E0 370 Statistical Learning Theory **Lecture 20 (Nov 17, 2011)**

Online Learning from Experts: Weighed Majority and Hedge

Lecturer: Shivani Agarwal *Scribe: Saradha R*

1 Introduction

In this lecture, we will look at the problem of learning from multiple experts in an online fashion. There are finite number of experts, who give their predictions $\xi_1, \ldots, \xi_N$. The learning algorithm has to use the predictor values and come up with an outcome $\hat{y}$. The total number of mistakes made by the algorithm is compared with the performance of the best expert in consideration.

2 Online Prediction from Experts

A general online prediction problem proceeds as follows.

Online (binary) prediction using multiple experts

For $t = 1, \ldots, T$:
- Receive instance $x^t \in \mathcal{X}$
- Receive expert predictors $\xi_1(x^t), \ldots, \xi_N(x^t) \in \{\pm 1\}$
- Predict $\hat{y}^t \in \{\pm 1\}$
- Receive true label $y^t \in \{\pm 1\}$
- Incur loss $\ell(y^t, \hat{y}^t)$

2.1 Halving Algorithm

Here we assume that the set of experts that we consider has an expert which would give the correct label for all instances. In the halving algorithm, for every iteration, only the consistent experts are retained. If a predictor makes a mistake it will no more be contributing in the prediction process.

Halving Algorithm

Initiate weights $w_i^1 = 1 \forall i \in [N]$

For $t = 1, \ldots, T$:
- Receive instance $x^t \in \mathcal{X}$
- Receive expert predictors $\xi_1(x^t), \ldots, \xi_N(x^t) \in \{\pm 1\}$
- Predict $\hat{y}^t = sign(\sum_{j=1}^n w_j^t . \xi_j^t)$ (majority vote)
- Receive true label $y^t \in \{\pm 1\}$
- Incur loss $\ell(y^t, \hat{y}^t)$
- Update:-
- Update:- $\forall i \in 1 \ldots N : If \xi_i^t \neq y^t$ then
  $\mathbf{w}_i^{t+1} \leftarrow 0$
  else
  $\mathbf{w}_i^{t+1} \leftarrow \mathbf{w}_i^t$

Thus the maximum number of mistakes, or the sum of loss over any given sequence is bounded by the logarithm of number of predictors. i.e.

$$L_S^{0\text{-}1}[Halving] \;\leq\; \log_2 N\,.$$

2.2 Weighted Majority (WM) Algorithm

In the halving algorithm, when a predictor makes even one mistake, it will not be able to contribute to the prediction in the successive iterations. When we don't have an expert that would predict correctly for all samples, this would not be a suitable approach. The weighted majority algorithm works well in such situations. Here every predictor is assigned equal weight, say 1, initially. Later as they make binary predictions on instances, the weights of the predictors are decreased using multiplicative update, when they commit mistakes. The rate at which the weights are updated is governed by the parameter η.

Weighted majority Algorithm

Initiate weights $w_i^1 = 1 \forall i \in [N]$
Choose parameter $\eta > 0$ that would be used in the weight update rule.

For $t = 1, \ldots, T$:
- Receive instance $x^t \in \mathcal{X}$
- Receive expert predictors $\xi_1(x^t), \ldots, \xi_N(x^t) \in \{\pm 1\}$
- Predict $\hat{y}^t = sign(\sum_{j=1}^n w_j^t . \xi_j^t)$ (majority vote)
- Receive true label $y^t \in \{\pm 1\}$
- Incur loss $\ell(y^t, \hat{y}^t)$
- Update:- If $\hat{y}^t \neq y^t$
  $\forall i \in 1 \ldots N$
  $\mathbf{w}_i^{t+1} \leftarrow \mathbf{w}_i^t \exp^{(\eta . \mathbf{I}(y^t \neq \xi_i^t))}$

Theorem 2.1. Let $\xi_1, \ldots, \xi_N \in \{\pm 1\}^T$. Let $S = (y^1, \ldots, y^T) \in \{\pm 1\}$ and let $\eta > 0$.
Then the total number of mistakes

$$L_S^{0\text{-}1}[WeightedMajority(\eta)] \;\leq\; \left(\frac{\eta}{\ln(\frac{2}{1+\exp(-\eta)})} \right) . \min_i L_S^{0\text{-}1}[\xi_i] + \frac{1}{\ln(\frac{2}{1+\exp(-\eta)})} . \ln(N)\,.$$

Proof. Denote $L_S^{0\text{-}1}[WeightedMajority] = L$
For each trial t on which there is a mistake, we have

$$W^{t+1} \;=\; \sum_{i=1}^N \mathbf{w}_i^{t+1} = \sum_{i=1}^N w_i^t . \exp^{(-\eta . I(y^t \neq \xi_i^t))}\,. \tag{1}$$

$$= \sum_{i:y^t \neq \xi_i^t} w_i^t . \exp^{-\eta} + \sum_{i:y^t = \xi_i^t} w_i^t \tag{2}$$

$$= \exp^{-\eta} . W_{maj} + W_{min} \tag{3}$$

$$\leq \exp^{-\eta} . W_{maj} + W_{min} + \frac{1 - \exp^{-\eta}}{2}(W_{maj} - W_{min}) \tag{4}$$

$$= \frac{1 + exp^{-\eta}}{2} . (W_{maj} + W_{min}) \;=\; \frac{1 + exp^{-\eta}}{2} . (W^t) \tag{5}$$

For all mistake trials t, we have $\frac{W^{t+1}}{W^t} \leq \frac{1+exp^{-\eta}}{2}$. For other trials, $\frac{W^{t+1}}{W^t} \leq 1$
Therefore summing over $t = 1, \ldots, T$ gives

$$\frac{W^{T+1}}{W^T} \;\leq\; (\frac{1 + exp^{-\eta}}{2})^L\,. \tag{6}$$

Taking logarithm, we get

$$L \;\leq\; \frac{\ln W^1 - \ln W^{T+1}}{\ln(\frac{2}{1+\exp^(-\eta)})}\,. \tag{7}$$

Finding the lower bound on $\ln W^{T+1}$

$$W^{t+1} \;=\; \sum_{j=1}^{N} w_j^{T+1} \;\geq\; w_j^{T+1} \;\geq\; \exp^{-\eta.L_i}\, w_i^1 (\forall i)\,. \tag{8}$$

$$L \;\leq\; \frac{\ln W^1 - \eta.L_i - \ln w_i^1}{\ln(\frac{2}{1+e^(-\eta)})} \;=\; \frac{\ln N + \eta.L_i}{\ln(\frac{2}{1+e^(-\eta)})} \tag{9}$$

$$\tag{10}$$

for all $w_j^1 > 0 \forall j$

Thus we obtain the result. □

2.3 Weighted Majority: Continuous Version (WMC)

We now see the continuous version of weighted majority algorithm. The final prediction is a weighted average of the expert predictor values.

Here $\tilde{y} = \hat{y} = y = [0, 1]$

Weighted majority Algorithm :Continuous Version (WMC)

Initiate weights $w_i^1 = 1 \forall i \in [N]$
Choose parameter $\eta > 0$ that would be used in the weight update rule.

For $t = 1, \ldots, T$:
- Receive instance $x^t \in \mathcal{X}$
- Receive expert predictors $\xi_1(x^t), \ldots, \xi_N(x^t) \in [0, 1]$
- Predict $\hat{y}^t = \frac{\sum_{i=1}^{N} w_i^t . \xi_i^t}{\sum_{i=1}^{N} w_i^t} \in [0, 1]$ (Weighted Average)
- Receive true label $y^t \in [0, 1]$
- Incur loss $\ell_{abs}(y^t, \hat{y}^t) = |y^t - \hat{y}^t|$
- Update:- $\forall i \in 1 \ldots N$
  $\mathbf{w}_i^{t+1} \leftarrow \mathbf{w}_i^t . \exp^{-\eta . |\xi_i^t - y^t|}$

Theorem 2.2. Let $\xi_1, \ldots, \xi_N \in [0, 1]^T$. Let $S = (y^1, \ldots, y^T) \in [0, 1]$ and let $\eta > 0$.
Then the total number of mistakes

$$L_S^{abs}[WMC(\eta)] \;\leq\; (\frac{\eta}{1 - \exp^{-\eta}}) . \min_i L_S^{abs}[\xi_i] + \frac{1}{1 - \exp^{-\eta}} . \ln(N)\,.$$

Proof. Denote $L_S^{abs}[WMC(\eta)] = L$
For each trial t we have

$$W^{t+1} \;=\; \sum_{i=1}^{N} \mathbf{w}_i^{t+1} = \sum_{i=1}^{N} w_i^t . \exp^{-\eta . |y^t - \xi_i^t|}\,. \tag{11}$$

$$\leq\; \sum_{i=1}^{N} w_i^t . [1 - (1 - \exp^{-\eta}) |y^t - \xi_i^t|]\,. \tag{12}$$

$$W^{t+1} \leq \sum_{i=1}^{N} w_i^t \left[1 - (1 - \exp^{-\eta}) \frac{\sum_{i=1}^{N} w_i^t |y^t - \xi_i^t|}{\sum_{i=1}^{N} w_i^t}\right] \tag{13}$$

$$\leq W^t . \left[1 - (1 - \exp^{-\eta}) \left|\frac{\sum_{i=1}^{N} w_i^t |y^t - \xi_i^t|}{\sum_{i=1}^{N} w_i^t}\right|\right] \tag{14}$$

$$= W^t . \left[1 - (1 - \exp^{-\eta}) |\hat{y}^t - y^t|\right] \tag{15}$$

$$W^{t+1} \leq W^t . \left[\exp^{-(1-\exp^{-\eta}).|\hat{y}^t - y^t|}\right] . \tag{16}$$

$$\frac{W^{t+1}}{W^t} \leq \left[\exp^{-(1-\exp^{-\eta}).|\hat{y}^t - y^t|}\right] \tag{17}$$

$$\leq \left[\exp^{-(1-\exp^{-\eta}).\sum_{t=1}^{T} |\hat{y}^t - y^t|}\right] = \exp^{-(1-\exp^{-\eta}).L} \tag{18}$$

Taking logarithm, we get

$$L \leq \frac{\ln W^1 - \ln W^{T+1}}{1 - (\exp^{-\eta})} . \tag{19}$$

Finding the lower bound on $\ln W^{T+1}$

$$W^{t+1} \geq \exp^{-\eta . L_i} w_i^1 (\forall i) . \tag{20}$$

Thus we obtain the result

$$L \leq \frac{\ln W^{T+1} + \eta . L_i - \ln w_i^1}{1 - (\exp^{-\eta})} \tag{21}$$

$$\leq \frac{\ln N + \eta . L_i}{1 - (\exp^{-\eta})} . \tag{22}$$

□

3 Online Allocation

The problem of online allocation occurs in scenarios where we need to allocate different fraction of resources into N different options. The loss associated with every option is available at the end of every iteration. We would like to reduce the total loss suffered for the particular allocation. The allocation for the next iteration is then revised, based on the total loss suffered in the current iteration using multiplicative update.

Hedge Algorithm(η)

Initiate weights $w_i^1 = 1 \forall i \in [N]$

Choose parameter $\eta > 0$ that would be used in the weight update rule.

For $t = 1, \ldots, T$:

- Make allocation $p^t \in \triangle_N$
- $p^t = \frac{w^t}{\sum_{i=1}^{N} w_i^t}$;
- Receive vector of loses $\ell^t = (\ell_1^t, \ldots, \ell_N^t) \in [0, 1]^N$
- Incur loss $p^t . \ell^t = \sum_{i=1}^{N} p_i^t . \ell_i^t$
- Update:- $\forall i \in 1 \ldots N$

  $\mathbf{w}_i^{t+1} \leftarrow \mathbf{w}_i^t . \exp^{-\eta(\ell_i^t)}$

Theorem 3.1. Let $\ell^1, \ldots, \ell^T \in [0,1]^N$ The cumulative loss of the algorithm is

$$L[A] \;=\; \sum_{i=1}^{T} p^t.\ell^t$$

If the loss of a particular option over the T iterations is given by

$$L_i \;=\; \sum_{i=1}^{T} \ell_i^t\,.$$

Then

$$L[Hedge(\eta)] \;\leq\; (\frac{\eta}{1-\exp^{-\eta}}).\min_i L_i + \frac{1}{1-\exp^{-\eta}}.\ln(N)\,.$$

Proof. Denote $L[Hedge(\eta)] = L$
For each trial t we have

$$W^{t+1} \;=\; \sum_{i=1}^{N} \mathbf{w}_i^{t+1} = \sum_{i=1}^{N} w_i^t.\exp^{-\eta.\ell_i^t}\;. \tag{23}$$

$$\leq\; \sum_{i=1}^{N} w_i^t.\left[1-(1-\exp^{-\eta}).\frac{\sum_{i=1}^{N} w_i^t.\ell_i^t}{\sum_{i=1}^{N} w_i^t}\right]. \tag{24}$$

$$W^{t+1} \;\leq\; \sum_{i=1}^{N} w_i^t\left[1-(1-\exp^{-\eta})p^t.\ell^t\right]. \tag{25}$$

$$\frac{W^{t+1}}{W^t} \;\leq\; \left[\exp^{-(1-\exp^{-\eta}).p^t.\ell^t}\right] \tag{26}$$

$$\leq\; \left[\exp^{-(1-\exp^{-\eta}).L}\right] \tag{27}$$

Taking logarithm, we get

$$L \;\leq\; \frac{\ln W^1 - \ln W^{T+1}}{1-(\exp^{-\eta})}\,. \tag{28}$$

Finding the lower bound on $\ln W^{T+1}$

$$W^{t+1} \;\geq\; \exp^{-\eta.L_i} w_i^1(\forall i)\,. \tag{29}$$

Thus we obtain the result

$$L \;\leq\; \frac{\ln W^{T+1} + \eta.L_i - \ln w_i^1}{1-(\exp^{-\eta})} \tag{30}$$

$$\leq\; \frac{\ln N + \eta.L_i}{1-(\exp^{-\eta})}\,. \tag{31}$$

□

4 Next Lecture

In the next lecture, we will introduce the idea of minimax regret, in an adversarial learning setting.

References