Multiobjective Prediction with Expert Advice

Alexey Chernov

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

GTP Workshop, June 2010
Example: Prediction of Sport Match Outcome

V. Vovk, F. Zhdanov. Predictions with Expert Advice for Brier Game. ICML’08

Bookmakers data:
4 bookmakers, odds for $\sim 10000$ tennis matches (2 outcomes)
8 bookmakers, odds for $\sim 9000$ football matches (3 outcomes)

Odds $a_i$ can be transformed to probabilities $\text{Prob}[i]$:

$$\text{Prob}[i] = \frac{1/a_i}{\sum_j 1/a_j}$$

The loss is measured by the square (Brier) loss function.

Learner’s strategy is the Aggregating Algorithm.
Theoretical bounds
Learner
Bookmakers

Graph of the negative regret $\text{Loss}_{\varepsilon_k}(T) - \text{Loss}(T)$, 4 Experts

Learner is the AA for the square loss
Theoretical bounds
Learner
Bookmakers

Graph of the negative regret \( \text{Loss}_{\epsilon_k}(T) - \text{Loss}(T) \), 4 Experts
Learner is the AA for the log loss (Bayes mixture)
Tennis Predictions: “Wrong” Losses

Graphs of the negative regret $\text{Loss}_{\epsilon_k}(T) - \text{Loss}(T)$

Learner optimizes for a “wrong” loss function
For the game with 2 outcomes, one can construct a sequence of predictions of 2 Experts and a sequence of outcomes with the following property. If Learner’s predictions are generated by the Aggregating Algorithm for the log loss then for almost all $T$:

$$\text{Loss}(T) \geq \text{Loss}_{E_1}(T) + \frac{T}{10},$$

where $\text{Loss}(T)$ and $\text{Loss}_{E_1}(T)$ are the square losses of Learner and Expert 1.

A similar statement holds for the Aggregating Algorithm for the square loss evaluated by the log loss.
New Settings

Experts: $\gamma_t^{(1)}, \ldots, \gamma_t^{(k)}$
Learner: $\gamma_t$
Reality: $\omega_t$

Many loss functions

$\text{Loss}_{E_k}^{(m)}(T) = \sum_{t=1}^{T} \lambda^{(m)}(\gamma_t^{(k)}, \omega_t)$

$\text{Loss}^{(m)}(T) = \sum_{t=1}^{T} \lambda^{(m)}(\gamma_t, \omega_t)$
New Settings

Experts: $\gamma_t^{(1)}, \ldots, \gamma_t^{(k)}$
Learner: $\gamma_t$
Reality: $\omega_t$

Many loss functions

\[ \text{Loss}^{(m)}_E(T) = \sum_{t=1}^{T} \lambda^{(m)}(\gamma_t^{(k)}, \omega_t) \]

\[ \text{Loss}^{(m)}(T) = \sum_{t=1}^{T} \lambda^{(m)}(\gamma_t, \omega_t) \]

Expert Evaluator’s advice

\[ \text{Loss}^{(k)}_E(T) = \sum_{t=1}^{T} \lambda^{(k)}(\gamma_t^{(k)}, \omega_t) \]

\[ \text{Loss}^{(k)}(T) = \sum_{t=1}^{T} \lambda^{(k)}(\gamma_t, \omega_t) \]
Bound for New Settings

**Theorem**

If $\lambda^{(k)}$ are $\eta^{(k)}$-mixable proper loss functions, $k = 1, \ldots, K$, Learner has a strategy (e.g. the Defensive Forecasting algorithm) that guarantees, for all $T$ and for all $k$, that

$$\text{Loss}^{(k)}(T) \leq \text{Loss}_{\epsilon_k}^{(k)}(T) + \frac{1}{\eta^{(k)}} \ln K.$$  

**Corollary**

If $\lambda^{(m)}$ are $\eta^{(m)}$-mixable proper loss functions, $m = 1, \ldots, M$, Learner has a strategy that guarantees, for all $T$, for all $k$ and for all $m$, that

$$\text{Loss}^{(m)}(T) \leq \text{Loss}_{\epsilon_k}^{(m)}(T) + \frac{1}{\eta^{(m)}} (\ln K + \ln M).$$
The graphs show the negative regret $\text{Loss}_k^{(m)}(T) - \text{Loss}^{(m)}(T)$ for both square loss and log loss. The theoretical bounds are compared with the learner's performance and the bookmakers' predictions.

Learner optimizes for both loss functions, using the DF algorithm.
Defensive Forecasting Algorithm

\[
\exists \pi \ \forall \omega \ \sum_{k=1}^{K} p_{t-1}^{(k)} e^{\eta(\lambda(\pi, \omega) - \lambda(\pi^{(k)}_t, \omega))} \leq 1 ,
\]

where \( p_{t-1}^{(k)} = p_0^{(k)} e^{\eta(\text{Loss}(t-1) - \text{Loss}e_k(t-1))} \)

To get this (from Levin's Lemma) we need that \( \lambda(\pi, \omega) \) is continuous and for all \( \pi, \pi' \)

\[
E_{\pi} e^{\eta(\lambda(\pi, \cdot) - \lambda(\pi', \cdot))} = \sum_{\omega \in \Omega} \pi(\omega) e^{\eta(\lambda(\pi, \omega) - \lambda(\pi', \omega))} \leq 1
\]
Defensive Forecasting Algorithm

\[ \exists \pi \ \forall \omega \ \sum_{k=1}^{K} p_{t-1}^{(k)} e^{\eta^{(k)}(\lambda^{(k)}(\pi,\omega) - \lambda^{(k)}(\pi^{(k)}_t,\omega))} \leq 1, \]

where \( p_{t-1}^{(k)} = p_0^{(k)} e^{\eta^{(k)}(\text{Loss}^{(k)}(t-1) - \text{Loss}_{\epsilon}^{(k)}(t-1))} \)

To get this (from Levin’s Lemma) we need that \( \lambda^{(k)}(\pi, \omega) \) is continuous and for all \( \pi, \pi' \)

\[ \mathbb{E}_{\pi} e^{\eta^{(k)}(\lambda^{(k)}(\pi, \cdot) - \lambda^{(k)}(\pi', \cdot))} = \sum_{\omega \in \Omega} \pi(\omega) e^{\eta^{(k)}(\lambda^{(k)}(\pi, \omega) - \lambda^{(k)}(\pi', \omega))} \leq 1 \]
The DFA and the AA

\[ \lambda \text{ is continuous and } \forall \pi, \pi' \ E_{\pi} e^{\eta(\lambda(\pi, \cdot) - \lambda(\pi', \cdot))} \leq 1 \quad \Rightarrow \quad \lambda \text{ is } \eta\text{-mixable} \]

\[ \lambda \text{ is } \eta\text{-mixable and } ? \quad \Rightarrow \quad \forall \pi, \pi' \ E_{\pi} e^{\eta(\lambda(\pi, \cdot) - \lambda(\pi', \cdot))} \leq 1 \]
The DFA and the AA

\[ \lambda \text{ is continuous and } \forall \pi, \pi' \ E_\pi e_\eta(\lambda(\pi, \cdot) - \lambda(\pi', \cdot)) \leq 1 \quad \Rightarrow \quad \lambda \text{ is } \eta\text{-mixable} \]

\[ \forall \pi, \pi' \ E_\pi e_\eta(\lambda(\pi, \cdot) - \lambda(\pi', \cdot)) \leq 1 \quad \Rightarrow \quad \lambda \text{ is proper} \]

\[ \lambda \text{ is } \eta\text{-mixable and proper} \quad \Rightarrow \quad \forall \pi, \pi' \ E_\pi e_\eta(\lambda(\pi, \cdot) - \lambda(\pi', \cdot)) \leq 1 \]
Proper Loss Functions

$\lambda$ is proper if for any $\pi, \pi' \in \mathcal{P}(\Omega)$

$$E_{\pi} \lambda(\pi, \cdot) \leq E_{\pi} \lambda(\pi', \cdot)$$

If $\omega \sim \pi$ then $E_{\pi} \lambda(\pi', \omega)$ is the expected loss for prediction $\pi'$.

The expected loss is minimal for the true distribution $\Rightarrow$ the forecaster is encouraged to give the true probabilities.

The square loss and the log loss are proper.
Example: Hellinger Loss

\[ \lambda^{\text{Hellinger}}(\gamma, \omega) = \frac{1}{2} \sum_{j=1}^{r} \left( \sqrt{\gamma(j)} - \sqrt{\mathbb{I}_{\{\omega=j\}}} \right)^2 \]

The Hellinger loss is \( \sqrt{2} \)-mixable

The Hellinger loss is not proper
Proper Mixable Loss Functions

Each mixable loss function $\lambda(\gamma, \omega)$ has a proper analogue $\lambda^{\text{proper}}(\pi, \omega)$ such that

1. $\forall \pi \exists \gamma \forall \omega \quad \lambda^{\text{proper}}(\pi, \omega) = \lambda(\gamma, \omega)$

2. $\forall \pi \forall \gamma \quad \mathbb{E}_{\pi} \lambda^{\text{proper}}(\pi, \cdot) \leq \mathbb{E}_{\pi} \lambda(\gamma, \cdot)$

For the Hellinger loss, the proper analogue is the spherical loss $\lambda^{\text{spherical}}(\pi, \omega)$

$$\lambda^{\text{spherical}}(\pi, \omega) = 1 - \sqrt{\sum_{j=1}^{r} \left(\frac{\pi(j)}{\pi_{\pi}(j)}\right)^2}$$

$$\lambda^{\text{spherical}}(\pi, \omega) = \lambda^{\text{Hellinger}}(\gamma, \omega)$$

for $\gamma(\omega) = (\pi(\omega))^2 P_{r} j=1 \left(\frac{\pi(j)}{\pi_{\pi}(j)}\right)^2$
Proper Mixable Loss Functions

Each mixable loss function $\lambda(\gamma, \omega)$ has a proper analogue $\lambda^{\text{proper}}(\pi, \omega)$ such that

1. $\forall \pi \exists \gamma \forall \omega \; \lambda^{\text{proper}}(\pi, \omega) = \lambda(\gamma, \omega)$
2. $\forall \pi \forall \gamma \; \mathbb{E}_{\pi} \lambda^{\text{proper}}(\pi, \cdot) \leq \mathbb{E}_{\pi} \lambda(\gamma, \cdot)$

For the Hellinger loss, the proper analogue is the spherical loss

$$\lambda^{\text{spherical}}(\pi, \omega) = 1 - \frac{\pi(\omega)}{\sqrt{\sum_{j=1}^{r} (\pi(j))^2}}$$

$$\lambda^{\text{spherical}}(\pi, \omega) = \lambda^{\text{Hellinger}}(\gamma, \omega) \text{ for } \gamma(\omega) = \frac{(\pi(\omega))^2}{\sum_{j=1}^{r} (\pi(j))^2}$$
Example: Mixable and Non-Mixable Losses

Experts 1, . . . , $K$ predict $\pi^{(k)} \in \mathcal{P} \{0, 1\}$.
Experts 1, . . . , $N$ predict $\gamma^{(n)} \in \{0, 1\}$.

Learner predicts $(\pi, \tilde{\pi}) \in \mathcal{P} \{0, 1\} \times \mathcal{P} \{0, 1\}$ such that

if $\pi(0) > 1/2$ then $\tilde{\pi}(0) = 1$ and if $\pi(1) > 1/2$ then $\tilde{\pi}(1) = 1$.

There exists a strategy for Learner that guarantees for any $k$

$$\sum_{t=1}^{T} \lambda^{\text{square}}(\pi_{t}, \omega_{t}) \leq \sum_{t=1}^{T} \lambda^{\text{square}}(\pi_{t}^{(k)}, \omega_{t}) + \ln(K + N)$$

and for any $m$

$$\sum_{t=1}^{T} \lambda^{\text{abs}}(\tilde{\pi}_{t}, \omega_{t}) \leq \sum_{t=1}^{T} \lambda^{\text{simple}}(\gamma_{t}^{(n)}, \omega_{t}) + O(\sqrt{T \ln(K + N) + T \ln \ln T})$$
Graphs of the negative regret $\text{Loss}^{(m)}_{\epsilon_k}(T) - \text{Loss}^{(m)}(T)$

Learner optimizes for both loss functions, using the DF algorithm with “mixability” and Hoeffding supermartingales.
The Mixed Supermartingale

\[
\frac{1}{K+N} \sum_{k=1}^{K} e^{2\sum_{t=1}^{T-1}( (p_t - \omega_t)^2 - (p_t^{(k)} - \omega_t)^2 )} \times e^{2((p - \omega)^2 - (p_T^{(k)} - \omega)^2 )} \\
+ \frac{1}{K+N} \sum_{n=1}^{N} \int_{0}^{1/e} \frac{d\eta}{\eta \left( \ln \frac{1}{\eta} \right)^2} \eta \sum_{t=1}^{T-1}( |\bar{p}_t - \omega_t| - [\gamma_t^{(n)} \neq \omega_t] ) - \eta^2 / 2 \\
\times e^{\eta(|\bar{p} - \omega| - [\gamma_T^{(n)} \neq \omega])} - \eta^2 / 2
\]

where \( p_t = \pi_t(1), p_t^{(k)} = \pi_t^{(k)}(1), \tilde{p}_t = \tilde{\pi}_t(1). \)

\([x \neq y] = 1 \text{ if } x \neq y \text{ and } [x \neq y] = 0 \text{ if } x = y.\)
References


http://onlineprediction.net/

Alexey Chernov (RHUL) Multiobjective Prediction with Expert Advice GTP Workshop, June 2010 18 / 18