Machine Learning Theory (CS 6783)

Lecture 3: Cover's Result, Rademacher Complexity and Betting Game

Cover's result is remarkable because it says that if a bound on error of some stable ϕ is possible under random coin flips, then such bound is possible under any adversary!

So using this result, we will as promised prove our claim, but more generally we can ask what are examples of stable ϕ's and how far can we take this machinery. While we set our initial goal low, now its time to consider more general case of regret against a class of prefixed models $\mathcal{F}$. Instead of majority (which is either all $+1$ or all -1 in hindsight), consider an arbitrary set $\mathcal{F} \subset \{\pm 1\}^n$. We would like to come up with a strategy that minimizes regret with respect to any choice in $\mathcal{F}$ as:

$$\mathrm{Reg}_n(\mathcal{F}) := \sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\, \mathbb{1}_{\{\hat{y}_t \neq y_t\}}\right] - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbb{1}_{\{f_t \neq y_t\}}$$

Our goal specifically is to answer, what is $C_n(\mathcal{F}) > 0$ such that, there exists a prediction algorithm such that against any adversary this algorithm can guarantee that:

$$\mathrm{Reg}_n(\mathcal{F}) \leq C_n(\mathcal{F})$$

To this end we can use Cover's lemma above with

$$\phi(y_1, \ldots, y_n) = \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbb{1}_{\{f_t \neq y_t\}} + C_n(\mathcal{F})$$

We will soon show that this function is stable but for now if you trust me that it is stable, note that cover's result tells us the smallest value of $C_n(\mathcal{F})$. It is given by

$$\begin{aligned}
C_n(\mathcal{F}) &= \frac{n}{2} - \mathbb{E}_\epsilon\left[\min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbb{1}_{\{f_t \neq \epsilon_t\}}\right] \\
&= \frac{n}{2} + \mathbb{E}_\epsilon\left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} - \mathbb{1}_{\{f_t \neq \epsilon_t\}}\right] \\
&= \frac{n}{2} + \mathbb{E}_\epsilon\left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} -\frac{1}{2}(1 - f_t \epsilon_t)\right] \\
&= \frac{n}{2} + \frac{1}{2}\mathbb{E}_\epsilon\left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t - 1\right] \\
&= \frac{1}{2}\mathbb{E}_\epsilon\left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t\right]
\end{aligned}$$

Claim 1. *For any $\mathcal{F} \subseteq \{\pm 1\}^n$, the function*

$$\phi(y_1, \dots, y_n) = \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbb{1}_{\{f_t \neq y_t\}} + C_n(\mathcal{F})$$

is stable.

Proof. W.l.o.g. pick the last bit to flip, we find,

$$\begin{aligned}
&\phi(y_1, \dots, y_{n-1}, +1) - \phi(y_1, \dots, y_{n-1}, -1) \\
&\qquad = \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} + \mathbb{1}_{\{f_n \neq +1\}} \right\} - \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} + \mathbb{1}_{\{f_n \neq -1\}} \right\} \\
&\qquad = \max_{f' \in \mathcal{F}} \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} + \mathbb{1}_{\{f_n \neq +1\}} - \sum_{t=1}^{n-1} \mathbb{1}_{\{f'_t \neq y_t\}} - \mathbb{1}_{\{f'_n \neq -1\}} \right\} \\
&\qquad \leq \max_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} + \mathbb{1}_{\{f_n \neq +1\}} - \sum_{t=1}^{n-1} \mathbb{1}_{\{f_t \neq y_t\}} - \mathbb{1}_{\{f_n \neq -1\}} \right\} \\
&\qquad = \max_{f \in \mathcal{F}} \left\{ \mathbb{1}_{\{f_n \neq +1\}} - \mathbb{1}_{\{f_n \neq -1\}} \right\} \leq 1
\end{aligned}$$

Similarly we have, $\phi(y_1, \dots, y_{n-1}, -1) - \phi(y_1, \dots, y_{n-1}, +1) \leq 1$ and so we have stability. □

Putting this together we have:

Lemma 2. *There exists a randomized prediction strategy that ensures that*

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbf{1}\{f_t \neq y_t\} \leq \frac{1}{2} \mathbb{E}_\epsilon \left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f_t\right]$$

against any adversary and any $\mathcal{F} \subset \{\pm 1\}^n$

Corollary 3. *There exists a randomized prediction strategy that ensures that*

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \min_{b \in \{\pm 1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq y_t\} \leq \frac{\sqrt{n}}{2}$$

1 Rademacher Complexity

Given a set $\mathcal{F} \subseteq \mathbb{R}^n$, we define the Rademcher complexity of this set as

$$\mathcal{R}_n(\mathcal{F}) = \frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t\right]$$

While we have already seen the Rademacher complexity as coming from cover's result, it turns out that this quantity or rather complexity measure is a key tool in Statistical learning theory.

Hence lets try to see what the quantity represents. Note that if $\mathcal{F}$ was binary labels, then for any vector $f \in \mathcal{F}$, $\|f\|_2 = \sqrt{n}$ and $\|\epsilon\|_2 = \sqrt{n}$. Hence we can interpret,

$$\frac{1}{n}\sum_{t=1}^{n} f_t \epsilon_t = \frac{1}{n} f^\top \epsilon = \frac{f^\top \epsilon}{\|f\|_2 \|\epsilon\|_2} = \cos(\epsilon, f)$$

Hence, we can think of $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}} \cos(\epsilon, f)\right]$, that is, how well we can correlate with random draw of labels using set $\mathcal{F}$.

Now before we go into statistical learning, let us get back to our bit prediction problem.

2 A Game of Betting

In the previous section, we assumed ϕ was stable. While stable ϕ's consist of a large number of benchmark, it might not be expressive enough for some problems. Unfortunately, if we want to do classification, it is not easy to get rid of such an assumption easily. Instead below we consider a slightly different betting game on binary outcomes where we are allowed to bet arbitrary amounts on outcomes. In such game, the same idea as above can be used without requiring stability.

Consider a gambler who bets on the outcomes of games one every round. Specifically, on any round t, the gambler can choose an amount $|\hat{y}_t|$ to bet on the outcome of game between two players or teams A and B. The gambler can choose to place this bet of $|\hat{y}_t|$ on either team A to win or on team B. If the chosen team wins, the gambler gains an additional amount of $\hat{y}_t$ and if the chosen team looses the gambler looses the bet amount of $\hat{y}_t$. This game of betting can be formalized as the following linear game between the gambler and the house. Specifically, we can view the choice of the gambler at round t as a real number $\hat{y}_t$. The magnitude $\hat{y}_t$ denotes the bet amount and the sign of $\hat{y}_t$ denotes whether the bet is placed on team A or team B. The corresponding outcome of the game is encoded by the variable $y_t \in \{\pm 1\}$ which indicates whether team A won or team B. At time t, $-\hat{y}_t \cdot y_t$ denotes the loss of the gambler. That is if the gambler guessed the outcome right, that is if $\text{sign}(\hat{y}_t) = y_t$, then the loss is the negative value of $-|\hat{y}_t|$ (or in other words the gambler gains) and if the outcome is guessed in correctly the gambler looses the amount of $|\hat{y}_t|$.

> At time $t = 1, \ldots, n$, the forecaster chooses $\hat{y}_t \in \mathbb{R}$ based on the history $y_1, \ldots, y_{t-1}$ and then observes the value $y_t \in \{\pm 1\}$.

Given some benchmark function $\phi : \{\pm 1\}^n \to \mathbb{R}_{\geq 0}$,, the goal of the gambler is to ensure that the loss of the gambler is smaller than this benchmark. In other words, the gambler would like to ensure that,

$$\forall \mathbf{y} \quad \sum_{t=1}^{n} -\hat{y}_t y_t \leq \phi(\mathbf{y}) \tag{1}$$

Lemma 4. *ϕ is achievable if and only if $\mathbb{E}\left[\phi(\epsilon_1, \ldots, \epsilon_n)\right] \geq 0$. Further, in this case, the strategy for the gambler is given by:* $\hat{y}_t = \frac{1}{2} \cdot \mathbb{E}[\phi(y_{1:t-1}, -1, \varepsilon_{t+1:n}) - \phi(y_{1:t-1}, +1, \varepsilon_{t+1:n})]$.

Remark: stability is not required.

Proof. Note that for any $y_1, \ldots, y_{n-1}$ and $y_n \in \{\pm 1\}$,

$$-\hat{y}_n \cdot y_n - \phi(y_{1:n}) = -\frac{y_n}{2}\left(\phi(y_{1:n-1}, -1) - \phi(y_{1:n-1}, +1)\right) - \phi(y_{1:n})$$

Hence, if $y_n = +1$, then

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

Similarly when $y_n = -1$,

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

Thus, for any y_n,

$$-\hat{y}_n \cdot y_n - \phi(y_{1:n}) = -\mathbb{E}_{\epsilon_n}[\phi(y_{1:n-1}, \epsilon_n)]$$

Next proceeding to $n-1$ we see that for the strategy prescribed, for any y_{n-1},

$$-\hat{y}_{n-1} \cdot y_{n-1} - \mathbb{E}_{\epsilon_n}[\phi(y_{1:n-1}, \epsilon_n)] = -\mathbb{E}_{\epsilon_{n-1}, \epsilon_n}[\phi(y_{1:n-2}, \epsilon_{n-1}, \epsilon_n)]$$

Thus continuing we conclude that:

$$\sum_{t=1}^{n} -\hat{y}_t y_t - \phi(\mathbf{y}) = -\mathbb{E}_{\epsilon}[\phi(\epsilon_1, \ldots, \epsilon_n)] \leq 0$$

from our premise and so we have proved the lemma. □

Example 2.1. *We have a gambler who likes to bet on games played between m teams. Assume that the information about which pairs of teams play each other for the n matches is announced in advance. Specifically, say we know that on round t, teams i_t and j_t play each other. Let us further denote by n_i the number of games played by player i. This game of betting can be formalized in the linear betting games framework above. As specific benchmark a gambler might consider is the one where each of the m team is given a score represented by an m dimensional vector $\mathbf{w}$. Further, when team i plays team j, a bet of amount of $|w[i] - w[j]|$ on the team with the larger score is placed. Further, assume that the largest bet amount is restricted to B. The goal of the gambler is to do as well as the best scoring of the teams selected in hindsight. This example, can be represented by the benchmark $\phi\{\pm 1\}^n \mapsto \mathbb{R}$ as follows:*

$$\begin{aligned} \phi(y_1, \ldots, y_n) &= \inf_{\mathbf{w} \in \mathbb{R}^m : \max_{i,j} \mathbf{w}[i] - \mathbf{w}[j] \leq B} \frac{1}{n} \sum_{t=1}^{n} y_t \cdot (\mathbf{w}[i_t] - \mathbf{w}[j_t]) + \frac{B}{2n} \sum_{i=1}^{m} \sqrt{n_i} && (2) \\ &\leq \inf_{\mathbf{w} \in \mathbb{R}^m : \max_{i,j} \mathbf{w}[i] - \mathbf{w}[j] \leq B} \frac{1}{n} \sum_{t=1}^{n} y_t \cdot (\mathbf{w}[i_t] - \mathbf{w}[j_t]) + \frac{B}{2}\sqrt{\frac{m}{n}} && (3) \end{aligned}$$

This benchmark satisfies the property that $\mathbb{E}\left[\phi(\epsilon_1,\dots,\epsilon_n)\right] \geq 0$. *This is because*

$$\begin{aligned}
\mathbb{E}\left[\phi(\epsilon_1,\dots,\epsilon_n)\right] &= \mathbb{E}\left[\inf_{\mathbf{w}\in\mathbb{R}^m:\max_{i,j}\mathbf{w}[i]-\mathbf{w}[j]\leq B} \frac{1}{n}\sum_{t=1}^{n} y_t\cdot(\mathbf{w}[i_t]-\mathbf{w}[j_t])\right] + \frac{B}{2n}\sum_{i=1}^{m}\sqrt{n_i} \\
&= \mathbb{E}\left[\inf_{\mathbf{w}\in[0,B]^m} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t(\mathbf{w}[i_t]-\mathbf{w}[j_t])\right] + \frac{B}{2n}\sum_{i=1}^{m}\sqrt{n_i} \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{m}\min_{\mathbf{w}[i]\in[0,B]}\sum_{t=1}^{n}\mathbf{w}[i]\epsilon_t\left(\mathbb{1}_{\{i_t=i\}}-\mathbb{1}_{\{j_t=i\}}\right)\right] + \frac{B}{2n}\sum_{i=1}^{m}\sqrt{n_i} \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{m}\min\left\{B\sum_{t=1}^{n}\epsilon_t\left(\mathbb{1}_{\{i_t=i\}}-\mathbb{1}_{\{j_t=i\}}\right),0\right\}\right] + \frac{B}{2n}\sum_{i=1}^{m}\sqrt{n_i} \\
&= \frac{B}{n}\sum_{i=1}^{m}\mathbb{E}\left[\min\left\{\sum_{j=1}^{n_i}\epsilon_j,0\right\}\right] + \frac{B}{2n}\sum_{i=1}^{m}\sqrt{n_i} \\
&\geq -\frac{B}{2n}\sum_{i=1}^{m}\sqrt{n_i} + \frac{B}{2n}\sum_{i=1}^{m}\sqrt{n_i} = 0
\end{aligned}$$

where in the last line we used the fact that for any integer N, $\mathbb{E}\left[\min\left\{\sum_{j=1}^{N}\epsilon_j,0\right\}\right] \geq -\sqrt{N}/2$. *Hence, from Lemma 4 this benchmark is achievable by the gambler using the strategy* $\hat{y}_t = n\cdot\mathbb{E}[\phi(y_{1:t-1},-1,\varepsilon_{t+1:n})-\phi(y_{1:t-1},+1,\varepsilon_{t+1:n})]$. *Finally, noting that square-root is a concave function and applying Jensen's inequality, yields that* $\frac{B}{2n}\sum_{i=1}^{m}\sqrt{n_i} \leq \frac{B}{2}\sqrt{\frac{m}{n}}$.