18.657: Mathematics of Machine Learning

Lecturer: Alexander Rakhlin
Scribe: Kevin Li

Lecture 19
Nov. 16, 2015

4. PREDICTION OF INDIVIDUAL SEQUENCES

In this lecture, we will try to predict the next bit given the previous bits in the sequence. Given completely random bits, it would be impossible to correctly predict more than half of the bits. However, certain cases including predicting bits generated by a human can be correct greater than half the time due to the inability of humans to produce truly random bits. We will show that the existence of a prediction algorithm that can predict better than a given threshold exists if and only if the threshold satifies certain probabilistic inequalities. For more information on this topic, you can look at the lecture notes at http://stat.wharton.upenn.edu/~rakhlin/courses/stat928/stat928_notes.pdf

4.1 The Problem

To state the problem formally, given a sequence $y_1, \dots, y_n, \dots \in \{-1, +1\}$, we want to find a prediction algorithm $\hat{y}_t = \hat{y}_t(y_1, \dots, y_{t-1})$ that correctly predicts y_t as much as possible.

In order to get a grasp of the problem, we will consider the case where $y_1, \dots, y_n \overset{iid}{\sim} Ber(p)$. It is easy to see that we can get

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}\right] \to \min\{p, 1-p\}$$

by letting $\hat{y}_t$ equal majority vote of the first $t-1$ bits. Eventually, the bit that occurs with higher probability will always have occurred more times. So the central limit theorem shows that our loss will approach $\min\{p, 1-p\}$ at the rate of $O(\frac{1}{\sqrt{n}})$.

Knowing that the distribution of the bits are iid Bernoulli random variables made the prediction problem fairly easy. More surprisingly is the fact that we can achieve the same for any individual sequence.

Claim: There is an algorithm such that the following holds for any sequence $y_1, \dots, y_n, \dots$.

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\} - \min\{\bar{y}_n, 1-\bar{y}_n\} \leq 0 \text{ a.s.}$$

It is clear that no deterministic strategy can achieve this bound. For any deterministic strategy, we can just choose $y_t = -\hat{y}_t$ and the predictions would be wrong every time. So we need a non-deterministic algorithm that chooses $\hat{q}_t = \mathbb{E}[\hat{y}_t] \in [-1, 1]$.

To prove this claim, we will look at a more general problem. Take a fixed horizon $n \geq 1$, and function $\phi : \{\pm 1\}^n \to \mathbb{R}$. Does there exist a randomized prediction strategy such that for any $y_1, \dots, y_n$

$$\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}] \leq \phi(y_1, \dots, y_n) \ ?$$

For certain ϕ such as $\phi \equiv 0$, it is clear that no randomized strategy exists. However for $\phi \equiv \frac{1}{2}$, the strategy of randomly predicting the next bit ($\hat{q}_t = 0$) satisfies the inequality.

Lemma: For a stable ϕ, the following are equivalent

a) $\exists(\hat{q}_t)_{t=1,\dots,n} \forall y_1, \dots, y_n \quad \mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}] \leq \phi(y_1, \dots, y_n)$

b) $\mathbb{E}[\phi(\epsilon_1, \dots, \epsilon_n)] \geq \frac{1}{2}$ where $\epsilon_1, \dots, \epsilon_n$ are Rademacher random variables

where stable is defined as follows

Definition (Stable Function): A function $\phi : \{\pm 1\}^n \to \mathbb{R}$ is stable if

$$|\phi(\dots, y_i, \dots) - \phi(\dots, -y_i, \dots)| \leq \frac{1}{n}$$

Proof. ($a \implies b$) Suppose $\mathbb{E}\phi < \frac{1}{2}$. Take $(y_1, \dots, y_n) = (\epsilon_1, \dots, \epsilon_n)$. Then $\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq \epsilon_t\}] = \frac{1}{2} > \mathbb{E}[\phi]$ so there must exist a sequence $(\epsilon_1, \dots, \epsilon_n)$ such that $\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq \epsilon_t\}] > \phi(\epsilon_1, \dots, \epsilon_n)$.

($b \implies a$) Recursively define $V(y_1, \dots, y_t)$ such that $\forall y_1, \dots, y_n$

$$V(y_1, \dots, y_{t-1}) = \min_{q_t \in [-1,1]} \max_{y_t \in \pm 1} \left(\frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{y}_t \neq y_t\}] + V(y_1, \dots y_n)\right)$$

Looking at the definition, we can see that $\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}] = V(\emptyset) - V(y_1, \dots, y_n)$. Now we note that $V(y_1, \dots, y_t) = -\frac{t}{2n} - \mathbb{E}[\phi(y_1, \dots, y_t, \epsilon_{t+1}, \dots, \epsilon_n)]$ satisfies the recursive definition since

$$\begin{aligned}
&\min_{\hat{q}_t} \max_{y_t} \frac{1}{n}\mathbb{E}[\mathbb{1}\{\hat{y}_t \neq y_t\}] - \mathbb{E}[\phi(y_1, \dots, y_t, \epsilon_{t+1}, \dots, \epsilon_n)] - \frac{t}{2n} \\
=&\min_{\hat{q}_t} \max_{y_t} \frac{-\hat{q}_t y_t}{2n} - \mathbb{E}[\phi(y_1, \dots, y_t, \epsilon_{t+1}, \dots, \epsilon_n)] - \frac{t-1}{2n} \\
=&\min_{\hat{q}_t} \max\{-\frac{\hat{q}_t}{2n} - \mathbb{E}[\phi(y_1, \dots, y_{t-1}, 1, \epsilon_{t+1}, \dots, \epsilon_n)] - \frac{t-1}{2n}, \frac{\hat{q}_t}{2n} - \mathbb{E}[\phi(y_1, \dots, y_{t-1}, -1, \epsilon_{t+1}, \dots, \epsilon_n)] - \frac{t-1}{2n}\} \\
=&- \mathbb{E}[\phi(y_1, \dots, y_{t-1}, \epsilon_t, \epsilon_{t+1}, \dots, \epsilon_n)] - \frac{t-1}{2n} \\
=&V(y_1, \dots, y_{t-1})
\end{aligned}$$

The first equality uses the fact that for $a, b \in \{\pm 1\}$, $\mathbb{1}\{a \neq b\} = \frac{1-ab}{2}$, the second uses the fact that $y_t \in \{\pm 1\}$, the third minimizes the entire expression by choosing $\hat{q}_t$ so that the two expressions in the max are equal. Here the fact that ϕ is stable means $\hat{q}_t \in [-1, 1]$ and is the only place where we need ϕ to be stable.

Therefore we have

$$\mathbb{E}[\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{\hat{y}_t \neq y_t\}] = V(\emptyset) - V(y_1, \dots, y_n) = -\mathbb{E}[\phi(\epsilon_1, \dots, \epsilon_n)] + \frac{1}{2} + \phi(y_1, \dots, y_n) \leq \phi(y_1, \dots, y_n)$$

by b). □

By choosing $\phi = \min\{\bar{y}, 1-\bar{y}\} + \frac{c}{\sqrt{n}}$, this shows there is an algorithm that satisfies our original claim.

4.2 Extensions

4.2.1 Supervised Learning

We can extend the problem to a regression type problem by observing x_t and trying to predict y_t. In this case, the objective we are trying to minimize would be

$$\frac{1}{n}\sum l(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum l(f(x_t), y_t)$$

It turns out that the best achievable performance in such problems is governed by martingale (or, sequential) analogues of Rademacher averages, covering numbers, combinatorial dimensions, and so on. Much of Statistical Learning techniques extend to this setting of online learning. In addition, the minimax/relaxation framework gives a systematic way of developing new prediction algorithms (in a way similar to the bit prediction problem).

4.2.2 Equivalence to Tail Bounds

We can also obtain probabilistic tail bound on functions ϕ on hypercube by using part a) of the earlier lemma. Rearranging part a) of the lemma we get $1 - 2\phi(y_1, \ldots, y_n) \le \frac{1}{n}\sum \hat{q}_t y_t$. This implies

$$\mathbb{P}\big(\phi(\epsilon_1, \ldots, \epsilon_n) < \frac{1-\mu}{2}\big) = \mathbb{P}\big(1 - 2\phi(\epsilon_1, \ldots, \epsilon_n) > \mu\big) \le \mathbb{P}\big(\frac{1}{n}\sum \hat{q}_t\epsilon_t > \mu\big) \le e^{-\frac{\mu^2}{2n}}$$

So $\mathbb{E}\phi \ge \frac{1}{2} \implies$ existence of a strategy $\implies$ tail bound for $\phi < \frac{1}{2}$.

We can extend the results to higher dimensions. Consider $z_1, \ldots, z_n \in B_2$ where B_2 is a ball in a Hilbert space. We can define recursively $\hat{y}_0 = 0$ and $\hat{y}_{t+1} = \mathrm{Proj}_{B_2}(\hat{y}_t - \frac{1}{\sqrt{n}} z_t)$. Based on the properties of projections, for every $y^* \in B_2$, we have $\frac{1}{n}\sum\langle \hat{y}_t - y^*, z_t\rangle \le \frac{1}{\sqrt{n}}$. Taking $y^* = \dfrac{\sum z_t}{\|\sum z_t\|}$,

$$\forall z_1, \ldots, z_n, \quad \|\sum_{t=1}^n z_t\| - \sqrt{n} \le \sum_{t=1}^n \langle \hat{y}_t, -z_t\rangle$$

Take a martingale difference sequence $Z_1, \ldots, Z_n$ with values in B_2. Then

$$\mathbb{P}\big(\|\sum_{t=1}^n Z_t\| - \sqrt{n} > \mu\big) \le \mathbb{P}(\sum_{t=1}^n \langle \hat{y}_t, -Z_t\rangle > \mu) \le e^{-\frac{n\mu^2}{2}}$$

Integrating out the tail,

$$\mathbb{E}\|\sum_{t=1}^n Z_t\| \le c\sqrt{n}$$

It can be shown using Von Neumann minimax theorem that

$$\exists(\hat{y}_t)\forall z_1,\ldots,z_n,y^* \in B_2 \quad \sum_{t=1}^{n}\langle \hat{y}_t - y^*, z_t\rangle \;\leq \sup_{\mathrm{MDS}W_1,\ldots,W_n} E\|\sum_{t=1}^{n} W_t\| \leq c\sqrt{n}$$

where the supremum is over all martingale difference sequences (MDS) with values in B_2. By the previous part, this upper bound is $c\sqrt{n}$. We conclude an interesting equivalence of (a) deterministic statements that hold for all sequences, (b) tail bounds on the size of a martingale, and (c) in-expectation bound on this size.

In fact, this connection between probabilistic bounds and existence of prediction strategies for individual sequences is more general and requires further investigation.

18.657 Mathematics of Machine Learning
Fall 2015