COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Miroslav Dudík

Lecture #12
3/13/2003

1 Learning with Expert Advice (cont.)

General framework:

- there are N experts
- at time $t = 1, 2, \ldots, T$:
  1. expert i predicts $\xi_i \in \{0, 1\}$
  2. learner predicts $\hat{y} \in \{0, 1\}$ based on experts' predictions
  3. outcome y is observed; mistake occurs if $\hat{y} \neq y$

Example 1 (from the previous lecture). We considered a special case reminiscent of the PAC model. The crucial difference is that samples are not picked according to a probability distribution $\mathcal{D}$ but they can be picked arbitrarily. We analyze the worst case scenario. The model is as follows:

- sample space $\mathcal{X}$, hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$, $h_i : \mathcal{X} \to \{0, 1\}$; expert i predicts according to h_i
- target concept $c \in \mathcal{H}$ (picked by adversary)
- on each round:
  1. observe $x \in \mathcal{X}$ (picked by adversary)
  2. $\xi_i = h_i(x)$
  3. predict $\hat{y}$
  4. observe $y = c(x)$

We will consider deterministic learning algorithms. For a deterministic algorithm A, let

$$M_A(\mathcal{H}) = \max_{\text{adversary}} (\#\text{mistakes of } A),$$

and define

$$\text{opt}(\mathcal{H}) = \min_A M_A(\mathcal{H}).$$

Theorem 1. $\text{opt}(\mathcal{H}) \leq M_{\text{halving}}(\mathcal{H}) \leq \lg |\mathcal{H}|$ *(proved in the previous lecture).*

Theorem 2. $\text{VCdim}(\mathcal{H}) \leq \text{opt}(\mathcal{H})$.

Proof. Let A^* be an optimal deterministic algorithm, i.e. $M_{A^*}(\mathcal{H}) = \text{opt}(\mathcal{H})$. Assume that $\text{VCdim}(\mathcal{H}) = d$. Let $x_1, \ldots, x_d \in \mathcal{X}$ be shattered by $\mathcal{H}$. The adversary can simulate computation of A^* on samples $x_1, \ldots, x_d$, always producing outcome $y_i \neq \hat{y}_i$. The concept c^* such that $c^*(x_i) = y_i$ for all $i = 1, \ldots, d$ is in $\mathcal{H}$ because $x_1, \ldots x_d$ are shattered. Thus if we choose c^* and samples $x_1, \ldots, x_d$ then the algorithm A^* will make d mistakes. $\square$

The bounds from the previous two theorems are tight. The tight example is $\mathcal{H} = \{h : \{1, \ldots, d\} \to \{0, 1\}\}$.

If we allow randomization then we obtain

$$\frac{\mathrm{VCdim}(\mathcal{H})}{2} \leq \mathrm{opt}^{\mathrm{rand}}(\mathcal{H}) \leq M^{\mathrm{rand}}_{\text{randomized halving}}(\mathcal{H}) \leq \frac{\lg |\mathcal{H}|}{2},$$

where $M^{\mathrm{rand}}_A = \mathbf{E}[\#\text{mistakes } A \text{ makes}]$.

The leftmost inequality can be obtained similarly to the Theorem 2. Each y_i is chosen to be the less likely value of $\hat{y}_i$ conditioned on the previously decided values $y_1, \ldots, y_{i-1}$. Note that we cannot condition on $\hat{y}_1 \neq y_1, \ldots, \hat{y}_{i-1} \neq y_{i-1}$, so we potentially need to consider all possibilities of $\hat{y}_1, \ldots, \hat{y}_{i-1}$ (with appropriate probabilities).

The second inequality in the line is trivial. The last one is somewhat involved and it will not be presented here.

2 Weighted Majority Algorithm

If all experts are allowed to make mistakes then the halving algorithm does not work. In the *weighted majority algorithm* we assign the weight w_i to each expert i, and predict according to the weighted majority of experts. Weights of experts who made a mistake in the given round are reduced by a factor of β, where $\beta \in [0, 1)$ is a parameter of the algorithm.

Weighted Majority Algorithm

initialize $w_i \leftarrow 1$ for $i = 1, \ldots, N$

in each round $t = 1, 2, \ldots, T$ do

let $q_0 = \sum_{i:\xi_i=0} w_i$ and $q_1 = \sum_{i:\xi_i=1} w_i$

$$\hat{y} = \begin{cases} 1 & \text{if } q_1 > q_0 \\ 0 & \text{otherwise} \end{cases}$$

observe y

for all i such that $\xi_i \neq y$ do: $w_i \leftarrow \beta w_i$

Theorem 3. #mistakes of learner $\leq a_\beta \cdot (\#\text{mistakes of best expert}) + c_\beta \lg N$, *where*

$$a_\beta = \frac{\lg(1/\beta)}{\lg\left(\frac{2}{1+\beta}\right)}, \quad c_\beta = \frac{1}{\lg\left(\frac{2}{1+\beta}\right)}.$$

Remark 1. Values of a_β, c_β for $\beta = 0, 1/2, 1$ are given in the following table:

β	a_β	c_β
$1/2$	≈ 2.4	≈ 2.4
$\to 0$	$\to \infty$	$\to 1$
$\to 1$	$\to 2$	$\to \infty$

The value of $\beta = 0$ corresponds to the halving algorithm.

Remark 2. Instead of the number of mistakes, we can consider the rate of mistakes, which is just the number of mistakes divided by the number of rounds. After T rounds we obtain:

$$\text{rate of learner} \leq a_\beta \cdot (\text{rate of best expert}) + c_\beta \cdot \frac{\lg N}{T},$$

hence the rate of learner approaches a_β-multiple of the rate of the best expert as $T \to \infty$.

Proof. Let $W = \sum_{i=1}^{N} w_i$ in each step. Initially $W = N$, during the execution of algorithm the value of W only decreases.

Suppose that the learner makes a mistake in the round t. Let W_{right} be the total weight of experts who provided a correct prediction and W_{wrong} the total weight of experts who made a mistake. Note that $W_{\text{right}} + W_{\text{wrong}} = W$ and $W_{\text{wrong}} \geq W_{\text{right}}$, so $W_{\text{wrong}} \geq W/2$. Therefore,

$$\begin{aligned} W_{\text{new}} = \beta W_{\text{wrong}} + W_{\text{right}} = \beta W_{\text{wrong}} + W - W_{\text{wrong}} = W - (1-\beta) W_{\text{wrong}} \\ \leq W - \frac{1-\beta}{2} \cdot W = \frac{1+\beta}{2} \cdot W \end{aligned}$$

Therefore, if m is the number of mistakes of learner, we obtain

$$W_{\text{final}} \leq N \left(\frac{1+\beta}{2} \right)^m.$$

Let m be the number of mistakes of the learner and m_i the number of mistakes of the expert i. Note that the final weights $w_i = \beta^{m_i}$. Thus for any fixed expert $\hat{\imath}$ we have

$$\beta^{m_{\hat{\imath}}} \leq \sum_{i=1}^{N} \beta^{m_i} = W_{\text{final}}.$$

Combine the two inequalities:

$$\forall 1 \leq \hat{\imath} \leq N: \quad \beta^{m_{\hat{\imath}}} \leq W_{\text{final}} \leq N \left(\frac{1+\beta}{2} \right)^m,$$

which yields

$$m \leq \frac{(\min_{\hat{\imath}} m_{\hat{\imath}}) \lg(1/\beta) + \lg N}{\lg \left(\frac{2}{1+\beta} \right)}.$$

□

3 Randomized Weighted Majority

The values of q_0 and q_1 in the weighted majority algorithm signify the learner's willingness to output 0 or 1, respectively, relative to the weights of experts. Instead of predicting according to the greater value of q_0 or q_1, we will predict 0 with probability q_0/W and 1 with probability q_1/W.

Randomized Weighted Majority Algorithm

initialize $w_i \leftarrow 1$ for $i = 1, \ldots, N$
in each round $t = 1, 2, \ldots, T$ do

let $q_0 = \sum_{i:\xi_i=0} w_i$ and $q_1 = \sum_{i:\xi_i=1} w_i$

$\hat{y} = \begin{cases} 1 & \text{with probability } q_1/W \\ 0 & \text{with probability } q_0/W \end{cases}$

observe y

for all i such that $\xi_i \neq y$ do: $w_i \leftarrow \beta w_i$

Theorem 4. $\mathbf{E}[\#\text{mistakes of learner}] \leq a_\beta \cdot (\#\text{mistakes of best expert}) + c_\beta \ln N$, *where*

$$a_\beta = \frac{\ln(1/\beta)}{1-\beta}, \quad c_\beta = \frac{1}{1-\beta}$$

Proof. Consider the round t. Similarly to the previous proof, let W_{right} be the total weight of experts giving a correct prediction and W_{wrong} the total weight of experts giving an incorrect prediction (i.e. $W_{\text{right}} = q_1, W_{\text{wrong}} = q_0$ if $y = 1$ and $W_{\text{right}} = q_0, W_{\text{wrong}} = q_1$ if $y = 0$). Then

$$W_{\text{new}} = \beta W_{\text{wrong}} + W_{\text{right}} = \beta W_{\text{wrong}} + W - W_{\text{wrong}} = W \cdot (1 - (1-\beta)\frac{W_{\text{wrong}}}{W}). \tag{1}$$

Denote the quantity W_{wrong}/W in round t by ℓ_t. Note that it corresponds to the probability that the learner will make a mistake in round t. Let L denote the number of mistakes of learner and let M_t be a binary random variable equal to 1 when the learner makes a mistake in round t, i.e. $L = \sum_t M_t$. The expected value $\mathbf{E}[M_t] = \ell_t$, so

$$\mathbf{E}[L] = \mathbf{E}[\textstyle\sum_t M_t] = \displaystyle\sum_t \mathbf{E}[M_t] = \sum_t \ell_t.$$

Using (1) we obtain

$$W_{\text{final}} = N \prod_t (1 - \ell_t(1-\beta)) \leq N \exp\left\{-(1-\beta)\sum_t \ell_t\right\} = N \exp\left\{-(1-\beta)\mathbf{E}[L]\right\}.$$

Let L_i be the number of mistakes of the i-th expert. Analogous to the previous proof we obtain

$$\forall 1 \leq \hat{\imath} \leq N : \quad \beta^{L_{\hat{\imath}}} \leq W_{\text{final}},$$

and combining the two inequalities yields

$$\mathbf{E}[L] \leq \frac{(\min_{\hat{\imath}} L_{\hat{\imath}})\ln(1/\beta) + \ln N}{1-\beta}.$$

□

Remark 3. In case that $\min_{\hat{\imath}} L_{\hat{\imath}} \leq K$, we can tune β to be $\beta = (1 + \sqrt{2\ln N/K})^{-1}$, which yields

$$\mathbf{E}[L] \leq \min_{\hat{\imath}} L_{\hat{\imath}} + \sqrt{2K\ln N} + \ln N,$$

and in terms of mistake rates $R = L/T$, $R_i = L_i/T$, $K = rT$,

$$\mathbf{E}[R] \leq \min_{\hat{\imath}} R_{\hat{\imath}} + \sqrt{\frac{2r\ln N}{T}} + \frac{\ln N}{T},$$

which tends to $\min_{\hat{\imath}} R_{\hat{\imath}}$ as $T \to \infty$ (because $K \leq T$).