Machine Learning Theory (CS 6783)

Lecture 14: Online Learning, Bit/Dice Prediction and Experts Algorithm

1 Extending Bir Prediction to Binary Classification

Last lectures we looked at the result by Thomas Cover on Bit prediction and its extension to multi-class. We specifically applied these ideas to the mind reading machine style problem and the node prediction problem on a fixed graph known to the learner. A natural question is whether the results can be extended to the classification setting where on any round t we first receive input instance x_t we then pick our prediction and receive true label y_t. Can we extend the node prediction result to the case where the graph is not known in advance? Only revealed on the fly?

To keep things simple lets stick to binary labels. To obtain Cover's result we worked backwards. Say we received $y_1, \ldots, y_{n-1}$ and $x_1, \ldots, x_n$. In this case, we need to pick a q_n such that

$$\frac{1}{n}\mathbb{E}\ \mathbb{1}_{\{\hat{y}_n \neq +1\}} - \phi(x_1, y_1, \ldots, x_n, +1) = \frac{1}{n}\mathbb{E}\ \mathbb{1}_{\{\hat{y}_n \neq -1\}} - \phi(x_1, y_1, \ldots, x_n, -1)$$

This is done just as in the earlier bit prediction case and we get

$$q_n = \frac{1}{2} + \frac{n}{2}\left(\phi(x_1, y_1, \ldots, x_n, -1) - \phi(x_1, y_1, \ldots, x_n, +1)\right)$$

which yields that for any y_n,

$$\frac{1}{n}\mathbb{E}\ \mathbb{1}_{\{\hat{y}_n \neq y_n\}} - \phi(x_1, y_1, \ldots, x_n, y_n) = \frac{1}{2n} - \mathbb{E}_{\epsilon_n}\phi(x_1, y_1, \ldots, x_n, \epsilon_n)$$

Now lets proceed to $n-1$'th round. To do this, notice that we still have to deal with the nth input instance x_n. If x_n can be arbitrarily chosen then the only choice is to take the max over it as the adversary/nature can inflict most loss or regret by choosing the worst x_n. Hence we conclude that for any x_n and any y_n chosen based on q_n,

$$\frac{1}{n}\mathbb{E}\ \mathbb{1}_{\{\hat{y}_n \neq y_n\}} - \phi(x_1, y_1, \ldots, x_n, y_n) = \frac{1}{2n} - \inf_{x_n}\mathbb{E}_{\epsilon_n}\phi(x_1, y_1, \ldots, x_n, \epsilon_n)$$

Great! Looks like we should be in shape: take the RHS above and proceed with the bit prediction style proof just as before:

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\ \mathbb{1}_{\{\hat{y}_{n-1} \neq y_{n-1}\}} + \frac{1}{n}\mathbb{E}\hat{y}_n \sim q_n\ \mathbb{1}_{\{\hat{y}_n \neq y_n\}} - \phi(x_1, y_1, \ldots, x_n, y_n) \\
&\quad = \frac{1}{n}\mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\ \mathbb{1}_{\{\hat{y}_{n-1} \neq y_{n-1}\}} - \inf_{x_n}\mathbb{E}_{\epsilon_n}\phi(x_1, y_1, \ldots, x_n, \epsilon_n)
\end{aligned}$$

Seems like this should yield

$$q_{n-1} = \frac{1}{2} + \frac{n}{2}\mathbb{E}_{\epsilon_n}\left(\inf_{x_n}\phi(x_1, y_1, \ldots, x_n, -1) - \inf_{x_n}\phi(x_1, y_1, \ldots, x_n, +1)\right)$$

and hence,

$$\frac{1}{n}\mathbb{E}_{\hat{y}_{n-1}\sim q_{n-1}}\mathbb{1}_{\{\hat{y}_{n-1}\neq y_{n-1}\}}+\frac{1}{n}\mathbb{E}_{\hat{y}_n\sim q_n}\mathbb{1}_{\{\hat{y}_n\neq y_n\}}-\phi(x_1,y_1,\ldots,x_n,y_n)=\frac{2}{2n}-\inf_{x_{n-1}}\mathbb{E}_{\epsilon_{n-1}}\inf_{x_n}\mathbb{E}_{\epsilon_n}\phi(x_1,y_1,\ldots,x_n,\epsilon_n)$$

and hence finally the result that

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\hat{y}_t\sim q_t}\mathbb{1}_{\{\hat{y}_t\neq y_t\}}-\phi(x_1,y_1,\ldots,x_n,y_n)\leq\frac{1}{2}-\inf_{x_1}\mathbb{E}_{\epsilon_1}\ldots\inf_{x_n}\mathbb{E}_{\epsilon_n}\phi(x_1,y_1,\ldots,x_n,\epsilon_n)$$

Right?

Wrong! We are missing stability.

For the n'th step the stability of ϕ implied q_n was a valid distribution. But think about the $n-1$th round. $\inf_{x_n}\mathbb{E}_{\epsilon_n}\phi(x_1,y_1,\ldots,x_n,\epsilon_n)$ depends on y_{n-1}. So if we change x_n and y_{n-1}, then this time we are varying two terms. This becomes worse at say $t=n/2$ round as $n/2$ future instances can change drastically if we flip $y_{n/2-1}$ and so stability of ϕ doesn't give us much. In fact, general ϕ functions of instance and outcomes has not really been explored in the literature. The special case of ϕ being defined as $\inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_{f(x_t)\neq y_t}$ we will explore in next few lectures. But before that, the story is not all that dire.

Consider the case where x's are drawn iid from some fixed known distribution (if distribution not known use unlabeled data). In this case, since we know how x's are drawn and it does not depend on the labels, at the nth step, we can use: for any y_n chosen based on q_n, in expectation over x_n,

$$\mathbb{E}_{x_n\sim D}\left[\frac{1}{n}\mathbb{E}\,\mathbb{1}_{\{\hat{y}_n\neq y_n\}}-\phi(x_1,y_1,\ldots,x_n,y_n)\right]=\frac{1}{2n}-\mathbb{E}_{x_n\sim D}\left[\mathbb{E}_{\epsilon_n}\phi(x_1,y_1,\ldots,x_n,\epsilon_n)\right]$$

and we can proceed as follows. Now notice that stability wont be an issue. This is because,

$$\begin{aligned}&\frac{n}{2}\left|\mathbb{E}_{x_{t+1:n}\epsilon_{t+1:n}}\left(\phi(x_1,y_1,\ldots,x_t,-1,x_{t+1},\epsilon_{t+1},\ldots,x_n,\epsilon_n)-\phi(x_1,y_1,\ldots,x_t,+1,x_{t+1},\epsilon_{t+1},\ldots,x_n,\epsilon_n)\right)\right|\\&\leq\frac{n}{2}\mathbb{E}_{x_{t+1:n}\epsilon_{t+1:n}}\left|\phi(x_1,y_1,\ldots,x_t,-1,x_{t+1},\epsilon_{t+1},\ldots,x_n,\epsilon_n)-\phi(x_1,y_1,\ldots,x_t,+1,x_{t+1},\epsilon_{t+1},\ldots,x_n,\epsilon_n)\right|\\&\leq\frac{1}{2}\end{aligned}$$

The same idea can be used for the case of the node prediction problem. Say we know the social network (or whatever other graph) evolves according to a know stochastic model (Eg. Preferential attachment model) and does not depend on the labels. In this case, at round t we sample the graph till round n and use this hallucinated graph for making predictions just as before. Cool right!

2 Prediction with (adversarial) Side-information: Experts setting

For $t = 1$ to n

Instance $x_t \in \mathcal{X}$ is provided

Learner picks $\hat{y} \in \mathcal{Y}$ (or randomized version $q_t \in \Delta(\mathcal{Y})$)

True label $y_t \in \mathcal{Y}$ is revealed and learner pays loss $\ell(\hat{y}_t, y_t)$

end

$$\mathbf{R}_n = \frac{1}{n}\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t)$$

If we use randomized algorithm then, on each round, label $\hat{y}_t$ is drawn from q_t. In this case, we wish to bound regret defined as :

$$\mathbf{R}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t, y_t)\right] - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t)$$

A simple application of Hoeffding-Azuma can in fact turn the above statement in to a high probability statement of form, for any $\delta > 0$ with probability at least $1 - \delta$ over the randomization of the learning algorithm,

$$\mathbf{R}_n \leq \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t, y_t)\right] - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t) + \sqrt{\frac{\log 1/\delta}{n}}$$

2.1 Experts/Exponential Weights Algorithm

Algorithm : $q_1(f) = 1/|\mathcal{F}|$. Further, each round we update the distribution over experts as,

$$q_{t+1}(f) \propto q_t(f) e^{-\eta \ell(f(x_t), y_t)}$$

Or in other words, $q_{t+1}(f) = \frac{e^{-\eta\sum_{i=1}^{t}\ell(f(x_t),y_t)}}{\sum_{f\in\mathcal{F}} e^{-\eta\sum_{i=1}^{t}\ell(f(x_t),y_t)}}$

Claim 1.

$$\sum_{t=1}^{n} \mathbb{E}_{f\sim q_t}\left[\ell(f(x_t), y_t)\right] - \min_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \sqrt{\frac{2\log|\mathcal{F}|}{n}}$$

Proof. We use the notation $L_t(f) = \sum_{i=1}^{t} \ell(f(x_i), y_i)$. Define $W_0 = |F|$ and define $W_t = \sum_{f\in\mathcal{F}} e^{-\eta L_t(f)}$. Note that

$$\begin{aligned}
\log\left(\frac{W_n}{W_0}\right) &= \log\left(\sum_{f\in\mathcal{F}} e^{-\eta L_n(f)}\right) - \log|\mathcal{F}| \\
&\geq \log\left(\max_{f\in\mathcal{F}} e^{-\eta L_n(f)}\right) - \log|\mathcal{F}| \\
&= -\eta \min_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \log|\mathcal{F}|
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\log\left(\frac{W_n}{W_0}\right) &= \sum_{t=1}^{n} \log\left(\frac{W_t}{W_{t-1}}\right) = \sum_{t=1}^{n} \log\left(\frac{\sum_{f\in\mathcal{F}} e^{-\eta L_t(f)}}{\sum_{f\in\mathcal{F}} e^{-\eta L_{t-1}(f)}}\right) \\
&= \sum_{t=1}^{n} \log\left(\sum_{f\in\mathcal{F}} \frac{e^{-\eta L_{t-1}(f)}}{\sum_{f\in\mathcal{F}} e^{-\eta L_{t-1}(f)}} e^{-\eta \ell(f(x_t),y_t)}\right) \\
&= \sum_{t=1}^{n} \log\left(\mathbb{E}_{f\sim q_t}\left[e^{-\eta \ell(f(x_t),y_t)}\right]\right) \\
&= \sum_{t=1}^{n} \log\left(\mathbb{E}_{f\sim q_t}\left[e^{-\eta\left(\ell(f(x_t),y_t)-\mathbb{E}_{f\sim q_t}[\ell(f(x_t),y_t)]\right)-\eta\mathbb{E}_{f\sim q_t}[\ell(f(x_t),y_t)]}\right]\right) \\
&= \sum_{t=1}^{n} \log\left(\mathbb{E}_{f\sim q_t}\left[e^{-\eta\left(\ell(f(x_t),y_t)-\mathbb{E}_{f\sim q_t}[\ell(f(x_t),y_t)]\right)}\right] \times e^{-\eta\mathbb{E}_{f\sim q_t}[\ell(f(x_t),y_t)]}\right) \\
&= \sum_{t=1}^{n} \log\left(\mathbb{E}_{f\sim q_t}\left[e^{-\eta\left(\ell(f(x_t),y_t)-\mathbb{E}_{f\sim q_t}[\ell(f(x_t),y_t)]\right)}\right]\right) - \eta\sum_{t=1}^{n} \mathbb{E}_{f\sim q_t}\left[\ell(f(x_t),y_t)\right]
\end{aligned}$$

Thus we conclude that

$$\sum_{t=1}^{n} \mathbb{E}_{f\sim q_t}\left[\ell(f(x_t),y_t)\right] - \min_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t),y_t) \le \frac{\log|\mathcal{F}|}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n} \log\left(\mathbb{E}_{f\sim q_t}\left[e^{-\eta\left(\ell(f(x_t),y_t)-\mathbb{E}_{f\sim q_t}[\ell(f(x_t),y_t)]\right)}\right]\right)$$

Note that for any zero mean RV X in the range $[-1,1]$, $\mathbb{E}\left[e^{-\eta X}\right] \le e^{\eta^2/2}$. Hence,

$$\sum_{t=1}^{n} \mathbb{E}_{f\sim q_t}\left[\ell(f(x_t),y_t)\right] - \min_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t),y_t) \le \frac{\log|\mathcal{F}|}{\eta} + \frac{n\eta}{2}$$

Picking $\eta = \sqrt{2\log|\mathcal{F}|/n}$ concludes the statement. □

2.2 Learning Thresholds

Not learnable, (even in realizable case) why ?