Machine Learning Theory (CS 6783)

Lecture 12: Online Learning, Bit Prediction and more

1 Bit Prediction

Say we have a sequence of coin flips $y_1 = +1, y_2 = -1, y_3 = +1, y_4 = +1, \ldots$. How do we predict the $t+1$'th outcome given the past t outcomes?

If the bits are produces iid by coin flips, then picking majority amongst outcomes so far, or estimating probability of $+1$ by its empirical frequency and using a randomized predictor q_t such that $\mathbb{E}_{\hat{y}_t \sim q_t}[\hat{y}_t] = \frac{1}{t-1}\sum_{j=1}^{t-1} y_j$ both work equally well. Specifically, for randomized predictor $q_t = \frac{1}{2}\frac{1}{t-1}\sum_{j=1}^{t-1} y_j + \frac{1}{2}$ we can show using Chernoff-type bounds that if y_t's were drawn iid from some fixed Bernoulli distribution with parameter p (unknown to learner), then:

$$\frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\{\hat{y}_t \neq y_t\}\right] \leq \min_{b\in\{\pm 1\}} \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{y_t \neq b\} + O\left(\frac{1}{\sqrt{n}}\right)$$

and indeed $\min_{b\in\{\pm 1\}} \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{y_t \neq b\}$ is close to the Bayes error and we can't really do better than this.

But here we made the crucial assumption that bits were drawn iid from a Bernoulli distribution. What if this were not true? In this case how do we make predictions? In this case, can we bound the below quantity referred to as regret? Maybe even by $1/\sqrt{n}$ like the iid case (as long as we are wishing)?

$$\frac{1}{n}\mathrm{Reg}_n = \frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \min_{b\in\{\pm 1\}} \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{y_t \neq b\}$$

When the bits are not drawn iid, this problem is far more complicated and interesting. First off, any deterministic algorithm can be made to incur maximal regret. Specifically, think of the process where learner deterministically on a round t predicts $\hat{y}_t \in \{\pm 1\}$, then setting $y_t = -\hat{y}_t$, we guarantee that our average loss is 1 while in hindsight, $\min_{b\in\{\pm 1\}} \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{y_t \neq b\}$ is at worst $1/2$. Hence deterministic algorithms like majority so far have to fail.

In fact, even the randomized algorithm that predicts based on estimated frequency so far $q_t = \frac{1}{2}\frac{1}{t-1}\sum_{j=1}^{t-1} y_j + \frac{1}{2}$ fails. To see this, say we flip coins and with probability $2/3$ we pick $+1$ and with probability $1/3$ its -1. But now say we sort these bits and present the $n/3$, bits of -1 first then the $2n/3$ bits of $+1$ next. In this case, note that the strategy $q_t = \frac{1}{2}\frac{1}{t-1}\sum_{j=1}^{t-1} y_j + \frac{1}{2}$ (after the very first round which we can ignore), makes 0 mistakes for the first $n/3$ rounds when -1 labels are

presented. But from then on, we have a larger expected error on every round. Specifically, we get,

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\mathbf{1}\{y_t\neq\hat{y}_t\} &\geq \frac{1}{n}\sum_{t=n/3+1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\mathbf{1}\{+1\neq\hat{y}_t\} = \frac{1}{n}\sum_{t=n/3+1}^{n}(1-q_t)\\ &= \frac{1}{n}\sum_{t=n/3+1}^{n}\left(1-\frac{1}{2}\frac{1}{t-1}\sum_{j=1}^{t-1}y_j-\frac{1}{2}\right)\\ &= \frac{1}{2n}\sum_{t=n/3+1}^{n}\left(1-\frac{1}{t-1}\left(t-1-\frac{2n}{3}\right)\right) = \frac{1}{3}\sum_{t=n/3+1}^{n}\left(\frac{1}{t-1}\right)\end{aligned}$$

Note that in the above, $\sum_{t=n/3+1}^{n}\left(\frac{1}{t-1}\right)$ is approximately $\log(3) > 1$ or at least is a fixed constant greater than 1 while $\min_{b\in\{\pm1\}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}\{y_t\neq b\} = 1/3$. Thus we see that for this algorithm, we can never hope to get regret that diminishes to 0.

So is it at all possible to get average regret to diminish to 0 with n?

2 Cover Result and Algorithm

Lets get back to the problem of predicting arbitrary (possibly adversarially chosen) sequence of bits as well as majority. In fact, think of the online learning problem where on each round t we we predict the next bit $y_t \in \{\pm1\}$. Also say $\mathcal{F} \subset \{\pm1\}^n$ and we want to minimize regret (in expectation) :

$$\mathrm{Reg}_n = \frac{1}{n}\sum_{t=1}^{n}\mathbb{1}_{\{\hat{y}_t\neq y_t\}} - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}_{\{f_t\neq y_t\}}$$

Note that for the majority case, $\mathcal{F} = \{(+1,+1,\ldots,+1),(-1,-1,\ldots,-1)\}$ are the two predictors. When can we ensure $\mathbb{E}[\mathrm{Reg}_n]\to 0$?

Lemma 1 (T. Cover'65)**.** *Let $\phi : \{\pm1\}^n \mapsto \mathbb{R}$ be a function such that, for any i, and any $y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n$,*

$$|\phi(y_1,\ldots,y_{i-1},+1,y_{i+1},\ldots,y_n) - \phi(y_1,\ldots,y_{i-1},-1,y_{i+1},\ldots,y_n)| \leq \frac{1}{n}\ ,\ \textit{(stability condition)}$$

then, there exists a randomized strategy such that for any sequence of bits,

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\left[\mathbf{1}\{\hat{y}_t\neq y_t\}\right] \leq \phi(y_1,\ldots,y_n)$$

if and only if,

$$\mathbb{E}_\epsilon\phi(\epsilon_1,\ldots,\epsilon_n) \geq \frac{1}{2}$$

and further, the strategy achieving this bound on expected error is given by:

$$q_t = \frac{1}{2} + \frac{n}{2}\,\mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n}\left[\phi(y_1,\ldots,y_{t-1},-1,\epsilon_{t+1},\ldots,\epsilon_n) - \phi(y_1,\ldots,y_{t-1},+1,\epsilon_{t+1},\ldots,\epsilon_n)\right]$$