Machine Learning Theory (CS 6783)

Lecture 13 : Bit Prediction and Multiclass Prediction

1 Bit Prediction

Claim 1. *There exists a randomized prediction strategy that ensures that*

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \leq \frac{1}{2n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n} f_t\epsilon_t\right]$$

To prove the above claim we first prove this following lemma, a result by Thomas Cover.

Lemma 2 (T. Cover'65)**.** *Let $\phi : \{\pm 1\}^n \mapsto \mathbb{R}$ be a function such that, for any i, and any $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$,*

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

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:

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

Proof of Lemma.
We start by proving that if there exists an algorithm that guarantees that

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

then, $\mathbb{E}_\epsilon\left[\phi(\epsilon_1, \ldots, \epsilon_n)\right] \geq 1/2$**.**

To see this, note that the regret bound implies that

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\left[\mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \phi(y_1, \ldots, y_n) \leq 0$$

for any $y_1, \ldots, y_n$. Now simply let the adversary pick $y_t = \epsilon_t$ as a Rademacher random variable. Thus, taking expectation, this implies that,

$$0 \geq \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} \left[\mathbb{E}_{\epsilon_t} \mathbf{1}\{\hat{y}_t \neq \epsilon_t\} \right] - \mathbb{E}_{\epsilon} \phi(\epsilon_1, \ldots, \epsilon_n) = \frac{1}{2} - \mathbb{E}_{\epsilon} \phi(\epsilon_1, \ldots, \epsilon_n)$$

Next we prove that if $\mathbb{E}_{\epsilon} \phi(\epsilon_1, \ldots, \epsilon_n) \geq \frac{1}{2}$, then $\exists$ strategy s.t. $\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} \left[\mathbf{1}\{\hat{y}_t \neq y_t\} \right] \leq \phi(y_1, \ldots, y_n)$.

The basic idea is to prove this statement starting from n and moving backwards. Say we have already played rounds up until round $n-1$ and have observed $y_1, \ldots, y_{n-1}$. Now let us consider the last round. On the last round we use,

$$q_n = \frac{1}{2} + \frac{n}{2}\ \phi(y_1, \ldots, y_{n-1}, -1) - \phi(y_1, \ldots, y_{n-1}, +1)$$

Now note that if $y_n = +1$ then $\mathbb{E}_{\hat{y}_n \sim q_n} \left[\mathbb{1}_{\{\hat{y}_n \neq y_n\}} \right] = \mathbb{E}_{\hat{y}_n \sim q_n} \left[\mathbb{1}_{\{\hat{y}_n = -1\}} \right] = 1 - q_n$ and if $y_n = -1$ then $\mathbb{E}_{\hat{y}_n \sim q_n} \left[\mathbb{1}_{\{\hat{y}_n \neq y_n\}} \right] = q_n$ and hence for the choice of q_n above, we can write

$$\mathbb{E}_{\hat{y}_n \sim q_n} \left[\mathbb{1}_{\{\hat{y}_n \neq y_n\}} \right] = \tfrac{1}{2n} - \tfrac{y_n}{2} \left(\phi(y_1, \ldots, y_{n-1}, -1) - \phi(y_1, \ldots, y_{n-1}, +1) \right)$$

Plugging in the above, note that for any y_n (possibly chosen adversarially looking at q_n), we have,

$$\begin{aligned}
&\tfrac{1}{n}\mathbb{E}_{\hat{y}_n \sim q_n} \left[\mathbb{1}_{\{\hat{y}_n \neq y_n\}} \right] - \phi(y_1, \ldots, y_n) && (1)\\
&\qquad = \frac{1}{2n} - \frac{y_n}{2} \left(\phi(y_1, \ldots, y_{n-1}, -1) - \phi(y_1, \ldots, y_{n-1}, +1) \right) - \phi(y_1, \ldots, y_n) \\
&\qquad = \frac{1}{2n} - \frac{1}{2} \left(\phi(y_1, \ldots, y_{n-1}, -1) + \phi(y_1, \ldots, y_{n-1}, +1) \right) \\
&\qquad = \frac{1}{2n} - \mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-1}, \epsilon_n) && (2)
\end{aligned}$$

Now recursively we continue just as above for $n-1$ to 0. Let us do the $n-1$th step and the rest follows. To this end, note that just as earlier, if $y_{n-1} = +1$ then $\mathbb{E}_{\hat{y}_{n-1} \sim q_{n-1}} \left[\mathbb{1}_{\{\hat{y}_{n-1} \neq y_{n-1}\}} \right] = \mathbb{E}_{\hat{y}_{n-1} \sim q_{n-1}} \left[\mathbb{1}_{\{\hat{y}_{n-1} = -1\}} \right] = 1 - q_{n-1}$ and if $y_{n-1} = -1$ then $\mathbb{E}_{\hat{y}_{n-1} \sim q_{n-1}} \left[\mathbb{1}_{\{\hat{y}_{n-1} \neq y_{n-1}\}} \right] = q_{n-1}$ and hence for the choice of $q_{n-1} = \frac{1}{2n} + \frac{n}{2}\ \mathbb{E}_{\epsilon_n} \left[\phi(y_1, \ldots, y_{n-2}, -1, \epsilon_n) - \phi(y_1, \ldots, y_{n-2}, +1, \epsilon_n) \right]$, we have

$$\tfrac{1}{n}\mathbb{E}_{\hat{y}_{n-1} \sim q_{n-1}} \left[\mathbb{1}_{\{\hat{y}_{n-1} \neq y_{n-1}\}} \right] = \tfrac{1}{2n} - \tfrac{y_{n-1}}{2} \left(\mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-2}, -1, \epsilon_n) - \mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-2}, +1, \epsilon_n) \right)$$

Thus we can conclude that,

$$\begin{aligned}
&\tfrac{1}{n}\mathbb{E}_{\hat{y}_{n-1} \sim q_{n-1}} \left[\mathbb{1}_{\{\hat{y}_{n-1} \neq y_{n-1}\}} \right] + \tfrac{1}{n}\mathbb{E}_{\hat{y}_n \sim q_n} \left[\mathbb{1}_{\{\hat{y}_n \neq y_n\}} \right] - \phi(y_1, \ldots, y_n) \\
&\quad = \frac{1}{2n} + \tfrac{1}{n}\mathbb{E}_{\hat{y}_{n-1} \sim q_{n-1}} \left[\mathbb{1}_{\{\hat{y}_{n-1} \neq y_{n-1}\}} \right] - \mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-1}, \epsilon_n) \qquad \text{(From Eq.2)} \\
&\quad = \frac{2}{2n} - \tfrac{y_{n-1}}{2} \left(\mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-2}, -1, \epsilon_n) - \mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-2}, +1, \epsilon_n) \right) - \mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-1}, \epsilon_n) \\
&\quad = \frac{2}{2n} - \frac{1}{2} \left(\mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-2}, +1, \epsilon_n) + \mathbb{E}_{\epsilon_n} \phi(y_1, \ldots, y_{n-2}, -1, \epsilon_n) \right) \\
&\quad = \frac{2}{2n} - \mathbb{E}_{\epsilon_{n-1}, \epsilon_n} \phi(y_1, \ldots, y_{n-2}, \epsilon_{n-1}, \epsilon_n)
\end{aligned}$$

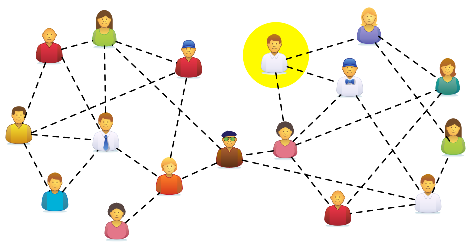
Proceeding in similar way we conclude that,

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\left[\mathbb{1}_{\{\hat{y}_t\neq y_t\}}\right]-\phi(y_1,\ldots,y_n)\leq\frac{n}{2n}-\mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\phi(\epsilon_1,\ldots,\epsilon_n)=\frac{1}{2}-\mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\phi(\epsilon_1,\ldots,\epsilon_n)$$

Hence, if $\mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\phi(\epsilon_1,\ldots,\epsilon_n)\geq 1/2$ then we can conclude that, $\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\left[\mathbb{1}_{\{\hat{y}_t\neq y_t\}}\right]\leq\phi(y_1,\ldots,y_n)$ as desired.

Hence we conclude the proof of this lemma. □

2 Application: Binary Node Classification



Let $G=(V,E)$ be a known undirected graph representing a social network. At each time step t, a user in the network opens her Facebook page, and the system needs to decide whether to classify the user as type "-1" or "$+1$", say, in order to decide on an advertisement to display. We assume here that the feedback on the "correct" type is revealed to the system after the prediction is made. Suppose we have a hunch that the type of the user ($+1$ or -1) is correlated with the community to which she belongs. For simplicity, suppose there are two communities, more densely connected within than across. To capture the idea of correlating communities and labels, we set ϕ to be small on labelings that assign homogenous values within each community. We make the following simplifying assumptions: (i) $|V|=n$, (ii) we only predict the label of each node once, and (iii) the order in which the nodes are presented is fixed (this assumption is easily removed). Smoothness of a labeling $f\in\{\pm1\}^n$ with respect to the graph may be computed via

$$\mathrm{Cut}(f)=\sum_{(u,v)\in E}\mathbb{1}_{\{f_u\neq f_v\}}=\frac{1}{4}\sum_{(u,v)\in E}(f_u-f_v)^2=f^\top Lf \tag{3}$$

where $L=D-A$, the diagonal matrix D contains degrees of the nodes, and A is the adjacency matrix and $f_v\in\{\pm1\}$ is the label in f that corresponds to vertex $v\in V$. This function in (3) counts the number of disagreements in labels at the endpoints of each edge. The value is also known as the size of the cut induced by f (the smallest possible being MinCut). As desired, the function in (3) gives a smaller value to the labelings that are homogenous within the communities.

Unfortunately, the function $\mathrm{Cut}(f)$ is not stable. Further, the cut size is $n-1$ for a star graph, where $n-1$ nodes, labeled as $+1$, are connected to the center node, labeled as -1. The large value of the cut does not capture the simplicity of this labeling, which is only one bit away from being a constant $+1$. Instead, we opt for the indirect definition:

$$F_\kappa = \left\{ f \in \{\pm 1\}^n \ : \ f^\top L f \le \kappa \right\} \tag{4}$$

for $\kappa \ge 0$, and then set

$$\phi(y_1, \ldots, y_n) = \inf_{f \in \mathcal{F}_\kappa} \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\{f_t \ne y_t\}} + \frac{1}{2n} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_\kappa} \sum_{t=1}^{n} f_t \epsilon_t \right] \tag{5}$$

Parameter κ should be larger than the value of MinCut, for otherwise the set F_κ is empty. This gives an interesting algorithm for the prediction problem .... What does this look like?

Well we want to use the strategy

$$\begin{aligned} q_t &= \frac{1}{2} + \frac{n}{2} \ \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \left[\phi(y_1, \ldots, y_{t-1}, -1, \epsilon_{t+1}, \ldots, \epsilon_n) - \phi(y_1, \ldots, y_{t-1}, +1, \epsilon_{t+1}, \ldots, \epsilon_n) \right] \\ &= \frac{1}{2} + \frac{n}{2} \ \mathbb{E}_{\epsilon_{t+1},\ldots,\epsilon_n} \left[\inf_{f \in \mathcal{F}_\kappa} \left\{ \frac{1}{n} \sum_{j=1}^{t-1} \mathbb{1}_{\{f_j \ne y_j\}} + \mathbb{1}_{\{f_t \ne -1\}} + \sum_{j=t+1}^{n} \mathbb{1}_{\{f_j \ne \epsilon_j\}} \right\} \right. \\ &\qquad\qquad \left. - \inf_{f \in \mathcal{F}_\kappa} \left\{ \frac{1}{n} \sum_{j=1}^{t-1} \mathbb{1}_{\{f_j \ne y_j\}} + \mathbb{1}_{\{f_t \ne +1\}} + \sum_{j=t+1}^{n} \mathbb{1}_{\{f_j \ne \epsilon_j\}} \right\} \right] \end{aligned}$$

It turns out that by concentration inequalities, it even suffices to take a single new sample of $\epsilon_{t+1}, \ldots, \epsilon_n$ for round t to compute q_t above. In this case the underlying strategy is peculiar: At time t, to predict label for vertex v_t, we fill seen entries by labels, unseen entries by random ϵ_v's and solve two optimization problems. One with labels set as mentioned and with label of v_t set to -1 we solve for $\inf_{f \in \mathcal{F}_\kappa} \left\{ \frac{1}{n} \sum_{j=1}^{t-1} \mathbb{1}_{\{f_j \ne y_j\}} + \mathbb{1}_{\{f_t \ne -1\}} + \sum_{j=t+1}^{n} \mathbb{1}_{\{f_j \ne \epsilon_j\}} \right\}$. Now we do the optimization with only changing the label of v_t to a $+1$. We can then set q_t by equation above. Here once can view the random signs we draw as a kind of regularization or protection against worst case adversarial future.

Of course two natural questions follow. First, what if outcomes are not binary. We will see this in the following section. Second, what if we did not know the graph in advance or worse yet the graph evolves with time, or more generally what if we didnt have just bit prediction but rather prediction of bit given some input x_t like in the classification setting?