CSIE5002 Prediction, learning, and games

Lecture 7: Aggregating algorithm

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

22.04.2019

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

Individual sequence prediction with the logarithmic loss is not the only problem that requires online prediction. In general, the loss is not necessarily logarithmic, and the goal might not be probability forecasting.

This lecture presents a general framework—called *learning with expert advice*—that addresses the issues.

Recommended reading

- N. Cesa-Bianchi and G. Lugosi. 2006. *Prediction, Learning, and Games*. Chapter 3.

- V. Vovk. 1998. A game of prediction with expert advice.

- K. Yamanishi. 1998. A decision-theoretic extension of stochastic complexity and its applications to learning.

- D. Haussler *et al.* 1998. Sequential prediction of individual sequences under general loss functions.

Table of contents

Individual sequence prediction as learning with expert advice

Recall: Individual sequence prediction with finitely many hypotheses

Individual sequence prediction with finitely many hypotheses. Let $\mathcal{A}$ be a finite alphabet. Let $\mathcal{H}$ be a finite class of hypotheses $h : \mathcal{A}^* \to \Delta \subset \mathbb{R}^{|\mathcal{A}|}$. Let $T \in \mathbb{N}$. For every $t = 1, \dots, T$, the following happen in order.

1. Learner outputs $\gamma_t = (\gamma_t(a))_{a \in \mathcal{A}} \in \Delta$.
2. Reality outputs $a_t \in \mathcal{A}$.

Regret.

$$R_T(h) := \sum_{t=1}^{T} \left[-\log \gamma_t(a_t) \right] - \sum_{t=1}^{T} \left[-\log h(a_t | a_{1:t-1}) \right], \quad \forall h \in \mathcal{H}.$$

Recall: Mixture forecaster

Let p_γ and p_h be the joint probability distributions induced by γ and h, respectively, i.e.,

$$p_\gamma(a_{1:t}) := \prod_{\tau=1}^{t} \gamma_\tau(a_\tau), \quad p_h(a_{1:t}) := \prod_{\tau=1}^{t} h(a_\tau | a_{1:\tau-1}).$$

Mixture forecaster. Let $(\pi_h)_{h\in\mathcal{H}}$ be a probability vector in $\mathbb{R}^{|\mathcal{H}|}$. The mixture forecaster is given by

$$p_\gamma(a_{1:t}) := \sum_{h\in\mathcal{H}} \pi(h) p_h(a_{1:t}), \quad \forall 1 \le t \le T, a_{1:t} \in \mathcal{A}^t.$$

Recall: Regret of the mixture forecaster

Theorem 1. The mixture forecaster satisfies

$$R_T(h) \leq \log\left(\frac{1}{\pi(h)}\right), \quad \forall h \in \mathcal{H}.$$

Proof. We write

$$\begin{aligned}
\sum_{t=1}^{T}\left[-\log \gamma_t(a_t)\right] &= -\log p_\gamma(a_{1:t}) \\
&= -\log \sum_{h\in\mathcal{H}} \pi(h) p_h(a_{1:t}) \\
&\leq -\log\left[\pi(h) p_h(a_{1:t})\right] \\
&= \log\left(\frac{1}{\pi(h)}\right) + \sum_{t=1}^{T}\left[-\log h(a_t|a_{1:t-1})\right], \quad \forall h \in \mathcal{H}.
\end{aligned}$$

Equivalent formulation as learning with expert advice

The following equivalent formulation is a special case of *learning with expert advice*.

Learning with expert advice with the logarithmic loss. Let $T \in \mathbb{N}$. Let the initial cumulative loss $L_0 = 0$. For every $1 \leq t \leq T$, the following happen in order.

1. Expert-h announces $\gamma_t(h) \in \Delta$, $h \in \mathcal{H}$.
2. Learner announces $\gamma_t \in \Delta$.
3. Reality announces $\omega_t \in \mathcal{A}$.
4. Update $L_t \leftarrow L_{t-1} + \lambda(\omega_t, \gamma_t)$, where $\lambda : (\omega, \gamma) \mapsto -\log \gamma[\omega]$.

Notice the abuse of notation above.

Equivalent definition of the regret

Regret. (Learning-with-expert-advice ver.) The regret compares the cumulative losses of LEARNER and any expert:

$$R_T(h) := \sum_{t=1}^{T} \lambda_t(\gamma_t) - \sum_{t=1}^{T} \lambda_t(\gamma_t(h)), \quad \forall h \in \mathcal{H}.$$

Aggregating algorithm. Let $(\pi_1(h))_{h\in\mathcal{H}}$ be a probability vector. For every $1 \le t \le T$, announce

$$\gamma_t := \sum_{h\in\mathcal{H}} \pi_t(h)\gamma_t(h),$$

and then, after seeing $\lambda_t(\gamma_t)$, update

$$\pi_{t+1}(h) := \frac{\pi_t(h)\mathrm{e}^{-\lambda_t(\gamma_t(h))}}{\sum_{h\in\mathcal{H}} \pi_t(h)\mathrm{e}^{-\lambda_t(\gamma_t(h))}}.$$

Equivalence to the mixture forecaster

Proposition 1. The strategy of LEARNER defined in the previous slide is equivalent to the mixture forecaster.

Proof. It is easily checked (by, e.g., induction) that

$$\pi_t(h) = \frac{\pi_1(h) \prod_{\tau=1}^{t-1} \gamma_\tau(h)[a_\tau]}{\sum_{h \in \mathcal{H}} \pi_1(h) \prod_{\tau=1}^{t-1} \gamma_\tau(h)[a_\tau]}, \quad \forall h \in \mathcal{H},$$

where $\gamma_\tau(h)[a_\tau]$ denotes $h(a_\tau | a_{1:\tau-1})$ in the language of individual sequence prediction. Then, we have

$$\gamma_t = \sum_{h \in \mathcal{H}} \pi_t(h) \gamma_t(h) = \frac{\sum_{h \in \mathcal{H}} \pi_1(h) \prod_{\tau=1}^{t} \gamma_\tau(h)[a_\tau]}{\sum_{h \in \mathcal{H}} \pi_1(h) \prod_{\tau=1}^{t-1} \gamma_\tau(h)[a_\tau]} = \frac{p_\gamma(a_{1:t})}{p_\gamma(a_{1:t-1})}.$$

Recall that p_γ is the joint probability distribution given by the mixture forecaster.

Regret of the aggregating algorithm

Corollary 1. The aggregating algorithm satisfies

$$R_T(h) \leq \log\left(\frac{1}{\pi(h)}\right), \quad \forall h \in \mathcal{H}.$$

Proof. By Proposition 1, it suffices to apply Theorem 1.

Question. What if the γ_t's are not in Δ, and the loss is not logarithmic?

Learning with expert advice

General learning with expert advice problem

Learning with expert advice. Let $T \in \mathbb{N}$. Let $\lambda : \Omega \times \Gamma \to \mathbb{R}$ be the loss function. Let $L_0 = 0$. For every $1 \leq t \leq T$, the following happen in order.

1. Expert-i announces $\gamma_t(i) \in \Gamma$, $1 \leq i \leq n$.
2. Learner announces $\gamma_t \in \Gamma$.
3. Reality announces $\omega_t \in \Omega$.
4. Update $L_t \leftarrow L_{t-1} + \lambda(\omega_t, \gamma_t)$.

Remark. We call Γ the *prediction space* and Ω the *outcome space*.

V. Vovk. 1998. A game of prediction with expert advice.

Regret

Regret. The regret is given by

$$R_T(i) := \sum_{t=1}^{T} \gamma(\omega_t, \gamma_t) - \min_{1 \le i \le n} \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t(i)), \quad \forall 1 \le i \le n.$$

Example. (Individual sequence prediction with the logarithmic loss) This has been shown in the previous section.

Example. (Majority vote) This corresponds to $\Omega = \Gamma = \{ 0, 1 \}$ and λ is the 0-1 loss.

Example: Brier & absolute losses

Consider the binary sequence prediction problem, which corresponds to $\Omega = \{0, 1\}$ and $\Gamma = [0, 1]$. The *Brier loss* is given by

$$\lambda_{\text{Brier}}(\omega, \gamma) := (\omega - \gamma)^2 .$$

The absolute loss is given by

$$\lambda_{\text{absolute}}(\omega, \gamma) := |\omega - \gamma|.$$

Question. How do we choose the loss function?

G. W. Brier. 1950. Verification of forecasts expressed in terms of probability.

D. Haussler *et al.* 1995. Tight worst-case loss bounds for predicting with expert advice.

V. Vovk. 2015. The fundamental nature of the log loss function.

Many other examples

Example. (Universal source coding) Obvious.

Example. (AdaBoost) To be introduced later.

Example. (Portfolio selection) To be introduced later.

Example. (Learning in games) See, e.g., Y. Freund and R. E. Schapire. 1999. Adaptive game playing using multiplicative weights.

Example. (Learning the learning rate) See, e.g., T. van Erven and W. Koolen. 2016. MetaGrad: Multiple learning rates in online learning.

Exercise. (Will be included in HW #2) Find an example of learning with expert advice that has not been listed.

Aggregating algorithm

Learning with expert advice as individual sequence prediction

Learning with expert advice as individual sequence prediction.
Let $T \in \mathbb{N}$. Let $\lambda : \Omega \times \Gamma \to \mathbb{R}$ be the loss function. *Let* $\tilde{L}_0 = 0$. *Let* $\eta > 0$. For every $1 \leq t \leq T$, the following happen in order.

1. Expert-i announces *a generalized prediction* $g_t(i) : \Omega \to \mathbb{R}$ *given by*
$$g_t(i)[\omega] := \mathrm{e}^{-\eta\lambda(\omega\gamma_t(i))}, \forall \omega \in \Omega.$$
2. Learner announces *a generalized prediction* $g_t : \Omega \to \mathbb{R}$.
3. Reality announces $\omega_t \in \Omega$.
4. *Update* $\tilde{L}_t \leftarrow \tilde{L}_{t-1} - \frac{1}{\eta}\log\{g_t[\omega_t]\}$.

Remark. The *generalized prediction* à la Vovk is $-(1/\eta)\log g_t$.

V. Vovk. 1998. A game of prediction with expert advice.

Aggregating pseudo-algorithm vs. mixture forecaster

Algorithm. (Aggregating pseudo-algorithm) Let $(\pi_i)_{1\le i\le n}$ be a probability vector. For every $1 \le t \le T$, announce g_t as

$$g_t[\omega] = \frac{\sum_{1\le i\le n} \pi_i g_t(i)[\omega] \prod_{\tau=1}^{t-1} g_\tau(i)[\omega_\tau]}{\sum_{1\le i\le n} \pi_i \prod_{\tau=1}^{t-1} g_\tau(i)[\omega_\tau]}, \quad \forall \omega \in \Omega.$$

Remark. View $g_t(i)[\omega]$ as *pseudo-conditional probability forecasts*. Then, the aggregating pseudo-algorithm (APA) corresponds to the mixture forecaster.

V. Vovk. 1990. Aggregating strategies.

V. Vovk. 1998. A game of prediction with expert advice.

V. Vovk. 2001. Competitive on-line statistics.

Regret of the APA

Lemma 1. The APA satisfies, for every $1 \leq i \leq n$,

$$\sum_{t=1}^{T} \left(-\frac{1}{\eta} \log g_t[\omega_t] \right) - \sum_{t=1}^{T} \left(-\frac{1}{\eta} \log \{g_t(i)[\omega_t]\} \right) \leq \frac{1}{\eta} \log \frac{1}{\pi_i}.$$

Remark. Notice that

$$\begin{aligned} \sum_{t=1}^{T} \left(-\frac{1}{\eta} \log \{g_t(i)[\omega_t]\} \right) &= \sum_{t=1}^{T} \left[-\frac{1}{\eta} \log \mathrm{e}^{-\eta \lambda(\omega_t, \gamma_t(i))} \right] \\ &= \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t(i)). \end{aligned}$$

Proof of Lemma 1

Proof. (Lemma 1) Let p and p_i be the *pseudo-joint probabilities* defined by g_t's and $g_t(i)$'s, respectively; that is,

$$p(\omega_{1:T}) := \prod_{t=1}^{T} g_t[\omega_t], \quad p_i(\omega_{1:T}) := \prod_{t=1}^{T} g_t(i)[\omega_t].$$

Notice that

$$p(\omega_{1:T}) = \sum_{1 \le i \le n} \pi_i p_i(\omega_{1:T}).$$

Then, we write

$$\begin{aligned} -\frac{1}{\eta} \log p(\omega_{1:T}) &\le -\frac{1}{\eta} \log \left[\pi_i p_i(\omega_{1:T})\right] \\ &= -\frac{1}{\eta} \log p_i(\omega_{1:T}) + \frac{1}{\eta} \log \frac{1}{\pi_i}, \quad \forall 1 \le i \le n. \end{aligned}$$

Mixability

We desire to have

$$\lambda(\omega_t, \gamma_t) \leq -\frac{1}{\eta} \log g_t[\omega_t], \quad \forall 1 \leq t \leq T.$$

Definition. (η-mixability) We say a loss function $\lambda : \Omega \times \Gamma \to \mathbb{R}$ is *η-mixable*, if and only if for every probability vector $(q_i)_{1 \leq i \leq n}$ and $\gamma_1, \ldots, \gamma_n \in \Gamma$, there exists some $\gamma \in \Gamma$ such that

$$\lambda(\omega, \gamma) \leq -\frac{1}{\eta} \log \left(\sum_{1 \leq i \leq n} q_i \, \mathrm{e}^{-\eta \lambda(\omega, \gamma_i)} \right), \quad \forall \omega \in \Omega.$$

Aggregating algorithm

Algorithm. (Aggregating algorithm) Let $(\pi_1(i))_{1\le i\le n}$ be a probability vector. For every $1 \le t \le T$, announce any γ_t such that

$$\lambda(\omega, \gamma_t) \le -\frac{1}{\eta} \log \left(\sum_{1\le i\le n} \pi_t(i) \mathrm{e}^{-\eta\lambda(\omega,\gamma_t(i))} \right), \quad \forall \omega \in \Omega,$$

and update

$$\pi_{t+1}(i) \propto \pi_t \mathrm{e}^{-\eta\lambda(\omega_t,\gamma_t(i))}, \quad \forall 1 \le i \le n.$$

Remark. Recall the APA:

$$g_t[\omega] = \frac{\sum_{1\le i\le n} \pi_i g_t(i)[\omega] \prod_{\tau=1}^{t-1} g_\tau(i)[\omega_\tau]}{\sum_{1\le i\le n} \pi_i \prod_{\tau=1}^{t-1} g_\tau(i)[\omega_\tau]}, \quad \forall \omega \in \Omega.$$

Regret of the aggregating algorithm

Theorem 2. If the loss λ is η-mixable, then the aggregating algorithm (AA) is well-defined and satisfies

$$R_T := \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) - \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t(i)) \leq \frac{1}{\eta} \log \frac{1}{\pi_1(i)}, \quad \forall 1 \leq i \leq n.$$

Remark. Notice that $R_T = O(1)$.

Proof. (Theorem 2) It is obvious that AA is well-defined with the mixability condition. It remains to apply Lemma 1.

V. Vovk. 1990. Aggregating strategies.

V. Vovk. 1998. A game of prediction with expert advice.

Vovk's game (1/3)

Consider the following game $\mathcal{G}(c, a)$ between LEARNER and REALITY. Let $(\Omega, \Gamma, \lambda)$ be given. We say LEARNER wins if LEARNER can achieve

$$\sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) \leq c \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t(i)) + a \log n,$$

for all T and n, and REALITY wins otherwise.

Question. Is every game $\mathcal{G}(c, a)$ *determined*? When will LEARNER win?

V. Vovk. 1990. Aggregating strategies.

V. Vovk. 1998. A game of prediction with expert advice.

Vovk's game (2/3)

Define

$$c(\eta) := \inf \left\{ c \;\middle|\; \forall P \, \exists \gamma^* \in \Gamma \, \forall \omega : \lambda(\omega, \gamma^*) \le \frac{-c}{\eta} \log \mathsf{E}_{\gamma \sim P} \mathrm{e}^{-\eta \lambda(\omega, \gamma)} \right\},$$

where the probability distributions P can only have finite supports.

Definition. (Separation curve) The *separation curve* is given by

$$\{ (c(\eta), a(\eta)) \mid \eta \in [0, +\infty] \},$$

where

$$a(\eta) := \frac{c(\eta)}{\eta}.$$

V. Vovk. 1990. Aggregating strategies.

V. Vovk. 1998. A game of prediction with expert advice.

Vovk's game (3/3)

Theorem 3. (Vovk) Suppose the following hold.

- The prediction set Γ is a compact topological space.
- The mapping $\gamma \mapsto \lambda(\omega, \gamma)$ is continuous for all $\omega \in \Omega$.
- There exists some γ such that $\lambda(\omega, \gamma) < +\infty$ for all ω.
- There does not exists any γ such that $\lambda(\omega, \gamma) = 0$ for all ω.

Then, each game $\mathcal{G}(c, a)$ is determined; $\textsc{Learner}$ wins the game $\mathcal{G}(c, a)$, if and only if (c, a) is northeast to (or in) the separation curve; Furthermore, $c(\eta) \geq 1$.

V. Vovk. 1990. Aggregating strategies.

V. Vovk. 1998. A game of prediction with expert advice.

General cumulative loss bound for AA (1/2)

Definition. ((c,η)-mixability) We say a loss $\lambda : \Omega \times \Gamma \to \mathbb{R}$ is (c,η)-mixable, if and only if for every probability vector $(q_i)_{1\le i\le n}$ and $\gamma_1, \ldots, \gamma_n \in \Gamma$, there exists some γ^* such that

$$\lambda(\omega, \gamma^*) \le \frac{-c}{\eta} \log \left(\sum_{1\le i\le n} q_i \mathrm{e}^{-\eta\lambda(\omega,\gamma_i)} \right), \quad \forall \omega \in \Omega.$$

General cumulative loss bound for AA (2/2)

Algorithm. (Aggregating algorithm) Let $(\pi_1(i))_{1\le i\le n}$ be a probability vector. For every $1 \le t \le T$, announce any γ_t such that

$$\lambda(\omega, \gamma_t) \le -\frac{c}{\eta} \log \left(\sum_{1\le i\le n} \pi_t(i) \mathrm{e}^{-\eta\lambda(\omega,\gamma_t(i))} \right), \quad \forall \omega \in \Omega,$$

and update

$$\pi_{t+1}(i) \propto \pi_t \mathrm{e}^{-\eta\lambda(\omega_t,\gamma_t(i))}, \quad \forall 1 \le i \le n.$$

Theorem 4. Suppose the loss is (c, a)-mixable. Then, AA achieves

$$\sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) \le c \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t(i)) + \frac{c}{\eta} \log \frac{1}{\pi_1(i)}, \quad \forall 1 \le i \le n.$$

Proof. Exercise.

Mixable losses

0-1 loss

Recall the voting protocol. Let $\Omega = \Gamma = \{0, 1\}$. Consider the 0-1 loss:

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

Proposition 2. (Vovk) The 0-1 loss is (c, η)-mixable if and only if

$$c \geq \frac{\eta}{\log\left(\frac{2}{1+\mathrm{e}^{-\eta}}\right)}.$$

Corollary 2. For every $\alpha > 2$, there exists an α-competitive online algorithm for the voting protocol.

V. Vovk. 1998. A game of prediction with expert advice.

Logarithmic loss

Suppose that $\Omega = \{ 0, 1 \}$ and $\Gamma = [0, 1]$. Let $\lambda : \Omega \times \Gamma \to \mathbb{R}$ be the logarithmic loss:

$$\lambda(\omega, \gamma) = \omega \log \frac{\omega}{\gamma} + (1 - \omega) \log \frac{1 - \omega}{1 - \gamma}.$$

Proposition 3. The logarithmic loss is 1-mixable.

Remark. Notice the resulting AA becomes the mixture forecaster.

V. Vovk. 1998. A game of prediction with expert advice.

Brier loss

Suppose that $\Omega = \{ 0, 1 \}$ and $\Gamma = [0, 1]$. Let $\lambda : \Omega \times \Gamma \to \mathbb{R}$ be the Brier loss:

$$\lambda(\omega, \gamma) \coloneqq (\omega - \gamma)^2 .$$

Proposition 4. The Brier loss is 2-mixable.

Remark. See Section 3.6 of *Prediction, Learning, and Games* for the closed-form expression of the associated AA.

G. W. Brier. 1950. Verification of forecasts expressed in terms of probability.

V. Vovk. 1998. A game of prediction with expert advice.

Absolute loss

Suppose that $\Omega = \Gamma = [0, 1]$. Let $\lambda : \Omega \times \Gamma \to \mathbb{R}$ be the absolute loss:

$$\lambda(\omega, \gamma) := |\omega - \gamma|.$$

Proposition 5. The absolute loss is (c, η)-mixable if and only if

$$c \geq \frac{\eta}{2 \log\left(\frac{2}{1+\mathrm{e}^{-\eta}}\right)}.$$

V. Vovk. 1998. A game of prediction with expert advice.

Conclusions

Conclusions

- AA is a generalization of the mixture forecaster.
- AA requires the loss to be either η- or (c, η)-mixable.
- When the loss is η-mixable, AA achieves an $O(1)$ regret.
  When the loss is (c, η)-mixable, AA is c-competitive.

Next lecture

- Aggregating algorithm continued.