Universal Prediction with general loss fns Part 2

CWI Peter Grünwald Universiteit Leiden

Centrum Wiskunde & Informatica – Amsterdam
Mathematisch Instituut – Universiteit Leiden

Universal Prediction with log-loss

- On each i (day), Marjon and Peter **announce the probability** that $y_{i+1} = 1$,i.e. that it will rain on day $i+1$
- We would like to combine their predictions in some way such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as whoever turns out to be the best forecaster for that sequence in terms of their cumulative log-loss
  - If, with hindsight, Marjon was better, we predict as well as Marjon
  - If, with hindsight, Peter was better, we predict as well as Peter

Universal Prediction with 0/1-loss

- On each i (day), Marjon and Peter predict whether it will rain on day $i+1$, i.e. **they announce '1' or '0'**
- If their prediction is wrong, their loss is 1, otherwise 0

Universal Prediction with 0/1-loss

- On each i (day), Marjon and Peter predict whether it will rain on day $i+1$, i.e. **they announce '1' or '0'**
- If their prediction is wrong, their loss is 1, otherwise 0
- We would like to combine their predictions in some way such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as whoever turns out to be the best forecaster for that sequence in terms of cumulative 0/1-loss **(total nr of mistakes)**
  - If, with hindsight, Marjon was better, we predict as well as Marjon
  - If, with hindsight, Peter was better, we predict as well as Peter

Universal prediction with log loss

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as the best forecaster for that sequence
- It turns out that there exists a universal strategy $\bar{S}$ such that, for all $n, y_1, \ldots, y_n \in \{0,1\}^n$

$$\text{loss}(y_1 \ldots, y_n, \bar{S}) \leq$$
$$\min\{\text{loss}(y_1 \ldots, y_n, S_{\text{Marjon}}), \text{loss}(y_1 \ldots, y_n, S_{\text{Peter}})\} + 1.$$

Universal prediction with 0/1-loss

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as the best forecaster for that sequence
- It turns out that there exists a universal strategy $\bar{S}$ such that, for all $n, y_1, \ldots, y_n \in \{0,1\}^n$

$$\text{loss}(y_1 \ldots, y_n, \bar{S}) \leq$$
$$\min\{\text{loss}(y_1 \ldots, y_n, S_{\text{Marjon}}), \text{loss}(y_1 \ldots, y_n, S_{\text{Peter}})\} + \sqrt{n}$$

Today: Beyond Bayes

- Good Algorithm for log-regret was Bayesian
  - Algorithm is minimax optimal up to constant factor (Vovk '99)
- Bayes can fail dramatically for 0/1-loss (and somewhat less dramatically for many other loss functions)
- Yet, intriguingly, a "simple" modification of Bayes is again essentially minimax optimal for general loss functions, including 0/1 – this is the 'Hedge' algorithm
- Today we are going to see how Hedge and Bayes hang together!

Bayes is good with log-loss

- For all n, y^n, all θ :

  $$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) - \log W(\theta)$$

- For all sequences of each length n, regret of Bayes bounded by constant depending on θ, not on n
- For "nonmixable" loss functions like 0/1-loss and absolute loss, this does not work
  - standard Bayes does not perform well at all in worst-case
  - Optimal algorithm gets much larger regret of order $\sqrt{n(-\log W(\theta))}$ in worst-case

WHY??

Menu

1. From log-loss to 0/1 loss
2. Two Problems with Bayes Prediction for 0/1 loss
3. Generalizing Bayes and making it work for general losses
4. Final Remarks about General Losses

"Generalized Bayes" for general loss functions

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
  - Log-loss: $\mathcal{A} = \Delta(\mathcal{Y}), \text{loss}(y, p) = -\log p(y)$
  - 0/1-loss: $\mathcal{A} = \mathcal{Y} = \{0, 1\}, \text{loss}(y, a) = |y - a|$

The Plan: turn loss of interest into log loss

Fix the loss function of interest loss.

We will construct a mapping that sends each action a to be judged relative to loss to another action p_a (a probability distribution) such that $-\log p_a(y)$, the log loss of p_a is a linear function of $\text{loss}(y, a)$, the loss of interest

Then use 'telescoping' again

We will run into difficulties at some point...but these will lead to highly interesting theoretical development

Entropification (The Gauss Device)

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- For every action a and $\eta > 0$ define density

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)}$$

Entropification (The Gauss Device)

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- For every action a and $\eta > 0$ define density

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)} \qquad Z(\eta, \mathbf{a}) = \int_y e^{-\eta \text{loss}(y,a)} dy$$

We can do this for just about every loss function, and will do it here for 0/1-loss. Complications with $Z(\eta)$ for nonsymmetric losses can be solved (G., 2008)

Entropification (The Gauss Device)

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- For every action a and $\eta > 0$ define density

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)} \qquad Z(\eta, \mathbf{a}) = \int_y e^{-\eta \text{loss}(y,a)} dy$$

- Example: 0/1-loss: $Z(\eta) = e^{-\eta \cdot 0} + e^{-\eta \cdot 1} = 1 + e^{-\eta}$

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)} = \begin{cases} \frac{e^{-\eta}}{1+e^{-\eta}} = \frac{1}{2} - b & \text{if } y \neq a \\ \frac{1}{1+e^{-\eta}} = \frac{1}{2} + b & \text{if } y = a \end{cases}$$

Entropification (The Gauss Device)

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- For every action a and $\eta > 0$ define density

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)} \qquad Z(\eta, \mathbf{a}) = \int_y e^{-\eta \text{loss}(y,a)} dy$$

- Example: squared loss:

$$\text{loss}(y, a) = (y - a)^2, \mathcal{Y} = \mathcal{A} = \mathbb{R}$$

$$p_{a,\eta}(y) = \frac{1}{Z(\eta)} e^{-\eta(y-a)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{\sigma^2}(y-a)^2}$$

Entropification (The Gauss Device)

- log loss of constructed distributions is affine (linear+constant) function of loss of interest

$$\text{log-loss}(y, p_{a,\eta}) = -\log p_{a,\eta}(y) = \eta \text{loss}(y, a) + \ln Z(\eta)$$

- log-loss difference is even linear in loss-difference:

$$\text{log-loss}(y, p_{a,\eta}) - \text{log-loss}(y, p_{a',\eta}) = \eta \left(\text{loss}(y, a) - \text{loss}(y, a') \right)$$

action is good for original loss iff transformed action is good for log-loss!

Entropification (The Gauss Device)

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- For every action a and $\eta > 0$ define prob. mass fn.

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)}$$

- A strategy relative to $\mathcal{A}$ is a function $S : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{A}$

$$S : \mathcal{Y}^* \to \mathcal{A}$$

Entropification (The Gauss Device)

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- For every action a and $\eta > 0$ define prob. mass fn.

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)}$$

- A strategy relative to $\mathcal{A}$ is a function $S : \mathcal{Y}^* \to \mathcal{A}$
- Recall the notation

$$\text{loss}(y^n, S) := \sum_{i=1}^{n} \text{loss}(y_i, S(y^{i-1}))$$

Entropification (The Gauss Device)

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- For every action a and $\eta > 0$ define prob. mass fn.

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)}$$

- A strategy relative to $\mathcal{A}$ is a function $S : \mathcal{Y}^* \to \mathcal{A}$
- Extend definition to strategies as:

$$p_{S,\eta}(y^n) := \prod_{i=1}^{n} p_{S(y^{i-1}),\eta}(y_i) = \frac{1}{Z(\eta)^n} \cdot e^{-\eta \text{loss}(y^n,S)}$$

Entropification (The Gauss Device)

- Let loss : $\mathcal{Y} \times \mathcal{A} \to [0, \infty]$ be arbitrary loss fn.
- For every action a and $\eta > 0$ define prob. mass fn.

$$p_{a,\eta}(y) := \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y,a)}$$

- A strategy relative to $\mathcal{A}$ is a function $S : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{A}$
- Extend definition to strategies as:

$$p_{S,\eta}(y^n \mid x^n) = \frac{1}{Z(\eta)^n} \cdot e^{-\eta \text{loss}(y^n,S)}$$

Entropification (The Gauss Device)

- accumulated log loss prediction error is affine (linear+constant) function of loss of interest

$$\text{log-loss}(y^n, p_{S,\eta}) = -\log p_{S,\eta}(y^n) = \eta \text{loss}(y^n, S) + n \ln Z(\eta)$$

- accumulated loss difference is even linear:

$$\text{log-loss}(y^n, p_{S,\eta}) - \text{log-loss}(y^n, p_{S',\eta}) = \eta \left(\text{loss}(y^n, S) - \text{loss}(y^n, S') \right)$$

prediction strategy is good for original loss iff transformed prediction strategy is good for log-loss!

Applying "Bayes" to general predictors

- Now consider a "model" of a finite or countably infinite set of (potentially black-box) predictors Θ,identified by their prediction strategies $\{S_\theta : \theta \in \Theta\}$
- Each of these is mapped to a probability distribution p_θ as just explained, such that

$$p_{\theta,\eta}(y^n) = \frac{1}{Z(\eta)^n} e^{-\eta \text{loss}(y^n, S_\theta)}$$

and

$$p_{\theta,\eta}(y_i \mid y^{i-1}) = \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y_i, S_\theta(y^{i-1}))}$$

Applying "Bayes" to general predictors

- Since we have

$$p_{\theta,\eta}(y_i \mid y^{i-1}) = \frac{1}{Z(\eta)} e^{-\eta \text{loss}(y_i, S_\theta(y^{i-1}))}$$

- We now get

$$P_{\text{Bayes},\eta}(y_{i+1} \mid y^i) = \sum_\theta W_\eta(\theta \mid y^i) \cdot \left(\frac{1}{Z(\eta)} e^{-\eta \text{loss}(y_i, S_\theta(y^{i-1}))} \right)$$

- ... with the **generalized posterior**

$$W_\eta(\theta \mid y^i) = \frac{e^{-\eta \text{loss}(y^i, S_\theta)} \cdot W(\theta)}{\sum_{\theta' \in \Theta} e^{-\eta \text{loss}(y^i, S'_\theta)} \cdot W(\theta')}$$

With $\eta = 1$ and log-loss this reduces to standard Bayes posterior

Bayes is good with log-loss

- For all n, y^n, all θ :

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) - \log W(\theta)$$

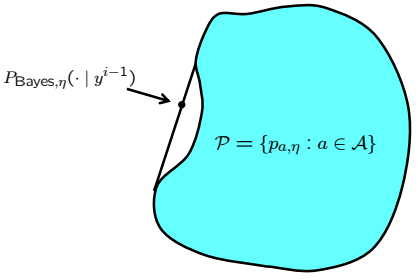
- For all sequences of each length n, regret of Bayes bounded by constant depending on θ, not on n
- For "nonmixable" loss functions like 0/1-loss and absolute loss, this does not work
  - standard Bayes does not perform well at all in worst-case
  - Optimal algorithm gets much larger regret of order $\sqrt{n(-\log W(\theta))}$ in worst-case

WHY??

Answer: Bayes goes beyond model

- Bayesian predictive distribution steps outside "model" $\mathcal{P} = \{p_{a,\eta} : a \in \mathcal{A}\}$: it predicts by mixture of $p_{a,\eta}$
- But if we want a prediction strategy applicable for original loss , we must always predict by p of form $p_{a,\eta} \propto \exp(-\eta \text{loss})$

Bayes goes Beyond Model



Forcing Bayes into the Model

- Bayesian predictive distribution steps outside model : it predicts by mixture of $p_{a,\eta}$
- But if we want a prediction strategy applicable for original loss , we must always predict by p of form $p_{a,\eta} \propto \exp(-\eta \text{loss})$
- We then need some algorithm **A** to turn Bayes posterior into allowed prediction. Examples:
  - Predict by MAP:
    $\text{log-loss}(y_i, \mathbf{map}(W(\cdot \mid y^{i-1}))) := -\log p_{\theta_{\text{map}}}(y_i \mid \theta)$
    where θ_{map} achieves maximum of $W(\theta \mid y^{i-1})$
  - Predict by $p_{a,\eta}$ minimizing posterior expected loss

The Most Important Notion: $\Delta^*_\eta(\mathbf{A})$

- We define the **mixability gap** (G. et al. 2011) to be:
  $$\Delta^*_\eta(\mathbf{A}) = \text{loss}(y^n, \mathbf{A}) - \frac{1}{\eta}\text{log-loss}(y^n, P_{\text{Bayes},\eta})$$
- Interpretation: amount of bits (loss units) lost by being forced to use allowed predictions instead of using the happily mixing Bayes prediction

The Clue

- We then have for all n, y^n, θ :
  $$\text{log-loss}(y^n, P_{\text{Bayes},\eta}) \leq \text{log-loss}(y^n, P_{\theta,\eta}) - \log W(\theta)$$

so

$$\text{log-loss}(y^n, \mathbf{A}) \leq \text{log-loss}(y^n, \theta) - \log W(\theta) + \text{log-loss}(y^n, \mathbf{A}) - \text{log-loss}(y^n, P_{\text{Bayes},\eta})$$

so

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \text{loss}(y^n, \mathbf{A}) - \frac{1}{\eta}\text{log-loss}(y^n, P_{\text{Bayes},\eta})$$

so

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \Delta^*_\eta(\mathbf{A})$$

The Clue

- We then have for all n, y^n, θ :

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \Delta^*_\eta(\mathbf{A})$$

Vovk '90: for so-called "mixable" loss fns, there exists an A (the aggregating algorithm) such that for some $\eta > 0$, we are guaranteed $\Delta^*_\eta(\mathbf{A}) \leq 0$ (hence name mixability gap)

The Clue

- We then have for all n, y^n, θ :

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \Delta^*_\eta(\mathbf{A})$$

Vovk '90: for so-called "mixable" loss fns, there exists an A (the aggregating algorithm) such that for some $\eta > 0$, we are guaranteed $\Delta^*_\eta(\mathbf{A}) \leq 0$ (hence name mixability gap)

Example: squared loss
$\text{loss}(y, a) = (y - a)^2$ is ½- mixable if $\mathcal{Y} = [-1, 1]$

Mixable Loss Functions

- All 'strictly convex' loss functions with bounded range are mixable for finite $\eta > 0$
- So we can still run generalized Bayes and get regret bounds that are still of order $O(-\log W(\theta))$
- Usually $\eta < 1$:
  prior becomes more, data less important
- Modifying Bayes may seem like 'hack' (relation to Bayes' theorem is lost) but: resulting procedure still minimax optimal up to a constant (and in practice preferable over minimax algorithm)
- Something deep going on...

The Clue, now for nonmixable loss

- We then have for all n, y^n, θ :

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \Delta^*_\eta(\mathbf{A})$$

0/1-loss and absolute loss are nonmixable

The Clue

- We then have for all n, y^n, θ :

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \Delta^*_\eta(\mathbf{A})$$

Hedge-style algorithms: for bounded nonmixable losses functions: there still exist **A** such that $\Delta^*_\eta(\mathbf{A}) \leq \eta \cdot n$
For uniform prior over K predictors, optimizing over η gives $\eta = \tilde{O}\left(\sqrt{\frac{\log K}{n}}\right)$
We get regret bound (and actual regret) $O\left(\sqrt{n \cdot (\log K)}\right)$
(Warmuth, Freund, Schapire)

The Clue

- We then have for all n, y^n, θ :

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \Delta^*_\eta(\mathbf{A})$$

Hedge-style algorithms: for bounded nonmixable losses functions: there still exist **A** such that $\Delta^*_\eta(\mathbf{A}) \leq \eta \cdot n$
For uniform prior over K predictors, optimizing over η gives $\eta = \tilde{O}\left(\sqrt{\frac{\log K}{n}}\right)$
We get regret bound (and actual regret) $O\left(\sqrt{n \cdot (\log K)}\right)$
(Warmuth, Freund, Schapire)

The Hedge Algorithm

- Basic Algorithm requires knowledge of 'horizon' n : at every time point $t < n$, you use generalized posterior with $\eta = \tilde{O}\left(\sqrt{\frac{\log K}{n}}\right)$
- If n unknown, can still achieve same bound by decreasing learning rate dynamically: at each time t, you use posterior weights you get if you had used $\eta = \tilde{O}\left(\sqrt{\frac{\log K}{t}}\right)$ from time 1 to t

"The older you get, the less attention you pay to all the experiences you had in life!"

Varying the Setting

- Predicting better than the Best Expert
  - As good as the best convex combination
  - As good as the best sequence of experts
  - Applied to Prediction of Electricity Consumption in Greater Paris Region by Electricité de France (Devaine, Goude, Stoltz 2012)
- Hedge with Infinitely Many Experts
- **Bandit** and Other Limited Feedback Settings

Google Ads

Major Open Problem(s)

Learning the Learning Rate