Machine Learning Theory (CS 6783)

Lecture 1 : A Bit of Fun

1 A Game of Mind Reading

Most of you guys would have played games like Rock-Paper-Scissors and Matching-Pennies while growing up. The excitement of these games is in trying to predict the future — the next choice of the opponent. It is the subtle cues from the other player's past behavior that make the game interesting. Does the opponent tend to play "Rock" after losing with "Scissors"?, do they try to play more heads than tails?, does the opponent tend to stick with the same choice after winning a round? We try to notice such patterns in behavior to tip the balance in our favor.

For the simplest concrete example let us consider the penny matching game. In this game both players on each round place their coins simultaneously and can choose to have either heads or tails faced up. If both coins match (either both heads or tails) then player 1 takes both pennies, if not player 2 gets both pennies. The game of each player can be seen as predicting what the other player would play and either match it (for player 1) or flip it (for player 2).

What is the optimal strategy for a player in this game?

What should the optimal strategy make or loose typically when playing n rounds of this game with an opponent? What would the opponent make or loose typically?

To introduce some formalism, we shall think of game proceeding for n rounds and variable t indicates the current round. On each round we will call player 1 the learner and player 1's prediction for round t we will denote by $\hat{y}_t \in \{\pm 1\}$. Player 2 we will call nature or adversary and denote their prediction as $y_t \in \{\pm 1\}$. The amount lost by player 1 (matching player) is then given by

$$\sum_{t=1}^{n} \mathbf{1}\{y_t \neq \hat{y}_t\} - \sum_{t=1}^{n} \mathbf{1}\{y_t = \hat{y}_t\} = 2\sum_{t=1}^{n} \mathbf{1}\{y_t \neq \hat{y}_t\} - n = 2\left(\sum_{t=1}^{n} \mathbf{1}\{y_t \neq \hat{y}_t\} - \frac{n}{2}\right)$$

Human beings are notorious at being bad at coming up with random sequences of bits. This led to a famous (informal, in house) competition between David Hagelbarger and Claude Shannon in the Bell Labs in the 1950's. The two wanted to design computer programs to beat humans at the game of penny matching. While at AT&T Bell Labs, they each built a machine—aptly called "mind reader"—to play the game of Matching-Pennies. According to various accounts, the machines were able to predict the sequence of heads/tails entered by an untrained human markedly better than random guess, picking up on a variety of patterns of the past play. What would have happened if the two machines were played against each other is still a mystery to us. But let us try to ask the question of how do we start with designing such algorithms?

Figure 1: Shannon's Mind Reading Machine, MIT Museum. (Source: http://william-poundstone.com/blog/2015/7/30/how-i-beat-the-mind-reading-machine)

Lets set our aims low first, say you wanted to develop an algorithm that minimizes the following notion of regret against not picking the majority over the n rounds.

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

If we could always make this quantity smaller than 0, then this would mean we have a strategy that only makes money and never looses it in expectation. But unfortunately this is not possible, a random adversary would inflict $\sqrt{n}$ loss with constant probability. **Why?**

But let us just ask to make this quantity as small as we can against any adversary. When y_t's are drawn from a fixed distribution I claim this is easy, Why, what is the strategy?

When the bits are not drawn iid, this problem is far more complicated and interesting.

Can you guys come up with strategies?

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 $\min_{b \in \{\pm 1\}} \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{y_t \neq b\} = 1/3$. 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 regret to be $o(n)$?

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

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \leq \sqrt{n}$$

against any adversary!

Specifically this means that we have a strategy that never looses worse than $\sqrt{n}$ against any adversary (which is the best we could hope for even for optimal) and further, if we have uneven number of heads than tails, we can win significantly more.

To prove the above claim and much more, we first prove this following lemma, a result by Thomas Cover (all the way back in 1965). In fact, the more general question we will answer will be roughly in the form: For what function ϕ's is it possible to ensure that, there exists forecaster s.t.,

```
for any sequence,
        number of mistakes made by forecaster $\leq \phi$(sequence).
```

The function ϕ controlling the number of mistakes is a measure of "complexity" or "predictiveness" of the sequence. It captures our prior belief of what kinds of patterns might appear. For the Penny-Matching game, ϕ may be related to the frequency of heads vs tails, or more fine-grained statistics, such as predictability of the next outcome based on the last three outcomes. In fact, Shannon's mind reading machine was based on only 8 such states. Which ϕ can one choose? How to develop an efficient algorithm for a given ϕ?

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:

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

Once we have the above lemma, using $\phi(y_1, \ldots, y_n) = \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{f_t \neq y_t\}} + \frac{1}{2n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t\right]$ which satisfies the stability condition, we can conclude the result.