CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 21**

Online Learning: Halving Algorithm and Exponential Weights

Lecturer: Sasha Rakhlin *Scribe: Ariel Kleiner*

This lecture introduces online learning, in which we largely eschew statistical assumptions and instead consider the behavior of individual sequences of observations and predictions.
See `http://seed.ucsd.edu/~mindreader` for a demonstration.
In general, we will think of an algorithm as a "player" and a source of data as an "adversary."

1 Halving Algorithm

Suppose that we (the player) have access to the predictions of N "experts." Denote these predictions by

$$f_{1,t}, \ldots, f_{N,t} \in \{0, 1\}.$$

At each $t = 1, \ldots, T$, we observe $f_{1,t}, \ldots, f_{N,t}$ and predict $p_t \in \{0, 1\}$. We then observe $y_t \in \{0, 1\}$ and suffer loss $1(p_t \neq y_t)$. Suppose $\exists j$ such that $f_{j,t} = y_t$ for all $t \in [T]$.

Halving Algorithm: predict $p_t = \text{majority}(C_t)$, where $C_1 = [N]$ and $C_t \subseteq [N]$ is defined below for $t > 1$.

Theorem 1.1. If $p_t = \text{majority}(C_t)$ and

$$C_{t+1} = \{i \in C_t : f_{i,t} = y_t\}$$

then we will make at most $\log_2 N$ mistakes.

Proof. For every t at which there is a mistake, at least half of the experts in C_t are wrong and so

$$|C_{t+1}| \leq \frac{|C_t|}{2}.$$

It follows immediately that

$$|C_T| \leq \frac{|C_1|}{2^M}$$

where M is the total number of mistakes. Additionally, because there is a perfect expert, $|C_T| \geq 1$. As a result, recalling that $C_1 = [N]$,

$$1 \leq \frac{N}{2^M}$$

and, rearranging,

$$M \leq \log_2 N.$$

□

2 Exponential Weights or Weighted Majority

We now change our assumptions about the game. For $t = 1, \ldots, T$, the player observes

$$f_{1,t}, \ldots, f_{N,t} \in [0, 1]$$

and predicts $p_t \in [0, 1]$. The outcome $y_t \in [0, 1]$ is then revealed, and the player suffers loss $l(p_t, y_t)$; the experts suffer losses $l(f_{i,t}, y_t), \forall i$. We assume that the loss function $l : [0, 1] \times [0, 1] \to [0, 1]$ is convex in its first argument. Our goal is to achieve low regret R_T, defined as

$$R_T = \underbrace{\sum_{t=1}^{T} l(p_t, y_t)}_{L_T} - \min_{i \in [N]} \underbrace{\sum_{t=1}^{T} l(f_{i,t}, y_t)}_{L_{i,T}}.$$

Exponential Weights (or Weighted Majority) Algorithm: Maintain an (unnormalized) distribution over $[N]$ given by the weights

$$w_{i,t} = e^{-\eta L_{i,t-1}}$$

and predict

$$p_t = \frac{\sum_{i=1}^{N} w_{i,t} f_{i,t}}{\sum_{i=1}^{N} w_{i,t}}.$$

Note that the weights can be defined equivalently by letting $w_{i,1} = 1$ and

$$w_{i,t+1} = w_{i,t} e^{-\eta l(f_{i,t}, y_t)}$$

Theorem 2.1. With an appropriate choice of η,

$$R_T = O(\sqrt{T}).$$

In fact, with $\eta = \sqrt{\frac{8 \ln N}{T}}$,

$$R_T \leq \sqrt{\frac{T}{2} \ln N}.$$

Proof. Define $W_t = \sum_{i=1}^{N} w_{i,t}$. Recall that, by definition, $w_{i,1} = 1, \forall i$ and so $W_1 = N$. Now,

$$\begin{aligned}
\ln \frac{W_{T+1}}{W_1} &= \ln \sum_{i=1}^{N} w_{i,T+1} - \ln N \\
&= \ln \sum_{i=1}^{N} e^{-\eta L_{i,T}} - \ln N \\
&\geq \ln \left(\max_{i=1,\ldots,N} e^{-\eta L_{i,T}} \right) - \ln N \\
&= -\eta \min_{i=1,\ldots,N} L_{i,T} - \ln N. \qquad (1)
\end{aligned}$$

Additionally,

$$\begin{aligned}
\ln \frac{W_{t+1}}{W_t} &= \ln \frac{\sum_{i=1}^N w_{i,t+1}}{\sum_{i=1}^N w_{i,t}} \\
&= \ln \frac{\sum_{i=1}^N e^{-\eta l(f_{i,t},y_t)} w_{i,t}}{\sum_{i=1}^N w_{i,t}} \\
&\leq -\eta \frac{\sum_{i=1}^N l(f_{i,t},y_t) w_{i,t}}{\sum_{i=1}^N w_{i,t}} + \frac{\eta^2}{8} \qquad (2) \\
&\leq -\eta l(p_t, y_t) + \frac{\eta^2}{8}. \qquad (3)
\end{aligned}$$

Inequality (2) holds because of Hoeffding's inequality:

$$\ln \mathbb{E} e^{sX} \leq s\mathbb{E}X + \frac{s^2(a-b)^2}{8}$$

for any random variable $X \in [a,b]$ and any $s \in \mathbb{R}$. The role of X in (2) above is played by $l(f_{i,t}, y_t)$, and the role of s is played by $-\eta$. Inequality (3) follows from Jensen's inequality because l is convex in its first argument.

Using (3), we find that

$$\begin{aligned}
\ln \frac{W_{T+1}}{W_1} &= \ln \frac{W_{T+1}}{W_T} + \ln \frac{W_T}{W_{T-1}} + \cdots + \ln \frac{W_2}{W_1} \\
&\leq -\eta \sum_{t=1}^T l(p_t, y_t) + T\frac{\eta^2}{8}.
\end{aligned}$$

Therefore, combining this inequality with the lower bound (1) obtained above, we have

$$-\eta \min_{i=1,\ldots,N} L_{i,T} - \ln N \leq -\eta L_T + T\frac{\eta^2}{8}$$

and so, rearranging,

$$L_T \leq \min_{i=1,\ldots,N} L_{i,T} + \frac{\ln N}{\eta} + T\frac{\eta}{8}.$$

Finally, optimizing over η (i.e., minimizing the last two terms with respect to η), we obtain the desired result. □