E0 370 Statistical Learning Theory **Lecture 18 (Oct 17, 2013)**

Online Prediction from/Allocation among Experts

Lecturer: Shivani Agarwal *Scribe: Avinash M*

1 Introduction

In this lecture, we look at the scenario of Online Learning in the presence of 'Experts'. Specifically, we study methods to predict labels based on the predicitons of these experts and methods to allocate resources among a group of experts. Such problems naturally arise, for example, in scenarios like mutual fund investments.

The lecture is divided into two parts, the first part dealing with the case of online prediction from experts and the second with the online decision/allocation problem.

2 Online Prediction from Expert Advice

We assume that the set of experts is finite ($n < \infty$ experts) and denote their predictions by $\{\mathcal{E}_1^t, \mathcal{E}_2^t, \ldots, \mathcal{E}_n^t\}$, where the superscript represents the time instant at which the prediction is obtained. A general online prediction problem is described in what follows.

Online Prediction from Experts
For t=1:T
Receive predictions $\mathcal{E}^t = (\mathcal{E}_1^t, \mathcal{E}_2^t, \ldots, \mathcal{E}_n^t) \in \hat{\mathcal{Y}}^n$
Predict $\hat{y}^t \in \hat{\mathcal{Y}}$
Receive true label $y^t \in \mathcal{Y}$
Incur Loss $l(y^t, \hat{y}^t)$

The criterion of performance is the total loss accumulated by the algorithm, as compared with the performance of the best (again in the cumulative loss sense) expert. To this end, we define two losses for any given sequence $S = \left((\mathcal{E}^1, y^1), (\mathcal{E}^2, y^2), \ldots, (\mathcal{E}^T, y^T)\right)$:

$$L_S[A] = \sum_{t=1}^{T} l(y_t, \hat{y}^t) \tag{1}$$

$$L_S[\mathcal{E}_i] = \sum_{t=1}^{T} l(y_t, \mathcal{E}_i^t) \tag{2}$$

Here $L_S[A]$ is the cumulative loss of the algorithm (A) which we shall also denote by L, and $L_S[\mathcal{E}_i]$ is the i^{th} expert's cumulative loss which we shall also denote by L_i.

2.1 Binary Classification: Halving Algorithm

We now consider a more specific scenario where $\mathcal{Y} = \hat{\mathcal{Y}} = \{\pm 1\}$ and the loss function is l_{0-1}.

Suppose one of the experts always gives the correct prediction. In this case, one of the ways to find the said expert would be to discard experts making wrong predictions until we are left with just one expert, who will then necessarily have to be the correct expert. The algorithm proceeds as follows:

Algorithm **Halving Algorithm**

Initialization: $w_i^1 \leftarrow 1, \forall i \in [n]$
For t=1:T
Receive predictions $\mathcal{E}^t = (\mathcal{E}_1^t, \mathcal{E}_2^t, \ldots, \mathcal{E}_n^t) \in \{\pm 1\}^n$
Predict $\hat{y}^t = \text{sgn}(\sum_{i=1}^n w_i^t \mathcal{E}_i^t)$
Receive true label y^t
Incur Loss $l_{0-1}(y^t, \hat{y}^t)$
Update:
$\forall i \in [n]$: If $(\mathcal{E}_i^t \neq y^t)$ then $w_i^{t+1} \leftarrow 0$
else $w_i^{t+1} \leftarrow w_i^t$

Since our predictions are based on a majority rule, at each time-step, for our prediction to be wrong, at least half the experts have to be wrong. This results in the discarding of at least half the experts during such a step. So,

$$L_S^{0-1}[Halving] \leq \log_2 n \tag{3}$$

2.2 Binary Classification: Weighted Majority 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 (Vovk, 1990; Littlestone and Warmuth, 1994) works well in such situations. Here every predictor is assigned equal weight, say 1, initially. As time progresses, when the experts commit mistakes, the weights of the predictors are decreased using a multiplicative update. The rate at which the weights are updated is governed by the parameter η.

Algorithm **Weighted Majority Algorithm**

Initialize: $w_i^1 \leftarrow 1 \forall i \in [n]$
$\eta \in (0, 1)$.
For t=1:T
Receive predictions $\mathcal{E}^t = (\mathcal{E}_1^t, \mathcal{E}_2^t, \ldots, \mathcal{E}_n^t) \in \{\pm 1\}^n$
Predict $\hat{y}^t = \text{sgn}(\sum_{i=1}^n w_i^t \mathcal{E}_i^t)$
Receive true label y^t
Incur Loss $l_{0-1}(y^t, \hat{y}^t)$
Update:
$\forall i \in [n] : w_i^{t+1} \leftarrow w_i^t e^{-\eta l_{0-1}(y^t, \mathcal{E}_i^t)} = e^{-\eta \sum_{s=1}^t l_{0-1}(y^s, \mathcal{E}_i^s)}$

The cumulative $0-1$ loss with this algorithm (as defined in (1)) can be bounded as follows.

Theorem 2.1. Let $S = ((\mathcal{E}^1, y^1), (\mathcal{E}^2, y^2), \ldots (\mathcal{E}^T, y^T)) \in (\{\pm 1\}^n \times \{\pm 1\})^T$. The cumulative $0-1$ loss of the weighted majority algorithm can be bounded by:

$$L_S^{0-1}[WM(\eta)] \leq \frac{\eta}{\ln(\frac{2}{1+e^{-\eta}})} \min_{i \in [n]} L_S^{0-1}[\mathcal{E}_i] + \frac{\ln n}{\ln(\frac{2}{1+e^{-\eta}})} \tag{4}$$

Proof. Define $W^t = \sum_{i=1}^n w_i^t$, noting that W^t decreases with t since each weight either remains the same if the corresponding expert's prediction is correct or decreases.
For each trial t on which $\hat{y}^t \neq y^t$:

$$W^{t+1} = \sum_{i=1}^{n} w_i^{t+1} = \sum_{i=1}^{n} w_i^t e^{-\eta \mathbf{I}_{\{\mathcal{E}_i^t \neq y^t\}}} \tag{5}$$

$$= e^{-\eta} \sum_{i=1}^{n} w_i^t \mathbf{I}_{\{\mathcal{E}_i^t \neq y^t\}} + \sum_{i=1}^{n} w_i^t \mathbf{I}_{\{\mathcal{E}_i^t = y^t\}} \tag{6}$$

$$= e^{-\eta} W_{maj}^t + W_{min}^t \tag{7}$$

$$\leq e^{-\eta} W_{maj}^t + W_{min}^t + \frac{1+e^{-\eta}}{2}(W_{maj}^t + W_{min}^t) \tag{8}$$

$$= \frac{1+e^{-\eta}}{2} W^t \tag{9}$$

Thus, for 'mistake'-trials, $\frac{W^{t+1}}{W^t} \leq \frac{1+e^{-\eta}}{2}$, and $\frac{W^{t+1}}{W^t} \leq 1$ for others. Hence, multiplying over t=1:T, we have,

$$\frac{W^{T+1}}{W^1} \leq \left(\frac{1+e^{-\eta}}{2}\right)^L \tag{10}$$

$$\tag{11}$$

also,

$$\frac{W^{T+1}}{W^1} = \frac{\sum_{i=1}^{n} w_i^{T+1}}{W^1} \geq \max_{i\in[n]} \frac{w_i^{T+1}}{n} \tag{12}$$

$$= \max_{i\in[n]} \frac{e^{-\eta \sum_{s=1}^{T} \mathbf{I}_{\{\mathcal{E}_i^s \neq y^s\}}}}{n} \tag{13}$$

$$= \max_{i\in[n]} \frac{e^{-\eta L_i}}{n} = \frac{e^{-\eta \min_{i\in[n]} L_i}}{n} \tag{14}$$

Hence, $-\eta \min_{i\in[n]} L_i - \ln n \leq L \ln\left(\frac{1+e^{-\eta}}{2}\right)$. Rearrange to get the claim. □

2.3 Weighted Average/Exponential Weighting Algorithm

So far we considered a binary classification setting, where the experts and algorithms made binary predictions. We now generalize the weighted majority algorithm to predictions in any convex set $\hat{\mathcal{Y}}$ (Vovk, 1990; Littlestone and Warmuth, 1994).

Algorithm **Weighted Average Algorithm**

Initialize: $w_i^1 \leftarrow 1 \forall i \in [n]$

Parameter $\eta > 0$

For t=1:T

Receive predictions $\mathcal{E}^t = (\mathcal{E}_1^t, \mathcal{E}_2^t, \ldots, \mathcal{E}_n^t) \in \hat{\mathcal{Y}}^n$

Predict $\hat{y}^t = \sum_{i=1}^{n} p_i^t \mathcal{E}_i^t$ where $p_i^t = \frac{w_i^t}{\sum_{j=1}^{n} w_j^t}$. Note that $\hat{y}^t \in \hat{\mathcal{Y}}$ due to convexity.

Receive true label y^t

Incur Loss $l(y^t, \hat{y}^t)$

Update:

$$\forall i \in [n] : w_i^{t+1} \leftarrow w_i^t e^{-\eta l(y^t, \mathcal{E}_i^t)} = e^{-\eta \sum_{s=1}^{t} l(y^s, \mathcal{E}_i^s)}$$

Theorem 2.2. Let $S = ((\mathcal{E}^1, y^1), (\mathcal{E}^2, y^2), \ldots (\mathcal{E}^T, y^T)) \in (\{\hat{\mathcal{Y}}\}^n \times \{\mathcal{Y}\})^T$. Let $l : \mathcal{Y} \times \hat{\mathcal{Y}} \mapsto \mathbb{R}_+$ be bounded and convex in the second argument, let $l(y, \hat{y}) \in [0, 1] \forall y, \hat{y}$. Then:

$$L_S^l[WA(\eta)] - \min_{i\in[n]} L_S^l[\mathcal{E}_i] \leq \frac{\ln n}{\eta} + \frac{T\eta}{8}. \tag{15}$$

In particular, if T is known in advance, then setting $\eta^* = 2\sqrt{\frac{2\ln n}{T}}$ to minimize the above bound yields

$$L_S^l[WA(\eta^*)] - \min_{i\in[n]} L_S^l[\mathcal{E}_i] \leq \sqrt{\frac{T\ln n}{2}}, \tag{16}$$

which is sublinear in T.

PROOF. Let $W^{t+1} = \sum_{i=1}^n w_i^{t+1}$. Then,

$$\ln\left(\frac{W^{t+1}}{W^t}\right) = \ln\left(\frac{\sum_{i=1}^n w_i^{t+1}}{W^t}\right) = \ln\left(\sum_{i=1}^n p_i^t e^{-\eta l(y^t,\mathcal{E}_i^t)}\right) \tag{17}$$

$$= \ln\left(\mathbb{E}_I\left[e^{-\eta l(y^t,\mathcal{E}_I^t)}\right]\right) \tag{18}$$

$$\overset{Hoeffding}{\leq} \ln\left(e^{\frac{\eta^2}{8}} e^{\eta\mathbb{E}_I[l(y^t,\mathcal{E}_I^t)]}\right) \tag{19}$$

$$\Rightarrow \ln\left(\frac{W^{t+1}}{W^t}\right) \leq \frac{\eta^2}{8} - \eta\mathbb{E}_I[l(y^t,\mathcal{E}_I^t)] \tag{20}$$

$$\overset{Jensen}{\leq} \frac{\eta^2}{8} - \eta l(y^t, \mathbb{E}_I[\mathcal{E}_I^t]). \tag{21}$$

(Here, I denotes a random variable taking values $i \in [n]$ with probability p_i^t).Recognizing that $\mathbb{E}_I[\mathcal{E}_I^t] = \hat{y}^t$, and accumulating over t=1:T,

$$\ln\left(\frac{W^{T+1}}{W^1}\right) \leq \frac{T\eta^2}{8} - \eta L \tag{22}$$

$$\tag{23}$$

also,

$$\ln\left(\frac{W^{t+1}}{W^t}\right) = \ln(\sum_{i=1}^n w_i^{T+1}) - \ln n \tag{24}$$

$$\geq \ln\left(\max_{i\in[n]} w_i^{T+1}\right) - \ln n \tag{25}$$

$$= -\eta \min_{i\in[n]} L_i - \ln \tag{26}$$

Combining the upper and lower bounds on $\ln\left(\frac{W^{T+1}}{W^1}\right)$ above, yields the desired result. □

3 Online Allocation among Experts

The problem of online allocation occurs in scenarios like mutual fund investments where we need to allocate different fractions of resources among n different options. The following algorithm was proposed in (Freund and Schapire, 1997).

Algorithm **The Hedge Algorithm**

Initialize: $w_i^1 \leftarrow 1$ *and* $p_i^1 \leftarrow 1/n, \forall i \in [n]$
Parameter $\eta > 0$
For t=1:T
  Allocate $p^t \in \Delta_n$
  Receive Loss Vector $l^t = (l_1^t, l_2^t, \ldots, l_n^t) \in \mathbb{R}_+^n$
  Incur Loss $(p^t)^T l^t = \sum_{i=1}^n p_i^t l_i^t$
  Update:
    $\forall i \in [n] : w_i^{t+1} \leftarrow w_i^t e^{-\eta l_i^t}$
    $p_i^{t+1} \leftarrow \frac{w_i^{t+1}}{\sum_{k=1}^n w_k^{t+1}}$

Define the algorithm's and the experts' losses respectively as:

$$L[A] = \sum_{t=1}^{T} (p^t)^T l^t \tag{27}$$

$$L[\mathcal{E}_i] = \sum_{t=1}^{T} l_i^t. \tag{28}$$

Comparing with the previous result, we see that taking $\hat{\mathcal{Y}} = \Delta_n$ to be the convex set of predictions, $\mathcal{Y} = \mathbb{R}^n$ to be the 'label' space (consisting of loss vectors) and $l : \mathcal{Y} \times \hat{\mathcal{Y}} \mapsto \mathbb{R}_+$ to be the loss, defined by $l(\lambda, p) = \lambda^T p$, which is certainly convex in the second argument we get the following result for bounded loss vectors.

Theorem 3.1. Assume losses are bounded; say $l_i^t \in [0, 1] \forall$ i,t. Then

$$L[Hedge(\eta)] - \min_{i \in [n]} L_i \leq \frac{\ln n}{\eta} + \frac{Tn}{8}. \tag{29}$$

For $\eta^* = 2\sqrt{\frac{2 \ln n}{T}}$

$$L[Hedge(\eta^*)] - \min_{i \in [n]} L_i \leq \sqrt{\frac{T \ln n}{2}}. \tag{30}$$

Finally, one also has the following matching lower bound, the proof of which we omit:

Theorem 3.2. For any algorithm A for the online allocation problem,

$$\sup_{T,n} \sup_{(l^1,\ldots,l^T) \in ([0,1])^n} \frac{L[A] - \min_{i \in [n]} L_i}{\sqrt{\frac{T \ln n}{2}}} \geq 1 \tag{31}$$

4 Next Lecture

The next lecture will focus on Online Convex Optimization.

5 References

1. (Freund and Schapire, 1997) Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of Computer and System Sciences, 55(1):119–139, (1997).

2. (Littlestone and Warmuth, 1994) N. Littlestone and M.K. Warmuth, "The Weighted Majority Algorithm", Information and Computation, 108(2):212 - 261, 1994.

3. (Vovk, 1990) V. Vovk, "Aggregating strategies", Proceedings of the Third Annual Workshop on Computational Learning Theory. Fulk, M. and Case, J. (eds.). San Mateo, CA: Morgan Kaufmann, p. 371-383 (1990).