COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Seyed Sobhan Mir Yoosefi

Lecture #15
April 2, 2018

1 Recap

Last time, we introduced an online learning model. Online learning is different from what we have seen before; in these models at each step we see only one example, make a prediction, then we get feedback about our prediction and we can possibly update our prediction rule for the next times. Despite what we did in batch learning we do not assume that examples are coming from a fixed distribution or even that they are random at all. The online learning model that we introduced is called “Learning with Expert Advice” and can be formally written as:

- $N = \#$ experts
- For $t = 1, 2, \ldots, T$
  - every expert i predicts $\xi_i \in \{0, 1\}$
  - learner predicts $\hat{y} \in \{0, 1\}$
  - learner observes true outcome $y \in \{0, 1\}$
  - there is a mistake if $\hat{y} \neq y$

For the case that we have a “perfect” expert that makes no mistake we gave an algorithm called ”Halving Algorithm” and we proved

$$(\#\ \textit{mistakes of Halving Algorithm}) \leq \lg N$$

where $\lg N = \log_2 N$.

2 Connection to PAC Learning

Consider an online version of PAC learning in which examples are coming from the domain $\mathcal{X}$, one at a time. Examples are not necessarily random and can be completely adversarial. The learner tries to predict the label of the example, and after its prediction, the true label will be revealed. The true labels are determined according to some unknown target concept $c \in \mathcal{H}$ where $\mathcal{H}$ is finite and known. It can be formally written as:

- finite hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$
- target concept $c \in \mathcal{H}$
- at each round
  - observe $x \in \mathcal{X}$ (can be adversarial)
  - predict $\hat{y}$
  - observe $y = c(x)$

We want to show that this online version of PAC learning is actually a special case of "Learning with Expert Advice". Consider an expert for each hypothesis h_i in $\mathcal{H}$ with prediction $\xi_i = h_i(x)$. Therefore it can be formally written as:

- finite hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$
- target concept $c \in \mathcal{H}$
- at each round
    - observe $x \in \mathcal{X}$ (can be adversarial)
    - $\forall i \in [n] : \ \xi_i = h_i(x)$
    - predict $\hat{y}$
    - observe $y = c(x)$

Since $c \in \mathcal{H}$, we know that there is a "perfect" expert. Therefore we can apply Halving Algorithm and we can be sure that the number of mistakes is at most $\lg N = \lg |\mathcal{H}|$.

2.1 Mistake Bound

Now it is natural to ask whether Halving Algorithm is the best we can do or not. We are going to partially answer this question. Let's start by defining

$$M_A(\mathcal{H}) = \max_{c,x}(\#\ \textit{mistakes made by } A)$$

$$\text{opt}(\mathcal{H}) = \min_{A} M_A(\mathcal{H})$$

where A is a deterministic algorithm in the definitions above. For a deterministic algorithm A, we define $M_A(\mathcal{H})$ as the maximum number of mistakes that A makes over adversarial choice of examples in $\mathcal{X}$ and target concept in $\mathcal{H}$. Now we can define $\text{opt}(\mathcal{H})$ as the minimum number of mistakes in the worst adversarial setting over the choice of deterministic algorithm A. By the definition we know that $\text{opt}(\mathcal{H}) \leq M_{\text{Halving}}(\mathcal{H})$ and we have previously proved that $M_{\text{Halving}}(\mathcal{H}) \leq \lg |\mathcal{H}|$. Therefore we have

$$\text{opt}(\mathcal{H}) \leq M_{\text{Halving}}(\mathcal{H}) \leq \lg |\mathcal{H}|$$

Now we want to prove that $\text{opt}(\mathcal{H})$ is at least $VCdim(\mathcal{H})$.

Theorem 1. *$VCdim(\mathcal{H}) \leq \text{opt}(\mathcal{H})$*

Proof. Let A be an arbitrary deterministic algorithm and let $d = VCdim(\mathcal{H})$. We know that there are some $x_1, x_2, \ldots, x_d$ that can be shattered by $\mathcal{H}$. Let's choose $c \in \mathcal{H}$ such that

- for $t = 1, \ldots, d$
    - adversary present x_t
    - $\hat{y_t} = A$'s prediction on x_t
    - $y_t = c(x_t) \neq \hat{y_t}$

Since $x_1, \ldots, x_d$ is shattered by $\mathcal{H}$ there exists a $c \in \mathcal{H}$ that for every t we have $y_t = c(x_i) \neq \hat{y_t}$. In addition since A is deterministic, the adversary can simulate A ahead of time to pick the right c. Therefore there is an adversarial setting that can force A to make d mistakes. □

Putting all together we have

$$VCdim(\mathcal{H}) \leq \text{opt}(\mathcal{H}) \leq M_{\text{Halving}}(\mathcal{H}) \leq \lg|\mathcal{H}|$$

We can see that our earlier complexity measures for PAC learning, $VCdim(\mathcal{H})$ and $\lg|\mathcal{H}|$, turn out to also largely control learnability in the online model as well.

3 Weighted Majority Algorithm

Let's go back to "Learning with Expert Advice" that we had. It is not always the case that we have a perfect expert, and in this case, if we apply the Halving Algorithm which eliminates experts as soon as they make a single mistake, we will end up with no expert to listen to. The general idea then is to instead keep the experts who make mistakes, but listen less to their advice. Therefore we keep a weight for each expert, and if some expert makes a mistake, we lower its weight. Then our prediction becomes a weighted vote of experts. This algorithm is called the Weighted Majority Algorithm (WMA), and can be written formally as:

- Parameter $\beta \in [0, 1)$
- weight w_i for each expert i
- initially we have $w_i = 1\ \forall i$
- at round t:
  - get $\xi_i \in \{0, 1\}\ \forall i$
  - $q_0 = \sum_{i:\xi_i=0} w_i$ and $q_1 = \sum_{i:\xi_i=1} w_i$
  - $\hat{y} = \begin{cases} 1 & q_1 > q_0 \\ 0 & \text{otherwise} \end{cases}$
  - learner observes $y \in \{0, 1\}$
  - for each i that $\xi_i \neq y$ update $w_i \leftarrow w_i\beta$

Theorem 2.

$$(\#\textit{mistakes of WMA}) \leq a_\beta(\#\textit{mistakes of best expert}) + c_\beta \lg N$$

where

$$a_\beta = \frac{\lg(\frac{1}{\beta})}{\lg(\frac{2}{1+\beta})} \quad \textit{and} \quad c_\beta = \frac{1}{\lg(\frac{2}{1+\beta})}$$

Let's first look at some values of β.

β	a_β	c_β
$\frac{1}{2}$	≈ 2.4	≈ 2.4
$\to 0$	$+\infty$	1
$\to 1$	2	$+\infty$

As β gets closer to 0, c_β converges to 1, but on the other hand a_β goes to $+\infty$. As β gets closer to 1, a_β converges to 2, but on the other other hand c_β goes to $+\infty$. $\beta = \frac{1}{2}$ seems to be the point in which $a_\beta \approx c_\beta$. Therefore, there is a trade-off here: smaller β results in smaller c_β, but larger a_β; on the other hand, larger β results in smaller a_β, but larger c_β. If we divide both sides by T we get

$$\frac{(\#\textit{mistakes of WMA})}{T} \le a_\beta \frac{(\#\textit{mistakes of best expert})}{T} + c_\beta \frac{\lg N}{T}$$

As $T \to \infty$, we can see that $c_\beta \frac{\lg N}{T} \to 0$. Note that term $\frac{(\#\textit{mistakes of WMA})}{T}$ is the average rate at which the learner makes mistakes. The term $\frac{(\#\textit{mistakes of best expert})}{T}$ can also be interpreted as the average rate in which the best expert makes mistakes. Therefore as T gets large, WMA's mistake rate gets bounded by a constant factor a_β times the mistake rate of the best expert.

Proof. Define $W = \sum_{i=1}^{N} w_i$. Initially we have $W = N$. On some round, let y be the true outcome, w_i's be the current weights, and w_i^{new}'s be the updated weights after this round. Suppose $y = 0$ (the case $y = 1$ is similar). By letting W^{new} be the sum of updated weights we have:

$$\begin{aligned} W^{\text{new}} &= \sum_{i=1}^{N} w_i^{\text{new}} \\ &= \sum_{i:\xi_i=0} w_i + \sum_{i:\xi_i=1} w_i\beta \\ &= \beta q_1 + q_0 \\ &= W - (1-\beta)q_1 \end{aligned}$$

If the learner made a mistake on this round then we know $q_1 \ge q_0$ or equivalently $q_1 \ge \frac{W}{2}$ which results in

$$\begin{aligned} W^{\text{new}} &= W - (1-\beta)q_1 \\ &\le W - (1-\beta)\frac{W}{2} \\ &= \frac{1+\beta}{2} W \end{aligned}$$

Our initial W is N, and after each mistake W decreases at least by a factor of $\frac{1+\beta}{2}$. Note that on all rounds, including rounds where we make no mistake, W remains the same or decreases since $0 \le \beta < 1$. Therefore after m mistakes, we have the following bound:

$$W \le \left(\frac{1+\beta}{2}\right)^m N$$

Let L_i be the number of mistakes that the i-th expert makes. After each mistake, the new weight is the old weight times β; therefore at the end we have

$$w_i = \beta^{L_i}$$

It is clear that $W \ge w_i$ for each i, since weights are non-negative. Thus we have

$$\forall i : \beta^{L_i} \le W \le \left(\frac{1+\beta}{2}\right)^m N$$

Solving for m will give us the bound

$$\forall i : m \leq \frac{\lg(\frac{1}{\beta})L_i + \lg N}{\lg(\frac{2}{1+\beta})}$$

Since it is true for all L_i, it's also true for $\min_i L_i$:

$$m \leq \frac{\lg(\frac{1}{\beta}) \min_i L_i + \lg N}{\lg(\frac{2}{1+\beta})}$$

□

4 Randomized Weighted Majority Algorithm

We can see that a_β in WMA's mistake bound is always greater than or equal to 2; therefore we only proved that our mistake rate is at most twice the rate of the best expert which gives us a weak result if the best expert makes many mistakes (for example if the best expert has error rate of 30%, we will end up with 60%, which is even worse than random guessing). In order to improve the bound, we should introduce randomness. The next algorithm is called Randomized Weighted Majority Algorithm which is very similar to WMA. The only difference is that rather than predicting $\hat{y}$ by weighted majority vote, we pick some expert with probability proportional to its weight and let its prediction be ours for this round. The algorithm is as follows:

- Parameter $\beta \in [0, 1)$
- weight w_i for each expert i
- initially we have $w_i = 1 \ \forall i$
- at round t:
  - get $\xi_i \in \{0, 1\} \ \forall i$
  - $q_0 = \sum_{i:\xi_i=0} w_i$ and $q_1 = \sum_{i:\xi_i=1} w_i$
  - $\hat{y} = \begin{cases} 1 & \text{with probability } \frac{\sum_{i:\xi_i=1} w_i}{\sum_{i=1}^{N} w_i} = \frac{q_1}{W} \\ 0 & \text{with probability } \frac{\sum_{i:\xi_i=0} w_i}{\sum_{i=1}^{N} w_i} = \frac{q_0}{W} \end{cases}$
  - learner observes $y \in \{0, 1\}$
  - for each i that $\xi_i \neq y$ update $w_i \leftarrow w_i\beta$

Or equivalently, let $\hat{y} = \xi_i$ with probability $\frac{w_i}{W}$. Now let's analyze this algorithm

Theorem 3.

$$\mathbb{E}[(\#\textit{mistakes of RWMA})] \leq a_\beta(\#\textit{mistakes of best expert}) + c_\beta \ln N$$

where

$$a_\beta = \frac{\ln(\frac{1}{\beta})}{1-\beta} \quad \textit{and} \quad c_\beta = \frac{1}{1-\beta}$$

In this case we can show that $\alpha_\beta \to 1$ as $\beta \to 1$, which means for large T, the expected mistake rate of RWMA will be close to the mistake rate of the best expert.

Proof. Define ℓ as the probability that RWMA makes a mistake on a round

$$\ell = \Pr[\hat{y} \neq y] = \frac{\sum_{i:\xi_i \neq y} w_i}{W}$$

As before let's see how W changes in a round:

$$W^{\text{new}} = \sum_{i:\xi_i \neq y} w_i \beta + \sum_{i:x_i = y} w_i$$

From the definition of ℓ, we know that $\sum_{i:\xi_i \neq y} w_i = \ell W$.

$$= (\ell W)\beta + (W - \ell W)$$
$$= W(1 - \ell(1-\beta))$$

Therefore our final W can be written as

$$W^{\text{final}} = N \cdot ((1 - \ell_1(1-\beta))) \cdot (1 - \ell_2(1-\beta)) \cdots (1 - \ell_T(1-\beta))$$

where ℓ_t is the probability of mistake on round t. Applying $1 - x \leq e^{-x}$ $\forall x$:

$$\leq N \cdot \prod_{t=1}^{T} \exp(-\ell_t(1-\beta))$$
$$= N \cdot \exp\left(-(1-\beta)\sum_{t=1}^{T} \ell_t\right)$$

Note that $L_A = \sum_{t=1}^{T} \ell_t$ is the expected number of mistakes that RWMA makes. As before, we have $W^{\text{final}} \geq w_i = \beta^{L_i}$, so we can write

$$\forall i: \ \beta^{L_i} \leq W^{\text{final}} \leq Ne^{-(1-\beta)L_A}$$

Solving for L_A will give us the bound

$$L_A \leq \frac{\ln(\frac{1}{\beta})}{1-\beta} \min_i L_i + \frac{1}{1-\beta} \ln N$$

□

5 More Discussion

We want to talk about how to choose β. If we know some value K which is an upper bound on the number of mistakes of the best expert, that is, for which $\min_i L_i \leq K$, then putting $\beta = \frac{1}{1+\sqrt{\frac{2\ln N}{K}}}$ will give us the bound

$$L_A \leq \min_i L_i + \sqrt{2K \ln N} + \ln N$$

A natural question to ask is whether we could do better or not. The answer is yes — we can achieve this by doing something between WMA and RMWA. By looking at Figure 1, we can see three different prediction rules. The y-axis is the probability that the learner predicts

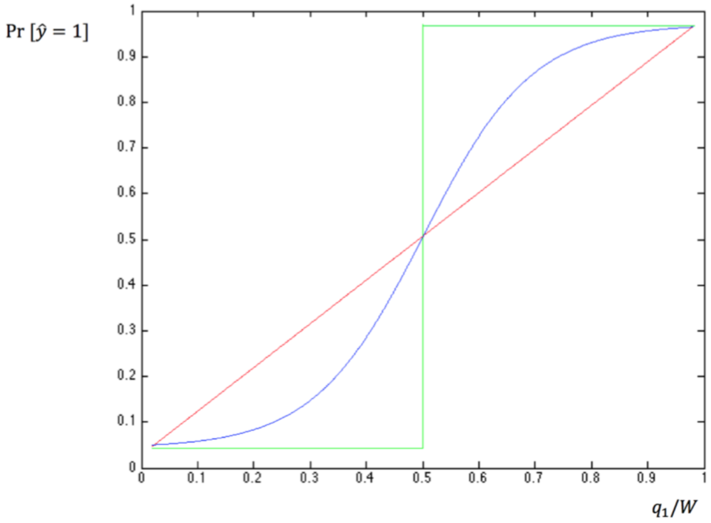

Figure 1: comparison between different prediction rules

1, and the x-axis is the weighted fraction of experts predicting 1. The green curve shows WMA's prediction, and the red curve shows RWMA's. By choosing a different algorithm, something like the blue curve in the figure, we can achieve

$$L_A \leq \frac{\ln(\frac{1}{\beta}) \min_i L_i + \ln N}{2 \ln(\frac{2}{1+\beta})}$$

whose constants are exactly half of the constants we had in the WMA bound. By tuning β for the case that we know $\min_i L_i \leq K$, we can achieve

$$L_A \leq \min_i L_i + \sqrt{K \ln N} + \frac{\lg N}{2}$$

If we have a "perfect" expert, it means we can set $K = 0$ and get the following bound

$$L_A \leq \frac{\lg N}{2}$$

which is half the Halving Algorithm's bound.