Online Prediction: Bayes versus Experts

Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

PASCAL-2004, July 19-21

Table of Contents

Abstract

We derive a very general regret bound in the framework of prediction with expert advice, which challenges the best known regret bound for Bayesian sequence prediction. Both bounds of the form $\sqrt{\text{Loss} \times \text{complexity}}$ hold for any bounded loss-function, any prediction and observation spaces, arbitrary expert/environment classes and weights, and unknown sequence length.

Keywords

Bayesian sequence prediction;
Prediction with Expert Advice;
general weights, alphabet and loss.

Sequential/online predictions

In sequential or online prediction, for $t = 1, 2, 3, ...$,

our predictor p makes a prediction $y_t^p \in \mathcal{Y}$

based on past observations $x_1, ..., x_{t-1}$.

Thereafter $x_t \in \mathcal{X}$ is observed and p suffers loss $\ell(x_t, y_t^p)$.

The goal is to design predictors with small total loss or cumulative $\text{Loss}_{1:T}(p) := \sum_{t=1}^{T} \ell(x_t, y_t^p)$.

Applications are abundant, e.g. weather or stock market forecasting.

Example:

Loss $\ell(x, y)$	$\mathcal{X} = \{$sunny	rainy$\}$
$\mathcal{Y} = \{$ umbrella	0.1	0.3
sunglasses $\}$	0.0	1.0

Setup also includes: Classification and Regression problems.

Bayesian Sequence Prediction

Bayesian Sequence Prediction - Setup

- Assumption: Sequence $x_1...x_T$ is sampled from some distribution μ, i.e. the probability of $x_{<t} := x_1...x_{t-1}$ is $\mu(x_{<t})$.
- The probability of the next symbol being x_t, given $x_{<t}$, is $\mu(x_t|x_{<t})$
- Goal: minimize the μ-expected-Loss $=:$ $\bar{\text{L}}\text{oss}$.
- More generally: Define the Bayes_ρ sequence prediction scheme

$$y_t^\rho \;:=\; \arg\min_{y_t\in\mathcal{Y}} \sum_{x_t} \rho(x_t|x_{<t})\ell(x_t,y_t),$$

  which minimizes the ρ-expected loss.

- If μ is known, Bayes_μ is obviously the best predictor in the sense of achieving minimal expected loss: $\bar{\text{L}}\text{oss}_{1:T}(\text{Bayes}_\mu) \leq \bar{\text{L}}\text{oss}_{1:T}(\text{Any } p)$

The Bayes-mixture distribution ξ

- Assumption: The true (objective) environment μ is unknown.
- Bayesian approach: Replace true probability distribution μ by a Bayes-mixture ξ.
- Assumption: We know that the true environment μ is contained in some known (finite or countable) set $\mathcal{M}$ of environments.
- The Bayes-mixture ξ is defined as

$$\xi(x_{1:m}) := \sum_{\nu\in\mathcal{M}} w_\nu \nu(x_{1:m}) \quad \text{with} \quad \sum_{\nu\in\mathcal{M}} w_\nu = 1, \quad w_\nu > 0 \; \forall\nu$$

- The weights w_ν may be interpreted as the prior degree of belief that the true environment is ν, or $k^\nu = \ln w_\nu^{-1}$ as a complexity penalty (prefix code length) of environment ν.
- Then $\xi(x_{1:m})$ could be interpreted as the prior subjective belief probability in observing $x_{1:m}$.

Bayesian Loss Bound

Under certain conditions, $\bar{\text{Loss}}_{1:T}(\text{Bayes}_\xi)$ is bounded by $\bar{\text{Loss}}_{1:T}(\text{Any } p)$ (and hence by the loss of the best predictor in hindsight Bayes_μ):

$$\boxed{\bar{\text{Loss}}_{1:T}(\text{Bayes}_\xi) \leq \bar{\text{Loss}}_{1:T}(\text{Any } p) + 2\sqrt{\bar{\text{Loss}}_{1:T}(\text{Any } p) \cdot k^\mu} + 2k^\mu \quad \forall\mu \in \mathcal{M}}$$

Note that $\bar{\text{Loss}}_{1:T}$ depends on μ. Proven for countable $\mathcal{M}$ and $\mathcal{X}$, finite $\mathcal{Y}$, any k^μ, and any bounded loss function $\ell : \mathcal{X} \times \mathcal{Y} \to [0,1]$ [H'01–03]

For finite $\mathcal{M}$, the uniform choice $k^\nu = \ln|\mathcal{M}| \ \forall\nu \in \mathcal{M}$ is common.

For infinite $\mathcal{M}$, k^ν = complexity of ν is common (Occam,Solomonoff).

Prediction with Expert Advice

Prediction with Expert Advice (PEA) - Setup

Given a countable class of $\mathcal{E}$ experts,

each $\text{Expert}_e \in \mathcal{E}$ at times $t = 1, 2, ...$ makes a prediction y_t^e.

The goal is to construct a master algorithm, which exploits the experts, and predicts asymptotically as well as the best expert in hindsight.

More formally, a PEA-Master is defined as:

For $t = 1, 2, ..., T$

- Predict $y_t^{\text{PEA}} := \text{PEA}(x_{<t}, \mathbf{y}_t, \text{Loss})$
- Observe $x_t := \text{Env}(\mathbf{y}_{<t}, x_{<t}, y_{<t}^{\text{PEA}}?)$
- Receive $\text{Loss}_t(\text{Expert}_e) := \ell(x_t, y_t^e)$ for each $\text{Expert}_e \in \mathcal{E}$
- Suffer $\text{Loss}_t(\text{PEA}) := \ell(x_t, y_t^{\text{PEA}})$

Notation: $x_{<t} := (x_1, ..., x_{t-1})$ and $\mathbf{y}_t = (y_t^e)_{e\in\mathcal{E}}$.

Goals

BEH := Best Expert in Hindsight = Expert of minimal total Loss.

$\text{Loss}_{1:T}(\text{BEH}) = \min_{e\in\mathcal{E}} \text{Loss}_{1:T}(\text{Expert}_e)$.

0) Regret := $\text{Loss}_{1:T}(\text{PEA}) - \text{Loss}_{1:T}(\text{BEH})$ shall be small ($O(\sqrt{\text{Loss}_{1:T}(\text{BEH})})$).

1) Any bounded Loss function (w.l.g. $0 \leq \text{Loss}_t \leq 1$). Literature: Mostly specific Loss (absolute, 0/1, log, square)

2) Neither (non-trivial) upper bound on total Loss, nor sequence length T is known. Solution: Adaptive learning rate.

3) Infinite number of Experts. Motivation:
   - Expert_e = polynomial of degree $e = 1, 2, 3, ...$ through data -or-
   - $\mathcal{E}$ = class of all computable (or finite state or ...) Experts.

Weighted Majority (WM)

Take expert which performed best in past with high probability and others with smaller probability.

At time t, select Expert I_t^{WM} with probability

$$\boxed{P[I_t^{\text{WM}}=e] \;\propto\; w^e \cdot \exp[-\eta_t \cdot \text{Loss}_{<t}(\text{Expert}_e)]}$$

η_t = learning rate, w^e = initial weight.

[Littlestone&Warmuth'90 (Classical)]: 0/1 loss and η_t=const.

[Freund&Shapire'97 (Hedge)] and others: General Loss, but η_t=const.

[Cesa-Bianchi et al.'97]: Piecewise constant η_t. Only $1/w^e=|\mathcal{E}|<\infty$.

[Auer&Gentile'00, Yaroshinsky et al.'04]: Smooth $\eta_t \searrow 0$, but only 0/1 Loss and $1/w^e=|\mathcal{E}|<\infty$.

Follow the Perturbed Leader (FPL)

Select expert of minimal perturbed and penalized Loss.

Let Q_t^e be i.i.d. random variables and k^e complexity penalty.

Select expert $\boxed{I_t^{\text{FPL}} := \arg\min_e\{\eta_t \text{Loss}_{<t}(\text{Expert}_e) + k^e + Q_t^e\}}$

[Hannan'57]: $Q_t^e \overset{d.}{\sim} \text{Uniform}[0,1]$, [Kalai&V.'03]: $P[Q_t^e = u] = \frac{1}{2}\exp(-|u|)$

Both: $k^e = 0$, $|\mathcal{E}| < \infty$, $\eta_t \propto 1/\sqrt{t} \quad \Longrightarrow \quad \text{Regret}{=}O(\sqrt{|\mathcal{E}|\cdot T})$.

[Hutter&Poland'04]: $P[Q_t^e = -u] = \exp(-u) \quad (u \geq 0)$,

General k^e and $\mathcal{E}$ and $\eta_t \propto 1/\sqrt{\text{Loss}} \quad \Longrightarrow \quad \text{Regret}{=}O(\sqrt{k^e\cdot\text{Loss}})$.

For all PEA variants (WM & FPL & others) it holds:

$P[I_t = e] = \left\{\begin{smallmatrix} large \\ small \end{smallmatrix}\right\}$ if Expert_e has $\left\{\begin{smallmatrix} small \\ large \end{smallmatrix}\right\}$ Loss.

$I_t \xrightarrow{\eta\to\infty}$ Best Expert in Past ($\eta =$ learning rate)

$I_t \xrightarrow{\eta\to 0}$ Uniform distribution among Experts.

FPL Regret Bounds for $|\mathcal{E}|<\infty$ and $k^e=\ln|\mathcal{E}|$

Since FPL is randomized, we need to consider expected-Loss $=:$ $\underline{\text{L}}$oss.

Regret $:=$ $\underline{\text{L}}\text{oss}_{1:T}(\text{FPL})-\text{Loss}_{1:T}(\text{BEH})$.

Static $\eta_t=\sqrt{\frac{\ln|\mathcal{E}|}{T}}$ $\Longrightarrow$	Regret $\leq 2\sqrt{T\cdot\ln|\mathcal{E}|}$
Dynamic $\eta_t=\sqrt{\frac{\ln|\mathcal{E}|}{2t}}$ $\Longrightarrow$	Regret $\leq 2\sqrt{2T\cdot\ln|\mathcal{E}|}$
Self-confident $\eta_t=\sqrt{\frac{\ln|\mathcal{E}|}{2(\underline{\text{L}}\text{oss}_{<t}(\text{FPL})+1)}}$ $\Longrightarrow$	Regret $\leq 2\sqrt{2(\text{Loss}_{1:T}(\text{BEH})+1)\cdot\ln|\mathcal{E}|}+8\ln|\mathcal{E}|$
Adaptive $\eta_t=\sqrt{\frac{1}{2}}\min\left\{1,\sqrt{\frac{\ln|\mathcal{E}|}{\text{Loss}_{<t}(\text{“BEH”})}}\right\}$ $\Longrightarrow$	Regret $\leq 2\sqrt{2\text{Loss}_{1:T}(\text{BEH})\cdot\ln|\mathcal{E}|}+5\ln|\mathcal{E}|\cdot\ln\text{Loss}_{1:T}(\text{BEH})+3\ln|\mathcal{E}|+6$

No hidden $O()$ terms!

FPL Regret Bounds for $|\mathcal{E}|=\infty$ and general k^e

Assume complexity penalty k^e such that $\sum_{e\in\mathcal{E}}\exp(-k^e)\leq 1$.

We expect $\ln|\mathcal{E}|\rightsquigarrow k^e$, i.e. Regret $=O(\sqrt{k^e\cdot(\text{Loss or }T)})$.

Problem: Choice of $\eta_t=\sqrt{k^e/...}$ depends on e. Proofs break down.

Choose: $\eta_t=\sqrt{1/...}$ $\Rightarrow$ Regret $\leq k^e\sqrt{\cdots}$, i.e. k^e not under $\sqrt{\ }$.

Solution: Two-Level **Hierarchy of Experts**:

Group all experts of (roughly) equal complexity.

- FPL^K over subclass of experts with complexity $k^e\in(K-1,K]$. Choose $\eta_t^K=\sqrt{K/2\text{Loss}_{<t}}=$ constant within subclass.
- Regard each FPL^K as a (meta)expert. Construct from them (meta) $\widetilde{\text{FPL}}$. Choose $\tilde\eta_t=\sqrt{1/\text{Loss}_{<t}}$ and $\tilde k^K=\frac{1}{2}+2\ln K$.

$$\Longrightarrow \quad \boxed{\text{Regret}\leq 2\sqrt{2\,k^e\cdot\text{Loss}_{1:T}(\text{Expert}_e)}\cdot\left(1+O\left(\frac{\ln k^e}{\sqrt{k^e}}\right)\right)+O(k^e)}$$

PEA versus Bayes

PEA versus Bayes Bounds – Formal

Formal similarity and duality between Bayes and PEA bounds is striking:

$$\bar{\text{L}}\text{oss}_{1:T}(\text{Bayes}_\xi) \;\leq\; \bar{\text{L}}\text{oss}_{1:T}(\text{Any }p) + 2\sqrt{\bar{\text{L}}\text{oss}_{1:T}(\text{Any }p)\cdot k^\mu} + 2k^\mu$$

$$\underline{\text{L}}\text{oss}_{1:T}(\text{PEA}) \leq \text{Loss}_{1:T}(\text{Expert}_e) + c\cdot\sqrt{\text{Loss}_{1:T}(\text{Expert}_e)\cdot k^e} + b\cdot k^e$$

$c = 2\sqrt{2}$ and $b = 8$ for PEA = FPL.

	beats predictors	in environment	expectation w.r.t.	function of
Bayes	all p	$\mu\in\mathcal{M}$	environment μ	$\mathcal{M}$
PEA	$\text{Expert}_e\in\mathcal{E}$	any $x_1...x_T$	prob. prediction	$\mathcal{E}$

Apart from these formal duality, there is a real connection between both bounds.

PEA Bound reduced to Bayes Bound

Regard class of Bayes-predictors $\{\text{Bayes}_\nu : \nu \in \mathcal{M}\}$ as class of experts $\mathcal{E}$.

The corresponding FPL algorithm then satisfies PEA bound

$$\underline{\text{L}}\text{oss}_{1:T}(\text{PEA}) \leq \text{Loss}_{1:T}(\text{Bayes}_\mu) + c\cdot\sqrt{\text{Loss}_{1:T}(\text{Bayes}_\mu)k^\mu} + b\cdot k^\mu.$$

Take the μ-expectation, and use $\bar{\text{L}}\text{oss}_{1:T}(\text{Bayes}_\mu) \leq \bar{\text{L}}\text{oss}_{1:T}(\text{Any } p)$ and Jensen's inequality, to get a Bayes-like bound for PEA

$$\boxed{\underline{\bar{\text{L}}}\text{oss}(\text{PEA}) \;\leq\; \bar{\text{L}}\text{oss}_{1:T}(\text{Any } p) + c\cdot\sqrt{\bar{\text{L}}\text{oss}_{1:T}(\text{Any } p)\cdot k^\mu} + b\cdot k^\mu \;\; \forall\mu \in \mathcal{M}}$$

Ignoring details, instead of using Bayes_ξ, one may use PEA with same/similar performance guarantees as Bayes_ξ.

Additionally, PEA has worst-case guarantees, which Bayes lacks.

So why use Bayes at all?

Open Problems

- We only compared *bounds* on PEA and Bayes. What about the actual (practical or theoretical) relative performance?

- Can FPL regret constant $c = 2\sqrt{2}$ be improved to $c = 2$?
For Hedge/FPL? Conjecture: Yes for Hedge, since Bayes has $c = 2$.

- Generalize existing bounds for WM-type masters (e.g. Hedge) to general $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{E}$, and $\ell \in [0, 1]$, similarly to FPL.

- Generalize FPL bound to infinite $\mathcal{E}$ and general k^e without the hierarchy trick (like for Bayes) (with expert dependent η_t^e?)

- Try first to prove weaker regret bounds with $\sqrt{\text{Loss}_{1:T}} \rightsquigarrow \sqrt{T}$.

More on (PEA) Regret Constant

Constant c in Regret $= c\cdot\sqrt{\text{Loss}\cdot k^e}$ for various settings and algorithms.

η	Loss	Optimal	LowBnd	Upper Bound
static	0/1	1?	1?	$\sqrt{2}$ [V'95]
static	any	$\sqrt{2}$!	$\sqrt{2}$ [V'95]	$\sqrt{2}$ [FS'97], 2 [FPL]
dynamic	0/1	$\sqrt{2}$?	1 [H'03]?	$\sqrt{2}$ [YEYS'04], $2\sqrt{2}$ [ACBG'02]
dynamic	any	2 ?	$\sqrt{2}$ [V'95]	$2\sqrt{2}$ [FPL], 2 [H'03]

Major open Problems

- Elimination of hierarchy (trick)
- Lower regret bound for infinite #Experts
- Same results (dynamic η_t, any Loss, $|\mathcal{E}|=\infty$) for WM

Some more FPL Results

Lower bound: $\underline{\text{Loss}}_{1:T}(\text{FPL}) \;\geq\; \text{Loss}_{1:T}(\text{BEH}) + \frac{\ln|\mathcal{E}|}{\eta_T}$ if $k^e = \ln|\mathcal{E}|$.

Bounds with high probability (Chernoff-Hoeffding):
$P[|\text{Loss}_{1:T} - \underline{\text{Loss}}_{1:T}| \geq \sqrt{3c\underline{\text{Loss}}_{1:T}}] \;\leq\; 2\exp(-c)$ is tiny for e.g. $c = 5$.

Computational aspects: It is trivial to generate the randomized decision of FPL. If we want to *explicitly* compute the probability we need to compute a 1D integral.

Deterministic prediction: FPL can be derandomized if prediction space $\mathcal{Y}$ and loss-function $\text{Loss}(x, y)$ are convex.

Thanks!

Questions?

Details:

http://www.idsia.ch/~marcus/ai/expert.htm [ALT 2004]

http://www.idsia.ch/~marcus/ai/spupper.htm [IEEE-TIT 2003]