

Competitive on-line prediction: minimax Bayesian approach vs defensive forecasting

Vladimir Vovk

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

MSR Cambridge
19 March, 2008



Outline

- 1 Prediction with expert advice
- 2 Minimax Bayesian approach
- 3 Large benchmark classes
- 4 Defensive forecasting



Outline

- 1 Prediction with expert advice
 - Weighted Majority Algorithm
 - Bayesian merging
 - Strong Aggregating Algorithm
- 2 Minimax Bayesian approach
- 3 Large benchmark classes
- 4 Defensive forecasting



Model-based approaches

Common belief: to make strong conclusions we need (strong) **statistical assumptions**.

The most popular assumption: i.i.d. data (independent identically distributed examples).

“Statistical learning theory” (Vapnik & Chervonenkis) is the dominant paradigm of learning under the i.i.d. assumption.

Competitive on-line prediction: no statistical assumptions.
Instead: a benchmark class of strategies we want to compete with. **Prediction with expert advice**: competing with free agents instead of prediction strategies.



Protocol of prediction with expert advice

The basic setting of competitive on-line prediction.

Prediction with K experts' advice

$$L_0 := 0, \quad L_0^k := 0, \quad k = 1, \dots, K$$

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

Expert k outputs $\gamma_n^k \in \Gamma$, $k = 1, \dots, K$

Learner outputs $\gamma_n \in \Gamma$

Nature outputs $y_n \in \mathbf{Y}$

$$L_n := L_{n-1} + \lambda(y_n, \gamma_n)$$

$$L_n^k := L_{n-1}^k + \lambda(y_n, \gamma_n^k), \quad k = 1, \dots, K$$

END FOR.

λ : **loss function**. Triple $(\mathbf{Y}, \Gamma, \lambda)$: **prediction game**.



Halving algorithm

The simplest is the **simple prediction game**: $\mathbf{Y} = \Gamma = \{0, 1\}$,

$$\lambda(y, \gamma) := \begin{cases} 0 & \text{if } y = \gamma \\ 1 & \text{otherwise.} \end{cases}$$

Suppose we know that one of the experts is perfect.

Theorem

There is a strategy for Learner that guarantees

$$\exists k : L_n^k = 0 \implies L_n \leq \log K.$$



Proof and idea of extension

Very easy to prove: Learner follows the majority of the experts who have not been eliminated so far. If he makes a mistake: at least half of the remaining experts is eliminated.

What can we do if none of the experts is perfect? Idea:

- start with equal weights for all experts;
- if an expert makes a mistake, multiply his weight by $\beta \in [0, 1)$ ($\eta := -\ln \beta$ is usually called the **learning rate**);
- predict with the weighted majority of the experts.

Halving Algorithm: $\eta = \infty$.



Weighted majority algorithm: result

Theorem

For any $\beta \in [0, 1)$, there is a strategy for Learner that guarantees

$$L_n \leq \frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}} L_n^k + \frac{1}{\ln \frac{2}{1+\beta}} \ln K$$

for all k and n .

Performance guarantee: very different from statistical learning theory.

Can replace $\ln K$ by $\ln \frac{1}{w_k}$, where $w_k \geq 0$ is the weight assigned to the k th expert ($\sum_{k=1}^K w_k = 1$).



Log loss game

The **log loss game**:

$$\mathbf{Y} = \{1, \dots, J\}$$

$$\Gamma = \left\{ (\gamma[1], \dots, \gamma[J]) \in [0, \infty)^J : \sum_{j=1}^J \gamma[j] = 1 \right\}$$

$$\lambda(\mathbf{y}, \gamma) = -\ln \gamma[\mathbf{y}].$$

Intuition: minus log likelihood, or the code length.



Bayesian prediction

Suppose: each expert follows a strategy (for simplicity, with $\gamma[j]$ always positive). In the log loss game:

strategy = probability measure P

loss $L_n =$ minus log likelihood ($-\ln P(y_1, \dots, y_n)$)

We know how to merge probability measures (Bayes):

$$P := \sum_{k=1}^K w_k P_k.$$

This way we will ensure

$$P \geq w_k P_k$$

$$L_n \leq L_n^k + \ln \frac{1}{w_k}$$



Bayesian statistics reinterpreted

Our framework: standard Bayesian. “Statistical model”:
 $\{P_k : k = 1, \dots, K\}$. “Prior probabilities”: w_k .

But it does not reflect our beliefs: we compete against the P_k ,
and the w_k reflect the head start we require on the k th expert.

The assumption that the experts follow strategies is
superfluous. At each step we can predict as the weighted
average of the experts’ predictions, with the weights constantly
updated using the Bayes formula.



General loss functions

Strong Aggregating Algorithm (SAA) and **Defensive Forecasting (DF)**: contain all prediction algorithms considered so far as special cases.

For many prediction games (in particular, all prediction games with bounded loss function) SAA guarantees

$$L_n \leq c(\eta)L_n^k + a(\eta) \ln K,$$

where $\eta \in (0, \infty]$ (the learning rate).



Optimality result

Theorem

If Learner has a strategy always achieving

$$L_n \leq cL_n^k + a \ln K$$

for any pool of experts, then $a(\eta) \leq a$ and $c(\eta) \leq c$ for some η .

Caveat: if you look at the proof, it is essential that K should be large.



Mixable games

An important class of prediction games: **mixable** games, for which Learner can guarantee

$$L_n \leq \min_k L_n^k + C_K.$$

Notice: C_K can be automatically replaced by $C \log K$.

Easy corollary from the previous result:

Corollary (Optimality result)

If Learner has a strategy always achieving

$$L_n \leq L_n^k + c$$

for any pool of experts of size K , then the SAA can also achieve this. (K is fixed.)



Important examples of mixable games

The **square loss game**:

$$\mathbf{Y} = \Gamma = [-Y, Y], \quad \lambda(y, \gamma) = (y - \gamma)^2.$$

The **Brier game**:

$$\mathbf{Y} = \{1, \dots, J\}, \quad \Gamma = \left\{ (\gamma[1], \dots, \gamma[J]) : \sum_{j=1}^J \gamma[j] = 1 \right\},$$

$$\lambda(y, \gamma) = \sum_{j=1}^J (1_{\{j=y\}} - \gamma[j])^2.$$



Performance guarantees

Theorem

The SAA guarantees

$$L_n \leq L_n^k + 2Y^2 \ln K \quad \forall k, n$$

in the square loss game and guarantees

$$L_n \leq L_n^k + \ln K \quad \forall k, n$$

in the Brier game. [The constants are optimal.]



Competing with bookmakers

Next slide: results of an experiment with 10,087 tennis matches (2004, 2005, 2006, and 2007 tournaments, including, e.g., Australian Open, French Open, Wimbledon, and US Open). There are four experts (bookmakers): Bet365, Centrebet, Expekt, and Pinnacle Sports.

Brier loss function.



Prediction results of tennis matches



Outline

- 1 Prediction with expert advice
- 2 **Minimax Bayesian approach**
 - Introduction to SAA
 - APA
 - SAA
- 3 Large benchmark classes
- 4 Defensive forecasting



Bayes-type aggregation

The SAA and its variants. Potential disadvantage: relative computational inefficiency.

Idea:

- maintain weights for the experts (or prediction strategies), at each step slashing the weights of experts who suffer large losses (“exponential weights”);
- at each step “merge” the experts’ predictions taking into account their weights.



SAA: Introduction

For simplicity: finitely many (K) experts. The initial weights:
 w_1, \dots, w_K .

Generalized prediction is a function $g : \mathbf{Y} \rightarrow \mathbb{R}$. The generalized prediction corresponding to $\gamma \in \Gamma$:

$$g(y) := \lambda(y, \gamma)$$

(**permitted prediction**).

$\eta > 0$: **learning rate**

$\beta := e^{-\eta}$: **exponential learning rate**



Aggregating Pseudo-Algorithm

Aggregating Pseudo-Algorithm (APA)

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

output the generalized prediction $g_n(y) := \log_{\beta} \sum_{k=1}^K \beta^{\lambda(y, \gamma_n^k)} w_k$

update the weights $w_k := w_k \beta^{\lambda(y_n, \gamma_n^k)}$, $k = 1, \dots, K$

normalize the weights $w_k := w_k / \sum_{j=1}^K w_j$, $k = 1, \dots, K$

END FOR.



APA's performance guarantee

APA achieves:

$$L_n = \log_{\beta} \sum_{k=1}^K \beta^{L_n^k} w_k$$

In particular:

$$L_n \leq L_n^k + \log_{\beta} w_k$$

Or:

$$L_n \leq L_n^k + \frac{1}{\eta} \ln \frac{1}{w_k}$$



From generalized to permitted predictions

However: a mixture of permitted predictions may not be a permitted prediction.

Sometimes it is: e.g., in the log loss game.



Mixable games

Even if it is not, it is possible that any mixture is dominated by a permitted prediction:

$$\forall \mathbf{w}_1, \dots, \mathbf{w}_K \quad \forall \gamma^1, \dots, \gamma^K \quad \exists \gamma \quad \forall \mathbf{y} :$$

$$\lambda(\mathbf{y}, \gamma) \leq \log_{\beta} \sum_{k=1}^K \beta^{\lambda(\mathbf{y}, \gamma^k)} \mathbf{w}_k$$

(and we can take 2 instead of K). Such games: **η -mixable** (recall: $\eta := -\ln \beta$). **Mixable** = η -mixable for some η (**fact**: this is equivalent to the old definition).

Examples:

- the square loss game is mixable ($\eta \leq 1/(2Y^2)$)
- the Brier game is mixable ($\eta \leq 1$)



Substitution functions

If the game is not η -mixable but $\lambda \geq 0$: find the smallest possible $c(\eta)$ such that

$$\forall w_1, \dots, w_K \quad \forall \gamma^1, \dots, \gamma^K \quad \exists \gamma \quad \forall y :$$

$$\lambda(y, \gamma) \leq c(\eta) \log_{\beta} \sum_{k=1}^K \beta^{\lambda(y, \gamma^k)} w_k$$

Substitution function: the corresponding $\Sigma : g \mapsto \gamma$.

SAA: $\gamma_n := \Sigma(g_n)$.

In this case:

$$L_n \leq c(\eta) L_n^k + \frac{c(\eta)}{\eta} \ln \frac{1}{w_k}$$

This is done for the simple prediction game.



Outline

- 1 Prediction with expert advice
- 2 Minimax Bayesian approach
- 3 Large benchmark classes**
 - Tracking
 - Linear decision rules
 - Function classes
- 4 Defensive forecasting



Tracking good experts

Classes of experts can be huge. For example: **superexperts**, predicting as a sequence of experts

$E = (e_1, \dots, e_n) \in \{1, \dots, K\}^n$ with a small number of **switches** $e_i \neq e_{i+1}$. Let L_n^E : loss of the superexpert over the first n steps.

Theorem

In the square loss game, the SAA (applied to the superexperts with a suitable prior) guarantees

$$L_n \leq L_n^E + 2Y^2 \left(\ln K + s \ln(K-1) + (s + 0.01) \ln n + 5 \right)$$

for all n and E , where s is the number of switches in E .

Can be done for many other loss functions and for many other kinds of superexperts.



Competing with linear decision rules

Square loss game. Add the objects $x_n \in \mathbb{R}^d$ to the protocol;
 $\|x_n\|_\infty \leq X$. “Experts”: $\theta \in \mathbb{R}^d$; predict $\theta \cdot x_n$ at step n .

Theorem

For each $a > 0$, the SAA (with a Gaussian prior) guarantees

$$L_n \leq \inf_{\theta} \left(L_n^{\theta} + a \|\theta\|_2^2 \right) + dY^2 \ln \left(\frac{nX^2}{a} + 1 \right)$$

for all n and θ .

The algorithm: almost Ridge Regression (the difference: add $(x_n, 0)$ to the data set when predicting y_n). The constant dY^2 is optimal (and cannot be achieved by RR).



Competing with function classes I

This can be generalized to much bigger classes:
 infinite-dimensional Hilbert and Banach spaces.

Regret term grows faster than $O(\ln n)$ (but much slower than n).

For example: Objects: $x_n \in \mathbf{X} := [0, 1]$. For $s \in (0, 1)$, the norm in $\mathcal{C}^s(\mathbf{X})$ is defined by

$$\|f\|_{\mathcal{C}^s(\mathbf{X})} = \max \left(\sup_{x \in \mathbf{X}} |f(x)|, \sup_{x, y \in \mathbf{X}: x \neq y} \left| \frac{f(x) - f(y)}{|x - y|^s} \right| \right)$$

(and the space $\mathcal{C}^s(\mathbf{X})$ consists of the functions f with finite norm).



Competing with function classes II

Prediction rule $f \in \mathcal{C}^s(\mathbf{X})$ predicts y_n with $f(x_n)$.

Theorem

Suppose $s \leq 1/2$ and fix an arbitrarily small $\epsilon > 0$. There exists a constant $C_{s,\epsilon} > 0$ such that Learner can guarantee

$$L_n \leq L_n^f + Y C_{s,\epsilon} \left(\|f\|_{\mathcal{C}^s(\mathbf{X})} + Y \right) n^{1-s+\epsilon}$$

for all n and all $f \in \mathcal{C}^s(\mathbf{X})$. [Method: DF]



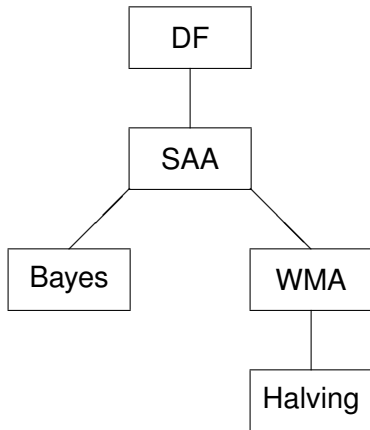
Outline

- 1 Prediction with expert advice
- 2 Minimax Bayesian approach
- 3 Large benchmark classes
- 4 **Defensive forecasting**
 - How DF works
 - Flexibility of DF



Defensive forecasting (DF)

This is a “dual” method to the SAA. But in simple cases it **becomes** the SAA.



Proper loss functions

Suppose that $\Gamma = [0, 1]$, $\mathbf{Y} = \{0, 1\}$, and $\lambda(\gamma, 0)$ and $\lambda(\gamma, 1)$ are continuous in $\gamma \in [0, 1]$.

The loss function λ is a **proper** if, for any $\pi, \pi' \in [0, 1]$,

$$\pi\lambda(1, \pi) + (1 - \pi)\lambda(0, \pi) \leq \pi\lambda(1, \pi') + (1 - \pi)\lambda(0, \pi').$$

(Encouraging honesty. Standard names for predictions.)

The log loss and square loss functions are proper.



Basic supermartingales

An alternative method to prove the SAA bound: for each $k = 1, \dots, K$, define

$$Q^k(\gamma_1^\bullet, \pi_1, y_1, \dots, \gamma_N^\bullet, \pi_N, y_N) := \prod_{n=1}^N e^{\eta(\lambda(y_n, \pi_n) - \lambda(y_n, \gamma_n^k))}.$$

If λ is η -mixable, Q^k is a (game-theoretic) **supermartingale**:

$$\begin{aligned} & \pi_N Q^k(\gamma_1^\bullet, \pi_1, y_1, \dots, \gamma_{N-1}^\bullet, \pi_{N-1}, y_{N-1}, \gamma_N^\bullet, \pi_N, 1) \\ & + (1 - \pi_N) Q^k(\gamma_1^\bullet, \pi_1, y_1, \dots, \gamma_{N-1}^\bullet, \pi_{N-1}, y_{N-1}, \gamma_N^\bullet, \pi_N, 0) \\ & \leq Q^k(\gamma_1^\bullet, \pi_1, y_1, \dots, \gamma_{N-1}^\bullet, \pi_{N-1}, y_{N-1}). \end{aligned}$$

Set

$$Q := \frac{1}{K} \sum_{k=1}^K Q^k.$$



Intuition behind non-negative supermartingales starting from 1

There is another player (“Sceptic”) who tries to prove Forecaster wrong. A non-negative supermartingale starting from 1 = the capital process of some strategy for Sceptic.

SAA (special case): we mix various strategies for Forecaster.
DF: we mix various strategies for Sceptic.

General phenomenon: duality often helps in machine learning (such as “generalized portrait” \rightarrow SVM). SAA \rightarrow DF might be another manifestation (on a smaller scale).



Non-increase

Levin, Takemura: for each supermartingale S there is a way to choose π_n such that S never increases, regardless of y_n . The defensive forecasting algorithm produces predictions π_n such that the sequence

$$Q_n := Q(\gamma_1^\bullet, \pi_1, y_1, \dots, \gamma_n^\bullet, \pi_n, y_n)$$

is non-increasing.

Since $Q_0 = 1$, we have: $Q \leq 1$ on the realized path; $Q^k \leq K$ on the realized path;

$$\sum_{n=1}^N \eta(\lambda(y_n, \pi_n) - \lambda(y_n, \gamma_n^k)) \leq \ln K.$$



Flexibility of defensive forecasting

So far: we have only reproduced the SAA result. Why is DF good?

For each mixable loss function λ the **optimal learning rate** for λ is defined to be the supremum of η such that λ is η -mixable.

Lemma: λ is η -mixable for the optimal learning rate η .

Let \mathcal{L} be the set of all mixable and proper loss functions with optimal learning rate 1. “Normalized loss functions using the same language.”



Prediction with expert evaluators' advice

Prediction with K experts' advice

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

Expert k outputs $\lambda_n^k \in \mathcal{L}$, $c_n^k \in [0, 1]$, and $\gamma_n^k \in [0, 1]$, $k = 1, \dots, K$

Learner outputs $\gamma_n \in [0, 1]$

Nature outputs $y_n \in \{0, 1\}$

END FOR.

λ_n^k : loss function chosen by the expert. c_n^k : the expert's confidence in his prediction. $c_n^k \in \{0, 1\}$: the framework of "sleeping experts".



Theorem

Theorem

Learner has a strategy (DF) that guarantees, for all N and for all $k = 1, \dots, K$, that

$$\sum_{n=1}^N c_n^k \lambda_n^k(y_n, \pi_n) \leq \sum_{n=1}^N c_n^k \lambda_n^k(y_n, \gamma_n^k) + \ln K.$$

Special cases: **multiobjective prediction with expert advice**; the standard protocol of **sleeping experts**. Easy to extend to **second-guessing experts**.



Prediction with K constant expert evaluators' advice

Suppose each expert always chooses the same $\lambda_n^k = \lambda^k$, mixable and proper but not required to have optimal learning rate 1.

Corollary

Learner has a strategy (DF) that guarantees, for all N and for all $k = 1, \dots, K$,

$$L_N^{(k)} \leq L_N^k + \frac{\ln K}{\eta^k},$$

where $L^{(k)}$ is Learner's cumulative loss relative to λ^k , L^k is Expert k 's cumulative loss relative to λ^k , and η^k is the optimal learning rate of λ^k .



Multiobjective prediction with expert advice

Now we have N experts and M loss functions $\lambda^1, \dots, \lambda^M$.

Multiobjective prediction with expert advice

$$L_0^{(m)} := 0, \quad m = 1, \dots, M$$

$$L_0^{k,m} := 0, \quad k = 1, \dots, K, m = 1, \dots, M$$

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

Expert k outputs $\gamma_n^k \in [0, 1]$, $k = 1, \dots, K$

Learner outputs $\gamma_n \in [0, 1]$

Nature outputs $y_n \in \{0, 1\}$

$$L_n^{(m)} := L_{n-1}^{(m)} + \lambda^m(y_n, \gamma_n), \quad m = 1, \dots, M$$

$$L_n^{k,m} := L_{n-1}^{k,m} + \lambda^m(y_n, \gamma_n^k), \quad k = 1, \dots, K, m = 1, \dots, M$$

END FOR.



Corollary

Corollary

Suppose that every λ^m is an η^m -mixable proper loss function, for some $\eta^m > 0$, $m = 1, \dots, M$. DF guarantees that

$$L_n^{(m)} \leq L_n^{k,m} + \frac{\ln MK}{\eta^m}$$

for all n , all $k = 1, \dots, K$, and all $m = 1, \dots, M$.

For example, suppose we are competing with K experts under the log loss and square loss functions simultaneously. DF ensures that the regret with respect to the logarithmic loss function is bounded by $\ln(2K) < \ln K + 0.7$, and the regret with respect to the square loss function by $0.5 \ln(2K) < 0.5 \ln K + 0.4$.



Prediction with specialist experts' advice

Let there be one loss function λ , mixable but not necessarily proper. Let a be any object that does not belong to $[0, 1]$ (an expert's decision to abstain).

Prediction with specialist experts' advice

$$L_0^{(k)} := 0, \quad k = 1, \dots, K$$

$$L_0^k := 0, \quad k = 1, \dots, K$$

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

Expert k outputs $\gamma_n^k \in ([0, 1] \cup \{a\})$, $k = 1, \dots, K$

Learner outputs $\gamma_n \in [0, 1]$

Nature outputs $y_n \in \{0, 1\}$

If $\gamma_n^k \neq a$, $L_n^{(k)} := L_{n-1}^{(k)} + \lambda(y_n, \gamma_n^k)$, $k = 1, \dots, K$

If $\gamma_n^k \neq a$, $L_n^k := L_{n-1}^k + \lambda(y_n, \gamma_n^k)$, $k = 1, \dots, K$

END FOR



Corollary



Corollary

Learner has a strategy (DF) that guarantees, for all n and all $k = 1, \dots, K$,

$$L_n^{(k)} \leq L_n^k + \frac{\ln K}{\eta}.$$



Further reading

-  Nicolò Cesa-Bianchi and Gábor Lugosi.
Prediction, Learning, and Games.
Cambridge: Cambridge University Press, 2006.
-  Alexey Chernov and Vladimir Vovk.
Prediction with expert evaluators' advice.
arXiv technical report.
-  On-line prediction wiki.
<http://www.onlineprediction.net>.



Summary

This talk: results by many people (Chernov, Dawid, Freund, Helmbold, Kalnishkan, Littlestone, Schapire, Shafer, Takemura, Vovk, Warmuth, Watkins, Zhdanov, . . .).

- You can develop a theory of prediction without any statistical assumptions.
- The Bayes scheme can be generalized using defensive forecasting.

