Competitive on-line prediction as approximation problem

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

San Antonio, TX, March 7, 2007
From approximation theory to learning theory

Both approximation theory and learning theory: you are trying to approximate a target function with an approximant.

Learning-theory terminology: the allowed target functions form the target class; the allowed approximants form the hypothesis class.
Difference between learning theory and some strands of approximation theory:

• in the most classical approximation theory, we are given the whole target function $F$;

• in learning theory, we are given not the whole function but only some observations (such as a set of pairs $(x, F(x))$).

Approximation from a finite set of observations is also popular (e.g., Ming-Jun Lai’s talk on Monday).
Optional differences between approximation theory and various strands of learning theory

1. A learning theory can be **agnostic**: the target class is the class of all functions. Instead, we choose a **benchmark class**, the class of functions we would like to compete with.

2. A learning theory can be **improper** (the name invented by the enemies of this approach?): the hypothesis class is the class of all functions. In proper learning, the hypothesis class often coincides with the target class (or the benchmark class, in the agnostic case).
3. A learning theory can be **statistical**: a statistical model is postulated. Almost exclusively, this is the **iid model**: the observations are assumed to be independent and identically distributed. All other talks in this minisymposium adopt the iid model.

4. A learning theory can be **on-line**: the observations arrive one by one while we are making predictions. (Approximation theory is always **off-line**, or **batch**.)

Dichotomies 1 and 3 are related: they express our assumptions about Nature (the agent producing the observations).
Varieties of learning theory

We have

\[ 2 \times 2 \times 2 \times 2 = 16 \]

varieties of learning theory.

[Will we ever reach Good’s 46,656 varieties of Bayesians?]

My variety (in this talk): agnostic, improper, non-statistical, on-line.

Andrea Caponnetto’s talk (?): non-agnostic, proper, statistical, off-line.
Competitive on-line prediction protocol

Square-loss regression:

FOR $n = 1, 2, \ldots$:
   Nature announces $x_n \in \mathbf{X}$
   Predictor announces $\mu_n \in \mathbb{R}$
   Nature announces $y_n \in [-1, 1]$
END FOR

$x_n$: signal (the data relevant for predicting $y_n$, perhaps including some of the previous $y_{n-1}, y_{n-2}, \ldots$)

$y_n$: observation
Example

$y_n$: (high) temperature in San Antonio on day $n$

$x_n$: the data available when the prediction is made
Our prediction protocol is:

- Obviously on-line.
- Non-statistical: it is a perfect-information game.
- Improper: no restrictions on how $\mu_n$ are obtained from $x_n$. 
Predictor’s goal in competitive on-line prediction

Our prediction protocol is also agnostic, so we will choose a benchmark class $\mathcal{F}$ of prediction rules $F : X \rightarrow \mathbb{R}$. We want Predictor to achieve

$$\frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \frac{1}{N} \sum_{n=1}^{N} (y_n - F(x_n))^2$$

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

There is a strategy for Predictor that asymptotically dominates every continuous prediction rule:

**Theorem** Let $X$ be a metric compact. There exists a strategy for Predictor that guarantees

$$\limsup_{N \to \infty} \left( \frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 - \frac{1}{N} \sum_{n=1}^{N} (y_n - F(x_n))^2 \right) \leq 0$$

for each continuous prediction rule $F$. 
Inequalities instead of asymptotics

If $\mathcal{F}$ is a suitable benchmark class (Banach space, not as big as $C(X)$), Predictor can guarantee

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + g(\|F\|_\mathcal{F}, N)$$

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

The regret term $g(\|F\|_\mathcal{F}, N)$ must be $o(N)$ (and not grow too fast with $\|F\|_\mathcal{F}$).
Proof techniques

The strongest results can be obtained using two completely different techniques:

• Making use of the small metric entropy of the unit ball in $\mathcal{F}$ (if applicable): entropy method.

• Making use of the uniform convexity of the unit ball in $\mathcal{F}$ (if applicable): convexity method.

The best the second method can give is

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + O(N^{1/2})$$

(for Hilbert spaces $\mathcal{F}$).
Example 1: The entropy method for Besov spaces

Suppose $X$ is a subset of Euclidean space, $X \subseteq \mathbb{R}^m$, which is a minimally regular domain (bounded and coincides with the interior of its closure).

Every $B^{s}_{p,q}(X)$ with $s > m/p$ is compactly embedded in $C(X)$.

Using Edmunds and Triebel’s (1996) result and the entropy method:

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C_{X,s,p,q} \left( \|F\|_{B^{s}_{p,q}(X)} + 1 \right)^{\frac{m}{m+s}} N^{\frac{m}{m+s}}$$

for all $F \in B^{s}_{p,q}(X)$ from some $N$ on.
Example 2: The convexity method for Besov spaces

Let $p \in [2, \infty)$, $q \in [p/(p - 1), p]$ and $s \in (m/p, \infty)$. Using Cobos and Edmunds’s (1988) result and the convexity method:

\[
\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C_{X,s,p,q} \left( \|F\|_{B^{s}_{p,q}} + 1 \right) N^{1 - 1/p}
\]

for all $N = 1, 2, \ldots$ and all $F \in B^{s}_{p,q}(X)$.

The two methods give results that are not comparable: the metric entropy of the unit ball essentially depends only on $s$, whereas the modulus of convexity is governed by $p$. 
Example 3: Smooth RKHS

Let $\mathcal{F}$ be an RKHS with a $C^\infty$ reproducing kernel on $\mathbb{X}^2$ for a compact set $\mathbb{X} \subseteq \mathbb{R}^m$.

Cucker and Smale’s (2001) result and the entropy method give: for an arbitrarily small $\delta > 0$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + N\delta$$

for all $F \in \mathcal{F}$ from some $N$ on.
Example 4

Let $K$ be a simply connected continuum in $\mathbb{C}$ containing more than one point and $G$ be a connected open set such that $K \subseteq G \subseteq \mathbb{C}$.

$A^K_G$: the set of all complex-valued functions on $K$ that admit a bounded analytic continuation to $G$.

A classical result (Kolmogorov, Babenko, Erokhin) and the entropy method give:

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C_{G,K} \log^2 N$$

for all real-valued $F \in A^K_G$ and from some $N$ on.
Example 5

Let $h > 0$.

$A_h$: the vector space of all periodic period $2\pi$ complex-valued functions on the real line $\mathbb{R}$ that admit a bounded analytic continuation to the strip $\{z \in \mathbb{C} | |\text{Im } z| < h\}$

Using Vitushkin’s result and using the entropy method:

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + \frac{C}{h} \log^2 N$$

for all real-valued $F \in A_h$ and from some $N$ on

Both $A_h$ and $A_{E,1}^{[-1,1]}$ are dense, and so give rise to universal prediction strategies.
Open problems

• Merge the entropy and convexity methods.

• Prove lower bounds.