E0 370 Statistical Learning Theory **Lecture 21 (Nov 24, 2011)**

Online Learning from Experts: Minimax Regret

Lecturer: Shivani Agarwal *Scribe: Nikhil Vidhani*

1 Introduction

In the last three lectures we have been discussing the online learning algorithms where we receive the instance x^t and then its label y^t for $t = 1, ..., T$. Specifically in the last lecture we talked about online learning from experts and online prediction. We saw many algorithms like Halving algorithm, Weighted Majority (WM) algorithm and lastly Weighted Majority Continuous (WMC) algorithm. We also saw bounds on the cumulative loss incurred by these algorithms. Today, we will focus on online prediction. For the WMC algorithm the setting is: we have N experts who predict the outcome (label) in [0,1], then we combine these predictions using a weighted average of these. Then we receive the true label and incur some loss (absolute loss in our setting), then we make an update to the weight vectors based on the loss. The intuition is that the higher the loss incurred by an expert, the more drastically we reduce its weight. For the WMC algorithm, we proved that:

$$L_S^{\text{abs}}[WMC(\eta)] \leq \frac{\ln(N)}{1 - e^{-\eta}} + \frac{\eta \min_i L_S^{\text{abs}}[\xi_i]}{1 - e^{-\eta}}$$

where the symbols have their meanings as explained in the last lecture. Also if the number of examples T is known in advance, then we can choose the WMC parameter η^* s.t.

$$L_S^{\text{abs}}[WMC(\eta^*)] \leq \min_i L_S^{\text{abs}}[\xi_i] + \sqrt{2T \ln(N)} + \ln(N)$$

2 Minimax Regret

Lets define the regret of an algorithm $\mathcal{A}$ w.r.t. the set of experts $\xi = \{\xi_1, ..., \xi_N\}$ on the sample $S = ((x^1, y^1), ..., (x^T, y^T))$ as:

$$\mathcal{R}_{\xi,S}^{\ell}[\mathcal{A}] = L_S^{\ell}[\mathcal{A}] - \min_i L_S^{\ell}[\xi_i]$$

i.e. the difference between the total loss of your algorithm and the total loss of the best expert. Thus if we had chosen the best expert always for our prediction then our regret would be zero; this makes sense as in online setting absolute loss is insignificant as the adversary may always choose the label other than what we predicted.

The goal of the predictor is to minimize the regret while the goal of adversary is to maximize it. In general, we can formalize it as a game where at each trial you receive expert's predictions, then you make your own prediction and finally the true label is revealed by the adversary. In this setting we want to get close to the best possible loss (minimum regret). Thus we want to choose an algorithm that minimizes the total regret, however we do not have any control over the set of expert predictions and the sample as adversary can choose a set of experts or sample which gives a larger regret over the algorithm. Here, we can define the minimax value (regret) of the game as:

$$V_{N,T}^{\ell} = \min_{\mathcal{A}} \max_{\xi,S} \mathcal{R}_{\xi,S}^{\ell}[\mathcal{A}]$$

It is the worst case guarantee on the regret over all the sequences of labels and expert predictions. From the last lecture we know that:

$$V_{N,T}^{\ell_{\text{abs}}} \leq \sqrt{2T \ln(N)} + \ln(N)$$

A natural question here is can we bound the minimax regret for other loss function? Is this the best possible minimax regret?

3 Minimax regret for various loss functions

In this section we will look at upper bound on minimax regret for some other loss functions which we have seen before. We will focus on binary outcomes $y^t \in \{0,1\}$ with experts predicting in [0,1]; $\xi_i^t \in [0,1]$ and $\hat{y}^t \in [0,1]$. ξ_i^t can be seen as probability of i^{th} expert predicting the label as 1. Define the loss function

$$\ell : \{0,1\} \times [0,1] \to [0,\infty)$$

Lets us also define $\ell_0(\hat{y}) = \ell(0,\hat{y})$ and $\ell_1(\hat{y}) = \ell(1,\hat{y})$.

3.1 Squared Loss

For squared loss $\ell(y,\hat{y}) = (y-\hat{y})^2$, it can be easily seen that $\ell_0(\hat{y}) = \hat{y}^2$ and $\ell_1(\hat{y}) = (1-\hat{y})^2$. Let us define some new quantities (assuming $\ell_0()$ and $\ell_1()$ being twice differentiable):

$$S_\ell(\hat{y}) = \ell_0'(\hat{y}) \cdot \ell_1''(\hat{y}) - \ell_1'(\hat{y}) \cdot \ell_0''(\hat{y})$$

$$R_\ell(\hat{y}) = \frac{\ell_0'(\hat{y}) \cdot \ell_1'(\hat{y})^2 - \ell_1'(\hat{y}) \cdot \ell_0'(\hat{y})^2}{S_\ell(\hat{y})}$$

$$c_\ell = \sup_{0<\hat{y}<1} R_\ell(\hat{y})$$

Theorem 3.1. (Haussler et al., 1998)
Let ℓ be such that $\ell_0(0) = \ell_1(1) = 0$. Let $\ell_0()$ be strictly increasing in (0,1) and $\ell_1()$ be strictly decreasing in (0,1). Let $\ell_0()$ and $\ell_1()$ be three times differentiable in (0,1). If $S_\ell(\hat{y}) > 0$ in (0,1) and $c_\ell < \infty$, then:

$$V_{N,T}^\ell = \Theta(\ln(N))$$

Specifically, we have

$$V_{N,T}^\ell \le c_\ell \ln(N) \quad \text{and} \quad V_{N,T}^\ell \ge (c_\ell - o(1)) \ln(N)$$

For the squared loss:

$$\ell_0'(\hat{y}) = 2\hat{y},\ \ell_1'(\hat{y}) = 2(\hat{y}-1),\ \ell_0''(\hat{y}) = 2,\ \ell_1''(\hat{y}) = 2$$

Thus, $S_\ell(\hat{y}) = 4 > 0$, and $R_\ell(\hat{y}) = 2\hat{y}(1-\hat{y})$. Clearly, $\sup_{0<\hat{y}<1} R_\ell(\hat{y}) = 1/2 < \infty$.
Therefore, for squared loss we have,

$$V_{N,T}^{\ell_{\text{sq}}} \le \frac{1}{2} \ln(N)$$

3.2 Logarithmic Loss

This type of loss function haven't been discussed before in the lectures. Let us define it first.

$$\ell_{log}(y,\hat{y}) = \begin{cases} -\ln(\hat{y}) & \text{if } y = 1 \\ -\ln(1-\hat{y}) & \text{if } y = 0 \end{cases}$$

For ℓ_{log} loss, $\ell_0(\hat{y}) = -\ln(1-\hat{y})$ and $\ell_1(\hat{y}) = -\ln(1-\hat{y})$. Simple mathematics will lead us to:

$$\ell_0'(\hat{y}) = \frac{1}{1-\hat{y}},\ \ell_1'(\hat{y}) = \frac{-1}{\hat{y}},\ \ell_0''(\hat{y}) = \frac{1}{(1-\hat{y})^2},\ \ell_1''(\hat{y}) = \frac{1}{\hat{y}^2}$$

Thus, we get

$$S_\ell(\hat{y}) = \frac{1}{(1-\hat{y})^2\hat{y}^2} \text{ and subsequently } R_\ell(\hat{y}) = 1, \Rightarrow c_\ell = 1$$

Therefore for the logarithmic loss we get,

$$V_{N,T}^{\ell_{log}} \le \ln(N)$$

3.3 Absolute Loss

For the absolute loss we have $\ell_{\text{abs}}(y, \hat{y}) = |y - \hat{y}|$, here it is easy to see that $\ell_0(\hat{y}) = \hat{y}$ and $\ell_1(\hat{y}) = 1 - \hat{y}$, for $\hat{y} \in (0, 1)$. Also

$$\ell_0'(\hat{y}) = 1,\ \ell_1'(\hat{y}) = -1,\ \ell_0''(\hat{y}) = 0,\ \ell_1''(\hat{y}) = 0$$

And therefore we get $S_\ell(\hat{y}) = 0 \Rightarrow c_\ell = \infty$. Thus, the above theorem doesn't apply in the absolute loss setting.
Thus for the absolute loss we have the result of WMC algorithm from the last lecture which gives an upper bound on the regret. Clearly for the squared and logarithmic loss we saw a much tighter bound; then a natural question arises: can we do better for the absolute loss too? It turns out that we can only improve the constants in the WMC regret bound for the absolute error case.

Theorem 3.2. Let ℓ be s.t. $\ell_0()$ is strictly increasing in (0,1) while $\ell_1()$ is strictly decreasing in (0,1) and both being three times differentiable in (0,1). Then,

- If $S_\ell(\hat{y}) = 0$ for some $\hat{y} \in (0, 1)$, then $V_{N,T} = \Omega\left(T^{1/6}\sqrt{\ln(N)}\right)$
- If $S_\ell(\hat{y}) < 0$ or $\exists\ a < b,$ s.t. $S_\ell(\hat{y}) = 0\ \forall\ \hat{y} \in [a, b],$ then $V_{N,T} = \Omega\left(\sqrt{T\ln(N)}\right)$

Note. Clearly, ℓ_{abs} falls in the second category of the above theorem. Since, we have already seen that $V_{N,T}^{\ell_{\text{abs}}} = O\left(\sqrt{T\ln(N)}\right)$. Thus, $V_{N,T}^{\ell_{\text{abs}}} = \Theta\left(\sqrt{T\ln(N)}\right)$ [1].

4 Vovk's Algorithm, 1990

This algorithm is very similar to the Weighted Majority algorithm. The difference lies in the update, i.e. how do we combine the predictions from the experts to make our own predictions.

Algorithm **Vovk**

Parameters: $c, \eta > 0$

Initial weight vector $\mathbf{w}_i^1 = 1\ \forall\ i \in [N]$
Loss function $\ell : \{0, 1\} \times [0, 1] \to [0, \infty)$
For $t = 1, \ldots, T$:
- Receive expert predictions $\xi_1^t, ..., \xi_N^t \in [0, 1]$
- Compute $\Delta_y = -c \ln\left(\frac{\sum_i \mathbf{w}_i^t e^{-\eta\ell(y,\xi_i^t)}}{\sum_i \mathbf{w}_i^t}\right)$ for $y = 0, 1$.
- Predict any $\hat{y}$ satisfying: $\ell(y, \hat{y}^t) \le \Delta_y \forall y \in \{0, 1\}$. If $\nexists$ such a $\hat{y}$, then the algorithm fails.
- Receive true label $y^t \in \{0, 1\}$
- Incur loss $\ell(y^t, \hat{y})$
- Update $\forall i \in [N]$:
  $\mathbf{w}_i^{t+1} \leftarrow \mathbf{w}_i^t \cdot e^{-\eta\ell(y^t,\xi_i^t)}$

Analysis of Vovk's Algorithm

Suppose that the algorithm doesn't fails. Define $\mathcal{U}^t = -c\ln(W^t)$, where $W^t = \sum_{i=1}^N \mathbf{w}_i^t$.
Then for each trial t,

$$\ell(y^t, \hat{y}^t) \le \mathcal{U}^{t+1} - \mathcal{U}^t \triangleq \Delta_y^t$$

Summing over $t = 1, ..., T$,

$$L_S^\ell[\text{Vovk}(c, \eta)] \le \mathcal{U}^{T+1} - \mathcal{U}^T$$

Now,

$$\mathcal{U}^{T+1} = -c\ln(W^{T+1}) \le -c\ln(\mathbf{w}_i^{T+1})\ \ \forall\ i,\ \ \because\ W^{T+1} \ge w_i^{T+1}$$

Also,

$$\mathcal{U}^1 = -c\ln(N)$$

Thus,

$$L_S^\ell[\text{Vovk}(c,\eta)] \le c\left(\ln(N) + \eta \min_i L_S^\ell[\xi_i]\right)$$

Note. The above is true for all S for which the algorithm with parameters (c,η) doesn't fail. Thus given the loss function of interest, if one can find the parameters c and η such that the algorithm never fails, then we get the above bound for all possible sequence S. The proof of above involves taking $c = c_\ell, \eta = 1/c_\ell$ with ℓ being (c,η)-realizable. Showing in general that any loss function satisfying these conditions is realizable w.r.t. these values involves a significant amount of work.

Definition. Define ℓ to be **$(\mathbf{c},\boldsymbol{\eta})$-realizable** if $\forall\ S,\xi$; Vovk(c,η) doesn't fail.

Now, we will see what conditions are sufficient for the Vovk's algorithm not to fail. We will verify it for some loss functions.

Logarithmic Loss

Run Vovk algorithm with $c = c_\ell$ and $\eta = 1/c_\ell$. For the ℓ_{log} loss we have, $c = 1, \eta = 1$. Define

$$p_i^t = \frac{\mathbf{w}_i^t}{\sum_{j=1}^N \mathbf{w}_j^t}$$

to be the normalized version of the weight vector.

$$\Delta_0 = -c\ln\left(\sum_i p_i^t e^{-\eta\ell_0(\xi_i^t)}\right) = -\ln\left(\sum_i p_i^t e^{\ln(1-\xi_i^t)}\right) = -\ln\left(1 - p^t\cdot\xi^t\right)$$

$$\Delta_1 = -c\ln\left(\sum_i p_i^t e^{-\eta\ell_1(\xi_i^t)}\right) = -\ln\left(\sum_i p_i^t e^{\ln(\xi_i^t)}\right) = -\ln\left(p^t\cdot\xi^t\right)$$

We basically need:

$$\ell_0(\hat{y}^t) \le -\ln\left(1 - p^t\cdot\xi^t\right)$$
$$\ell_1(\hat{y}^t) \le -\ln\left(p^t\cdot\xi^t\right)$$

which is same as requiring:

$$\ell_0^{-1}\left(-\ln\left(1 - p^t\cdot\xi^t\right)\right) \ge \hat{y}^t \iff \hat{y} \le p^t\cdot\xi^t$$
$$\ell_1^{-1}\left(-\ln\left(p^t\cdot\xi^t\right)\right) \ge \hat{y}^t \iff \hat{y} \ge p^t\cdot\xi^t$$

Thus there is only one value $\hat{y}^t = p^t\xi^t$ satisfying it.

Exercise. Work out the conditions for the Vovk's algorithm to be (c,η)-realizable in the squared loss setting.

Absolute Loss

We start with some $\eta > 0$ and $c_{\text{abs}}(\eta) = \frac{1}{2\ln\frac{2}{1+e^{-\eta}}}$.

For the above choice of c, ℓ_{abs} turns out to be realizable for any $\eta > 0$. The bound that we get here is

$$L_S^{\ell_{\text{abs}}}[\text{Vovk}(\eta, c_{\text{abs}}(\eta))] \le \left(\frac{\ln(N) + \eta\min_i L_S^{\ell_{\text{abs}}}[\xi_i]}{2\ln\left(\frac{2}{1+e^{-\eta}}\right)}\right)$$

The above holds for any $\eta > 0$. The bound seems quite similar to the one in Weighted Majority algorithm except for the denominator term. Since, it holds for all η, if we choose η carefully, we can get a bound of the form $L_S^{\ell_{\text{abs}}}[\text{Vovk}(\eta^*, c_{\text{abs}}(\eta^*))] \le \sqrt{T\ln(N)} + \ln(N)/2$. For the absolute loss the best possible bound happens to be $\sqrt{(T/2)\ln(N+1)} + \ln(N+1)/2$, which is obtained when we add an additional arbitrary expert ξ_{N+1}, who always predict the inverse of one of the experts (say ξ_1). Here, at least one of the two experts (ξ_1 and ξ_{N+1}) must have loss at most $T/2$ times.

5 Lower Bound on Minimax Regret for Absolute Loss

From earlier section we know that,

$$V_{N,T}^{\ell_{\text{abs}}} = \min_{\mathcal{A}} \max_{\xi,S} \left(L_S^{\ell_{\text{abs}}}[\mathcal{A}] - \min_i L_S^{\ell_{\text{abs}}}[\xi_i] \right) \equiv \min_{\mathcal{A}} \max_{\xi,S} \mathcal{R}_{\xi,S}^{\ell_{\text{abs}}}[\mathcal{A}]$$

Claim. For any $\mathcal{A}$ and any $\xi_i, ..., \xi_N \in [0,1]^T$ $\exists$ some label sequence S for which the regret follows:

$$\max_{S \in [0,1]^T} \mathcal{R}_{\xi,S}^{\ell_{\text{abs}}}[\mathcal{A}] \geq \frac{T}{2} - \mathbf{E}_{S \in [0,1]^T} \left[\min_i L_S^{\ell_{\text{abs}}}[\xi_i] \right]$$

PROOF. At each trial the algorithm is predicting some value $\hat{y}^t$. The label for the trial is uniformly either 0 or 1, i.e. w.p. $\frac{1}{2}$ the label is 0 and w.p. $\frac{1}{2}$ it is 1. Thus,

$$\mathbf{E}_{S \in [0,1]^T} \left[L_S^{\ell_{\text{abs}}} \right] = \left[\frac{1}{2}(\hat{y}^t - 0) + \frac{1}{2}(1 - \hat{y}^t) \right] \times T = T/2$$

The claim follows by observing that,

$$\max_{S \in [0,1]^T} \mathcal{R}_{\xi,S}^{\ell_{\text{abs}}}[\mathcal{A}] \geq \mathop{\mathbf{E}}_{S \in [0,1]^T} \left[\mathcal{R}_{\xi,S}^{\ell_{\text{abs}}}[\mathcal{A}] \right]$$

□

Now,

$$\begin{aligned}
V_{N,T}^{\ell_{\text{abs}}} &= \min_{\mathcal{A}} \max_{\xi} \max_{S} \left\{ \mathcal{R}_{\xi,S}^{\ell_{\text{abs}}}[\mathcal{A}] \right\} \\
&\geq \min_{\mathcal{A}} \max_{\xi \in (\{0,1\}^T)^N} \max_{S} \left\{ \mathcal{R}_{\xi,S}^{\ell_{\text{abs}}}[\mathcal{A}] \right\} \\
&\geq \min_{\mathcal{A}} \mathbf{E}_{\xi \in (\{0,1\}^T)^N} \left[\frac{T}{2} - \mathbf{E}_{S \in [0,1]^T} \left[\min_i L_S^{\ell_{\text{abs}}}[\xi_i] \right] \right] \\
&\geq \min_{\mathcal{A}} \left[\frac{T}{2} - \mathbf{E}_{\xi \in (\{0,1\}^T)^N, S \in [0,1]^T} \left[\min_i L_S^{\ell_{\text{abs}}}[\xi_i] \right] \right]
\end{aligned}$$

6 Next Lecture

In the next lecture we will study the connections of the online learning with the batch learning and also see some methods to transform the online learning problem into a batch learning problem.

References

[1] Nicolo Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire and Manfred K. Warmuth.
How to use expert advice. *Journal of the ACM*, 44(3):427-485, 1997.

[2] David Haussler, Jyrki Kivinen and Manfred K. Warmuth.
Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906-1925, 1998.