Online Learning

Once again, we follow Mohri, Rostamizadeh and Talwalkar (2012).

Online (sequential) prediction is amazing because it is completely assumption free. The basic setup is as follows:

1. For $t = 1, \ldots, T$:
   (a) Observe x_t.
   (b) Predict $\widehat{y}_t$.
   (c) Observe y_t.
   (d) Incur loss $L(\widehat{y}_t, y_t)$.
2. The cumulative loss is $\sum_t L(\widehat{y}_t, y_t)$.

Usually we will assume that $y_t \in \{0, 1\}$ and that $L(\widehat{y}_t, y_t) = I(\widehat{y}_t \neq y_t)$.

In the *expert advice* setting we have N algorithms (experts). The prediction from algorithm i is $y_{t,i}$. The goal in this case is to minimize the *regret*

$$R = \sum_t L(\widehat{y}_t, y_t) - \min_i L(y_{t,i}, y_t).$$

Halving Algorithm. This is the simplest case. We have a finite set of predictors $\mathcal{H}$. We assume there is one $h \in \mathcal{H}$ that makes perfect predictions. Let $M(h)$ be the maximum number of mistakes that our algorithm makes (over all $x_1, \ldots, x_T$). Let $M(\mathcal{H}) = \max_h M(h)$. The algorithm is as follows:

1. Set $\mathcal{H}_1 = \mathcal{H}$.
   (a) Observe x_t. Let $\widehat{y}_t$ be the majority vote of $\mathcal{H}_t$.
   (b) Observe y_t.
   (c) If $\widehat{y}_t \neq y_t$ set $\mathcal{H}_t = \{h : \ h(x_t) = y_t\}$.

Theorem 1 $M(\mathcal{H}) \leq \log_2 |\mathcal{H}|$.

Proof. If $\widehat{y}_t \neq y_t$ then we reduce $\mathcal{H}_t$ by at least half so that $|\mathcal{H}_{t+1}| \leq (1/2)|\mathcal{H}_t|$. So after $\log_2 |\mathcal{H}|$ mistakes there is only one expert left which must be the perfect expert and hence there will be no more mistakes. □

The assumption of a perfect predictor is unrealistic so let's move on to a more realistic setting.

Weighted Majority. The algorithm is:

1. Set $\beta \in [0, 1)$.
2. Set $w_{1,i} = 1$ for $i = 1, \ldots, N$.
3. For $t = 1, \ldots, T$:
   (a) Observe x_t.
   (b) If
   $$\sum_{y_{t,i}=1} w_{t,i} \geq \sum_{y_{t,i}=0} w_{t,i}$$
   then $\widehat{y}_t = 1$ else $\widehat{y}_t = 0$.
   (c) Observe y_t.
   (d) If $\widehat{y}_t \neq y_t$:
   $$\begin{array}{ll} \text{If } y_{t,i} \neq y_t & \text{set } w_{t+1,i} = \beta w_{t,i} \\ \text{If } y_{t,i} = y_t & \text{set } w_{t+1,i} = w_{t,i}. \end{array}$$

Let $m^* = \min_i \sum_t I(y_{t,i} \neq y_t)$ be the loss of the best expert. Let m be the loss of the algorithm.

Theorem 2 *We have that*
$$m \leq \frac{\log N + m^* \log(1/\beta)}{\log(2/(1+\beta))}.$$

Proof. Let $W_t = \sum_i w_{t,i}$. Note that $W_1 = N$. Because of the weighted majority rule, we have that if there is an error,
$$W_{t+1} \leq \left(\frac{1}{2} + \frac{\beta}{2}\right) W_t = \left(\frac{1+\beta}{2}\right) W_t.$$

Hence,
$$W_T \leq \left(\frac{1+\beta}{2}\right)^m N.$$

On the other hand, for each i,
$$W_T \geq w_{T,i} = \beta^{m(T,i)}$$

where $m(T,i)$ is the number of mistakes from expert i. In particular, this holds for the best expert so that

$$W_T \geq \beta^{m^*}.$$

Combining these two bounds,

$$\beta^{m^*} \leq W_T \leq \left(\frac{1+\beta}{2}\right)^m N.$$

Taking the log and re-arranging terms gives the result. □

This is a nice result but it does not guarantee that the loss is small. To see this, supposee there are two experts. The first outputs $0, 0, \ldots, 0$ and the second outputs, $1, 1, \ldots, 1$. Note that $m^* \leq T/2$. Now suppose that nature is evil and sets $y_t = 0$ when $\widehat{y}_t = 1$ and sets $y_t = 1$ when $\widehat{y}_t = 0$. Then $m = T$. So the regret is $R = m - m^* \geq T/2$. Can we make the regret smaller. Yes, as we now show.

Randomized Weighted Majority. For this algorithm we choose expert i with some probability $p_{t,i}$. We receive a vector of losses $\ell_t = (\ell_{t,1}, \ldots, \ell_{t,N})$. The expected loss is $L_t = \sum_i p_{t,i}\ell_{t,i}$ and the cumulative expected loss is $\mathcal{L}_T = \sum_{t=1}^T L_t$. We also define $\mathcal{L}_{T,i} = \sum_t \ell_{t,i}$ and the minimum loss $\mathcal{L}^* = \min_i \mathcal{L}_{T,i}$. Here is the algorithm:

1. Set $w_{i,1} = 1$ for $i = 1, \ldots, N$.
2. Set $p_{1,i} = 1/N$ for $i = 1, \ldots, N$.
3. For $t = 1, \ldots, T$:
   (a) If $\ell_{t,i} = 1$ set $w_{t+1,i} = \beta w_{t,i}$. If $\ell_{t,i} = 0$ set $w_{t+1,i} = w_{t,i}$.
   (b) Let $W_{t+1} \sum_i w_{t+1,i}$.
   (c) Set $p_{t+1,i} = w_{t+1,i}/W_{t+1}$.

Theorem 3 *We have*

$$\mathcal{L}_T \leq \mathcal{L}_* + 2\sqrt{T \log N}.$$

The remarkable thing about this result is that the regret only grows at rate $\sqrt{T}$. In other words, the average regrest is $\sqrt{\log N/T}$.

Proof. Set $p_{t,i} = w_{t,i}/W_t$ we have that $w_{t,i} = W_t p_{t,i}$. Hence,

$$\begin{aligned} W_{t+1} &= \sum_{i:\ \ell_{t,i}=0} w_{t,i} + \beta \sum_{i:\ \ell_{t,i}=1} w_{t,i} = W_t + (\beta - 1) \sum_{i:\ \ell_{t,i}=1} w_{t,i} \\ &= W_t + (\beta-1)W_t \sum_{i:\ \ell_{t,i}=1} p_{t,i} = W_t + (\beta-1)W_t L_t = W_t[1-(1-\beta)L_t]. \end{aligned}$$

Recalling that $W_1 = N$ we see that

$$W_{T+1} = N \prod_t [1 - (1-\beta) L_t].$$

On the other hand,

$$W_{T+1} \geq \max_i w_{T+1,i} = \beta^{\mathcal{L}_*}.$$

Combining these inequalities we get

$$\beta^{\mathcal{L}_*} \leq W_{T+1} \leq N \prod_t [1 - (1-\beta) L_t].$$

Hence,

$$\begin{aligned}
\mathcal{L}_* \log \beta &\leq \log N + \sum_t [1 - (1-\beta) L_t] \\
&\leq \log N - (1-\beta) \sum_t L_t \quad \text{since } \log(1-x) \leq -x \\
&= \log N - (1-\beta) \mathcal{L}_T.
\end{aligned}$$

Re-arranging terms we get

$$\mathcal{L}_T \leq \frac{\log N}{1-\beta} + (1-\beta) T + \mathcal{L}_*.$$

Now we set $\beta = 1 - \sqrt{\log N / T}$ and we have

$$\mathcal{L}_T \leq \mathcal{L}_* + 2\sqrt{T \log N}.$$

□

Exponential Weights. Now we allow the loss to take values in $[0,1]$. We handle this case by modifying the weights. We assume that the loss function L is convex in its first argument. In what follows, $L_{t,i}$ is the total loss of expert i after t steps. Here is the algorithm:

1. Set $w_{1,i} = 1$ for $i = 1, \ldots, N$.
2. For $t = 1, \ldots, T$:
   (a) Observe x_t.
   (b) Let
   $$\widehat{y}_t = \frac{\sum_i w_{t,i} y_{t,i}}{\sum_i w_{t,i}}.$$
   (c) Observe y_t. Set
   $$w_{t+1,i} = w_{t,i} e^{-\theta L(y_{t,i}, y_t)}.$$

Theorem 4 *If* $\theta = \sqrt{8 \log N / T}$ *then the regret satisfies*

$$R_T \le \sqrt{T \log N / 2}.$$

Remark: The interesting thing about the proof below is that it uses probabilistic ideas even though there is no probability distribution in the setup of the problem.

Proof. Let us begin by recalling the following fact: suppose that $a \le X \le b$ and $\mathbb{E}[X] = 0$. Then

$$\mathbb{E}[e^{tX}] \le e^{t^2 (b-a)^2/8}. \tag{1}$$

Define

$$\Phi_t = \log \sum_i w_{t,i}.$$

Then

$$\begin{aligned}\Phi_{t+1} - \Phi_t &= \log \frac{\sum_i w_{t+1,i}}{\sum_i w_{t,i}} = \log \frac{w_{t,i} e^{-\theta L(y_{t,i}, y_t)}}{\sum_i w_{t,i}} \\ &= \log \mathbb{E}_t e^{\theta X}\end{aligned}$$

where

$$X = -L(y_{t,i}, y_t) \in [-1, 0]$$

and $\mathbb{E}_t$ refers to expection with respect to the distribution with probability function $p_{t,i} = \frac{w_{t,i}}{\sum_i w_{t,i}}$. So

$$\begin{aligned}\Phi_{t+1} - \Phi_t &= \log \mathbb{E}_t e^{\theta(X - \mathbb{E}_t[X]) + \theta \mathbb{E}_t[X]} \\ &= \theta \mathbb{E}_t[X] + \log \mathbb{E}_t e^{\theta(X - \mathbb{E}_t[X])} \\ &\le \theta \mathbb{E}_t[X] + \frac{\theta^2}{8} \quad \text{using (1)} \\ &= -\theta \mathbb{E}_t[L(y_{t,i}, y_t)] + \frac{\theta^2}{8} \\ &\le -\theta L(\mathbb{E}_t[y_{t,i}], y_t) + \frac{\theta^2}{8} \quad \text{using convexity} \\ &= -\theta L(\widehat{y}_t, y_t) + \frac{\theta^2}{8} \quad \text{definition of } \widehat{y}_t.\end{aligned}$$

Now we sum over t to get

$$\Phi_{T+1} - \Phi_1 \le -\theta \sum_t L(\widehat{y}_t, y_t) + \frac{\theta^2 T}{8}.$$

Next we have the lower bound

$$\begin{aligned}\Phi_{T+1} - \Phi_1 &= \log \sum_i w_{T+1,i} - \log N = \log \sum_i e^{-\theta L_{T,i}} - \log N \\ &\ge \log \max_i e^{-\theta L_{T,i}} - \log N = -\theta \min_i L_{T,i} - \log N.\end{aligned}$$

Combining the lower and upper bound we have

$$-\theta \min_i L_{T,i} - \log N \leq \frac{\theta^2 T}{8} - \theta \sum_t L(\widehat{y}_t, y_t)$$

which implies that

$$\sum_t L(\widehat{y}_t, y_t) - \min_i L_{T,i} \leq \frac{\log N}{4} + \frac{\theta T}{8}.$$

The result follows by setting $\theta = \sqrt{8 \log N / T}$. □

Online to Batch. The setting we have focused on in class is the batch setting where we observe random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ from some distribution. It turns out that we can apply online algorithms to the batch setting. We again assume that the loss L is convex in its first argument.

Let $\mathcal{H}$ be a set of classifiers and assume that the loss function L is bounded by M. Suppose we have an online algorithm. Let h_i denote the classifier returned by the algorithm after observing (X_i, Y_i). As before, the regret is defined as

$$R_T \sum_i L(h_i(X_i), Y_i) - \min_{h \in \mathcal{H}} \sum_{i=1}^{T} L(h(X_i), Y_i).$$

Let $R(h) = \mathbb{E}[L(h(X), Y)]$. First we bound the average risk.

Theorem 5 *With probability at least* $1 - \delta$,

$$\frac{1}{T} \sum_i R(h_i) \leq \frac{1}{T} \sum_i L(h_i(X_i), Y_i) + M \sqrt{\frac{2 \log(1/\delta)}{T}}.$$

Before proceeding let us recall Azuma's inequality. If V_i is a sequence of random variables that satisfy

$$\mathbb{E}[V_{i+1} | X_1, \ldots, X_i] = 0$$

and $|V_i| \leq M$ then

$$P\left(\frac{1}{T} \sum_i X_i > \epsilon\right) \leq e^{-T\epsilon^2/(2M^2)}. \tag{2}$$

Proof. Let $V_i = R(h_i) - L(h_i(X_i), Y_i)$. Then

$$\mathbb{E}[V_i | X_1, \ldots, X_{i-1}] = R(h_i) - \mathbb{E}[L(h_i(X_i), Y_i) | h_i] = R(h_i) - R(h_i) = 0.$$

Also $|V_i| \leq M$. Let

$$\epsilon = M \sqrt{(2/T) \log(1/\delta)}.$$

By (2),

$$P\left(\frac{1}{T}\sum_i X_i > \epsilon\right) \le e^{-T\epsilon^2/(2M^2)} = \delta$$

and result follows from the definition of V_i. □

We define our batch classifier as

$$h = \frac{1}{T}\sum_{i=1}^{T} h_i.$$

Theorem 6 *We have, with probability at least* $1-\delta$ *that*

$$R(h) \le \inf_{h\in\mathcal{H}} + \frac{R_T}{T} + 2M\sqrt{\frac{2\log(1/\delta)}{T}}. \tag{3}$$

If we use the exponentially weigthed algorithm then $R_T \le \sqrt{T\log N/2}$. Plugging this into (3) we have

$$R(h) \le \inf_{h\in\mathcal{H}} + \sqrt{\frac{\log N}{2T}} + M\sqrt{\frac{2\log(1/\delta)}{T}}.$$

Proof. By convexity,

$$L\left(\frac{1}{T}\sum_i h(X_i), Y_i\right) \le \frac{1}{T}\sum_i L(h_i(X_i), Y_i).$$

By taking the expected value and using the fact that $h = \frac{1}{T}\sum_{i=1}^{T} h_i$,

$$R(h) \le \frac{1}{T}\sum_i R(h_i).$$

From the previous theorem, with probability at least $1-\delta/2$,

$$R(h) \le \frac{1}{T}\sum_i L(h_i(X_i), Y_i) + M\sqrt{\frac{2\log(2/\delta)}{T}}. \tag{4}$$

Since

$$R_T = \sum_i L(h(X_i), Y_i) - \min h \in \mathcal{H}\sum_i L(h(X_i), Y_i)$$

(4) implies that

$$\begin{aligned} R(h) &\le \frac{1}{T}\min h\in\mathcal{H}\sum_i L(h(X_i), Y_i) + \frac{R_T}{T} + M\sqrt{\frac{2\log(2/\delta)}{T}} \\ &= \frac{1}{T}\sum_i L(h_*(X_i), Y_i) + \frac{R_T}{T} + M\sqrt{\frac{2\log(2/\delta)}{T}}. \end{aligned}$$

By Hoeffding's inequality, with probability at least $1-\delta/2$,

$$\frac{1}{T}\sum_i L(h_*(X_i), Y_i) \leq R(h_*) + M\sqrt{\frac{2\log(2/\delta)}{T}}.$$

Hence,

$$R(h) \leq R(h_*) + \frac{R_T}{T} + 2M\sqrt{\frac{2\log(2/\delta)}{T}}.$$

□