{c}_{\mathcal{F}}\,\|D\|_{\mathcal{F}}\,.$$

Its main disadvantage: the asymptotic character of its more explicit version. The version involving Γ and U is not intuitive, but it can be evaluated using standard libraries.

Gradient descent

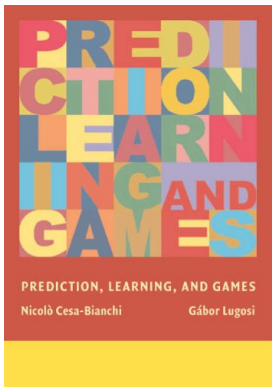
"Plug-in approach" (start from D_0, at each step n update D_{n-1} to D_n, and output $\mu_n := D_n(x_n)$): computationally efficient but tend to give weaker predictive performance guarantees than the "Bayesian" AA.

To apply GD: we need an *a priori* upper bound on the norm of the function; no universal consistency. Can be overcome (merging the algorithms corresponding to different bounds). Auer/Cesa-Bianchi/Gentile (2002): the idea is there but not pursued.

Full version of this paper with all proofs:

On-line regression competitive with reproducing kernel Hilbert spaces. Technical Report `arXiv:cs.LG/0511058` (version 2), `arXiv.org` e-Print archive, January 2006.

If you are interested in competitive on-line prediction in general:



Nicolò Cesa-Bianchi and Gábor Lugosi
Prediction, learning, and games
New York: Cambridge University Press, 2006

Neutral measures

Peter Gács. Uniform test of algorithmic randomness over a general space. Theoretical Computer Science, 341:91–137, 2005.

The original paper:

Leonid A. Levin. Uniform tests of randomness. Soviet Mathematics Doklady, 17:337–340, 1976.