

Defensive prediction with expert advice

Vladimir Vovk

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

vovk@cs.rhul.ac.uk

ALT'2005 (Singapore)
10 October 2005, 16:40

Prediction with expert advice: we are given a pool of decision strategies (more generally, of experts) and our goal is to perform almost as well as the best strategy in the pool. No assumptions about the environment.

Defensive forecasting \mapsto a new proof technique in prediction with expert advice.

Prediction \mapsto **forecasting** (previous talk) or **decision making** (this talk).

Decision-making protocol:

$\text{Loss}_0 := 0.$

FOR $n = 1, 2, \dots$:

 Reality announces $x_n \in \mathbf{X}.$

 Decision Maker announces $\gamma_n \in \Gamma.$

 Reality announces $y_n \in \mathbf{Y}.$

$\text{Loss}_n := \text{Loss}_{n-1} + \lambda(y_n, \gamma_n).$

END FOR.

x_n : datum (all relevant information); y_n : observation; λ : the loss function.

Let $\Gamma = [0, 1]$ (decision space) and $\mathbf{Y} = \{0, 1\}$ (observation space).

The difference between the two protocols

- In the forecasting protocol, our goal is to produce perfect probabilities (probabilities one cannot gamble against).
- In the decision-making protocol, we are merely minimizing our loss.

Decision rule $D : \mathbf{X} \rightarrow [0, 1]$.

We want to compete against decision rules that are not too irregular with no assumptions about Reality. Let $\mathbf{X} = [0, 1]$ at first.

The same Sobolev-type norm $\|f\|_{\mathcal{S}}$ of $f : [0, 1] \rightarrow \mathbb{R}$:

$$\|f\|_{\mathcal{S}}^2 := \left(\int_0^1 f(t) dt \right)^2 + \int_0^1 (f'(t))^2 dt$$

(∞ if f is not absolutely continuous etc.).

Proposition Suppose $\mathbf{X} = [0, 1]$ and $\lambda(y, \gamma) = |y - \gamma|$. Decision Maker has a strategy that guarantees

$$\frac{1}{N} \sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \frac{1}{N} \sum_{n=1}^N \lambda(y_n, D(x_n)) + \frac{\|2D - 1\|_{\mathcal{S}} + 1}{\sqrt{N}}$$

for all N and D .

$(2D(x) - 1 \in [-1, 1])$ is “symmetrized” $D(x)$.)

Intuitively: this is about a “small” decision maker.

When is Decision Maker competitive with D ?

We have (remember $D(x) \in [0, 1]$):

$$\begin{aligned} \|2D - 1\|_{\mathcal{S}} &\leq \left| \int_0^1 (2D(t) - 1) dt \right| + 2 \sqrt{\int_0^1 (D'(t))^2 dt} \\ &\leq 1 + 2 \text{ "mean slope of } D\text{".} \end{aligned}$$

OK if the mean slope $\ll \sqrt{N}$.

Can be extended to: many other \mathbf{X} , norms, and loss functions.

Similar results

Cesa-Bianchi, Conconi, Gentile (2003): specific (and different) loss functions for Decision Maker (counting mistakes) and the decision rules (hinge loss).

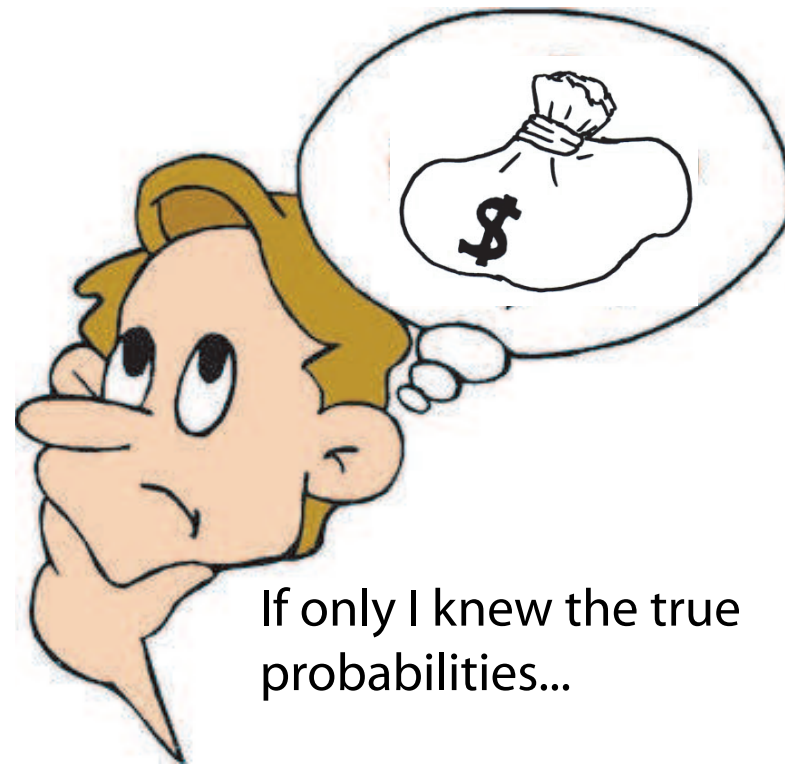
Long and Kinber (1994, 1997): there is a perfect decision rule.

Our approach inspired by: Foster and Vohra (1998); Fudenberg, Levine, Lehrer, Sandroni, Smorodinsky, Kakade, . . . ; no formal connections.

The rest of this talk: how to prove the proposition and many other results of this kind.

Research programme II (decision making)

- Choose a goal that could be achieved if you knew the true probabilities generating the observations.
- Construct a decision strategy provably achieving your goal.
- Isolate a continuous law of probability on which the proof depends. (Several laws can be merged into a single law.)
- Use defensive forecasting to get rid of the true probabilities.



The goal: (1) in terms of observables; (2) achievable regardless of what the true probabilities are.

The goal has to be **relative**.

Fix a choice function $G : [0, 1] \rightarrow \Gamma$:

$$G(p) \in \arg \min_{\gamma \in \Gamma} \lambda(p, \gamma),$$

where

$$\lambda(p, \gamma) := p\lambda(1, \gamma) + (1 - p)\lambda(0, \gamma).$$

For the square and log loss functions one can take $G(p) := p$.

The exposure of a decision: $\text{Exp}_\gamma := \lambda(1, \gamma) - \lambda(0, \gamma)$.

The exposure of a choice function G :

$$\text{Exp}_G(p) := \lambda(1, G(p)) - \lambda(0, G(p))$$

(assumed continuous; a modification of this definition also works for the absolute loss function).

The exposure of a decision rule $D : \mathbf{X} \rightarrow \Gamma$:

$$\text{Exp}_D(x) := \lambda(1, D(x)) - \lambda(0, D(x)).$$

Informal statement Suppose $\|\text{Exp}_G\|_{\mathcal{S}}$ is not too large. The decisions $\gamma_n := G(p_n)$ (“ELM principle”), with p_n output by ALN with a Sobolev kernel, satisfy

$$\frac{1}{N} \sum_{n=1}^N \lambda(y_n, \gamma_n) \lesssim \frac{1}{N} \sum_{n=1}^N \lambda(y_n, D(x_n))$$

for all N and all decision rules D with $\|\text{Exp}_D\|_{\mathcal{S}}$ not too large.

Proof Subtracting

$$\lambda(p, \gamma) = p\lambda(1, \gamma) + (1 - p)\lambda(0, \gamma)$$

from

$$\lambda(y, \gamma) = y\lambda(1, \gamma) + (1 - y)\lambda(0, \gamma)$$

gives

$$\begin{aligned}\lambda(y, \gamma) - \lambda(p, \gamma) &= (y - p)(\lambda(1, \gamma) - \lambda(0, \gamma)) \\ &= (y - p) \text{Exp}_\gamma.\end{aligned}$$

In conjunction with Theorem 3 (of the previous talk):

$$\begin{aligned}\sum_{n=1}^N \lambda(y_n, \gamma_n) &= \sum_{n=1}^N \lambda(y_n, G(p_n)) \\ &= \sum_{n=1}^N \lambda(p_n, G(p_n)) + \sum_{n=1}^N \left(\lambda(y_n, G(p_n)) - \lambda(p_n, G(p_n)) \right) \\ &= \sum_{n=1}^N \lambda(p_n, G(p_n)) + \sum_{n=1}^N (y_n - p_n) \text{Exp}_G(p_n) \\ &\approx \sum_{n=1}^N \lambda(p_n, G(p_n))\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{n=1}^N \lambda(p_n, D(x_n)) \\
&= \sum_{n=1}^N \lambda(y_n, D(x_n)) - \sum_{n=1}^N (\lambda(y_n, D(x_n)) - \lambda(p_n, D(x_n))) \\
&= \sum_{n=1}^N \lambda(y_n, D(x_n)) - \sum_{n=1}^N (y_n - p_n) \text{Exp}_D(x_n) \\
&\qquad\qquad\qquad \approx \sum_{n=1}^N \lambda(y_n, D(x_n)).
\end{aligned}$$

Summary of the proof technique: to show that the actual loss of our decision strategy does not exceed the actual loss of a decision rule D by much, we notice that

- the actual loss $\sum_{n=1}^N \lambda(y_n, G(p_n))$ of our decision strategy is approximately equal, by Theorem 3, to the (one-step-ahead conditional) expected loss $\sum_{n=1}^N \lambda(p_n, G(p_n))$ of our strategy;
- since we used the Expected Loss Minimization principle, the expected loss of our strategy does not exceed the expected loss of D ;
- the expected loss of D is approximately equal to its actual loss (again by Theorem 3).

Theorem 4 (special cases) Suppose $\lambda(y, \gamma) = (y - \gamma)^2$.

Decision Maker has a strategy that guarantees

$$\sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \sum_{n=1}^N \lambda(y_n, D(x_n)) + \frac{1}{\sqrt{3}} (\|2D - 1\|_{\mathcal{S}} + 1) \sqrt{N}$$

for all N and D .

Suppose $\lambda(y, \gamma) = |y - \gamma|$. Decision Maker has a strategy that guarantees

$$\sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \sum_{n=1}^N \lambda(y_n, D(x_n)) + \sqrt{\frac{7}{12}} (\|2D - 1\|_{\mathcal{S}} + 1) \sqrt{N}$$

for all N and D .

Suppose

$$\lambda(y, \gamma) = -y \ln \gamma - (1 - y) \ln(1 - \gamma).$$

Decision Maker has a strategy that guarantees

$$\sum_{n=1}^N \lambda(y_n, \gamma_n) \leq \sum_{n=1}^N \lambda(y_n, D(x_n)) + 0.75 \left(\left\| \ln \frac{D}{1-D} \right\|_{\mathcal{S}} + 1 \right) \sqrt{N}$$

for all N and D .

General theorem: functions with a “convex loss” (if unbounded, the tails must decay faster than $1/t$; in the log loss game, they decay exponentially fast).

Can be extended to non-convex loss functions and loss functions depending on several future observations (work in progress).

Further details

Game-theoretic probability:

Glenn Shafer and Vladimir Vovk, [Probability and finance: it's only a game](#), New York: Wiley, 2001

Defensive forecasting:

<http://www.probabilityandfinance.com>, Working Papers 13 (previous talk), 14 (this talk), 10 (Akimichi's talk).