Machine Learning Theory (CS 6783)

Lecture 1 : Learning Frameworks, a Bit of Fun

1 Setting up learning problems

1. $\mathcal{X}$ **: instance space or input space**
   Examples:
   - Computer Vision: Raw $M \times N$ image vectorized $\mathcal{X} = [0, 255]^{M \times N}$, SIFT features (typically $\mathcal{X} \subseteq \mathbb{R}^d$)
   - Speech recognition: Mel Cepstral co-efficients $\mathcal{X} \subset \mathbb{R}^{12 \times \text{length}}$
   - Natural Language Processing: Bag-of-words features ($\mathcal{X} \subset \mathbb{N}^{\text{document size}}$), n-grams
2. $\mathcal{Y}$**: Outcome space, label space**
   Examples: Binary classification $\mathcal{Y} = \{\pm 1\}$, multiclass classification $\mathcal{Y} = \{1, \ldots, K\}$, regression $\mathcal{Y} \subset \mathbb{R}$)
3. $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$**: loss function** (measures prediction error)
   Examples: Classification $\ell(y', y) = \mathbb{1}_{\{y' \neq y\}}$, Support vector machines $\ell(y', y) = \max\{0, 1 - y' \cdot y\}$, regression $\ell(y', y) = (y - y')^2$
4. $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$**: Model/ Hypothesis class** (set of functions from input space to outcome space)
   Examples:
   - Linear classifier: $\mathcal{F} = \{x \mapsto \text{sign}(f^\top x) : f \in \mathbb{R}^d\}$
   - Linear SVM: $\mathcal{F} = \{x \mapsto f^\top x : f \in \mathbb{R}^d, \|f\|_2 \leq R\}$
   - Neural Netoworks (deep learning): $\mathcal{F} = \{x \mapsto \sigma(W_{out}\sigma(W_K\sigma(\ldots\sigma(W_2(W_1\sigma(W_{in}x)))))\}$ where σ is some non-linear transformation (Eg. ReLU)

Learner observes sample: $S = (x_1, y_1), \ldots, (x_n, y_n)$

1.1 Statistical Learning

Generic $\mathcal{X}$, $\mathcal{Y}$, ℓ and $\mathcal{F}$

Samples generated as $(x_1, y_1), \ldots, (x_n, y_n) \sim D$ where D is some unknown distribution over $\mathcal{X} \times \mathcal{Y}$.
Goal: Find $\hat{\mathbf{y}}$ that minimizes

$$\mathbb{E}_{(x,y)\sim D}\left[\ell(\hat{\mathbf{y}}(x), y)\right] - \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim D}\left[\ell(f(x), y)\right]$$

For any mapping $g : \mathcal{X} \mapsto \mathcal{Y}$ we shall use the notation $L_D(g) = \mathbb{E}_{(x,y)\sim D}\left[\ell(g(x), y)\right]$ and so our goal is to minimize:

$$L_D(\hat{\mathbf{y}}) - \inf_{f\in\mathcal{F}} L_D(f)$$

Remarks: $\hat{\mathbf{y}}$ is a random quantity as it depends on the sample

1.1.1 PAC framework (Realizability)

$$\mathcal{Y} = \{\pm 1\}, \quad \ell(y', y) = \mathbb{1}_{\{y'\neq y\}}$$

Input instances generated as $x_1, \ldots, x_n \sim D_X$ where D_X is some unknown distribution over input space. The labels are generated as

$$y_t = f^*(x_t)$$

where target function $f^* \in \mathcal{F}$. Learning algorithm only gets sample S and does not know f^* or D_X.

Goal: Find $\hat{\mathbf{y}}$ that minimizes

$$\mathbb{P}_{x\sim D_X}\left(\hat{\mathbf{y}}(x) \neq f^*(x)\right)$$

1.2 Online Learning Framework

A multi-round game between nature and learner:

For $t = 1$ to n

Nature produces x_t

Learner predicts $\hat{y}_t$

Nature produces label y_t

End For

Goal: Minimize regret w.r.t. model class $\mathcal{F}$ defined as:

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

2 Prelude: Bit Prediction

2.1 Statistical Learning

We consider as a warmup example, the simplest statistical learning/prediction problem. That of learning coin flips! Let us consider the case where we don't receive any input instance (or $\mathcal{X} = \{\}$) and $\mathcal{Y} = \{\pm 1\}$. We receive ± 1 valued samples $y_1, \ldots, y_n \in \{\pm 1\}$ drawn iid from Bernoullis distribution with parameter p (ie. Y is $+1$ with probability p and -1 with probability $1 - p$). Our loss function is the zero-one loss function $\ell(y', y) = \mathbb{1}_{\{y'\neq y\}}$. Recall that our goal in statistical learning is to minimize $L_p(\hat{y}) - \inf_{f\in\{\pm 1\}} L_p(f)$. (Effectively our only choice of $\mathcal{F}$ for this problem is the set of constant mappings, $\mathcal{F} = \{\pm 1\}$). If 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 works well. Specifically, if we used this choice, it is not hard to show that:

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

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. This result is also true with high probability. (Excercise: Try to formally show this.)

But here we made the crucial assumption that bits were drawn iid from a Bernoulli distribution. What if this were not true? Can we even hope to handle this problem of bit prediction in the online setting?

2.2 Online Setting: Mind Reading Machine

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. Of course, if opponent is random, there is no good strategy, and the game becomes boring. This boring strategy is in fact minimax optimal. However, it is the subtle cues from the other player and their 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.

Can we program a computer to beat humans at these games? sThis question was asked by Claude Shannon and David Hagelbarger in the 1950's. 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.

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)

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 $\min_{b\in\{\pm 1\}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}\{y_t \neq b\} = 1/3$. 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 average regret to diminish to 0 with n?

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

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

Specifically, this means that if we want to do as well as majority, That is $\mathcal{F} = \{\pm 1\}$ the two constant predictions of only heads or only tails, then we can easily conclude that just like in the statistical learning setting for this problem, regret can be bounded by $O(1/\sqrt{n})$!

To prove the above claim 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 ≤ φ(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.

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,

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

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 $\mathbb{E}_{\hat{y}_{n-1} \sim q_{n-1}}\left[\ \mathbb{1}_{\{\hat{y}_{n-1} \neq y_{n-1}\}}\right] = 1 - q_{n-1}$ 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,

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

Thus we can conclude that,

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

Proceeding in similar way we conclude that,

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

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. □