CIS 320: Algorithms November 15, 2021

Lecture 20

Lecturer: Aaron Roth *Scribe: Aaron Roth*

Calibrated Prediction

Suppose you turn on your local morning show, and the weatherman tells you that tomorrow there is a 10% chance of rain in your neighborhood. What does this mean? Tomorrow will only happen once, so this is not a repeatable event. If it rains, this is not an indictment of the weatherman — he did allow that there was some chance that it would. So how can you distinguish between a weatherman who knows what he is doing, from one who does not?

Lets write down a simple model — the weather prediction game. In rounds $t = 1$ to T:

1. The prediction player predicts some probability p_t of rain, for $p_t \in \{1/m, 2/m, \ldots, (m-1)/m, 1\}$.
2. The outcome $y_t \in \{0, 1\}$ is revealed: it either rains ($y_t = 1$) or it does not ($y_t = 0$).

Lets think about devising a test to determine whether the weatherman knows what he is doing. First, what should this mean? Suppose that there was really some true probabilistic process that governed rain, that the weatherman was privy to: every day, a probability p_t^* was revealed to the weatherman, and then it rained with that probability: $\Pr[y_t = 1] = p_t^*$. We would want that a weatherman who predicted $p_t = p_t^*$ every day should pass the test. Lets call this the oracular weatherman. It should also be possible to fail the test.

Here is a first attempt:

Definition 1 (Average Consistency) *A prediction strategy satisfies ϵ average consistency if for every sequence of outcomes, the sequence of predictions it generates $(p_1, y_1, \ldots, p_T, y_T)$ satisfies*

$$\mathrm{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T} p_t - \sum_{t=1}^{T} y_T\right|\right] \leq \epsilon$$

We say it satisfies average consistency if $\epsilon \to 0$ as $T \to \infty$.

Certainly the oracular weatherman would pass this test, but its also clear that this is not stringent enough, because the following strategy ("The yesterday weatherman") also passes the test: "On day 1, predict $p_t = 0$, and on day t, predict $p_t = y_{t-1}$". i.e. just always predict that what happened yesterday will happen today. In this case we have $\left|\frac{1}{T}\sum_{t=1}^{T} p_t - \sum_{t=1}^{T} y_T\right| = y_T/T \leq 1/T$.

But it is easy to differentiate the yesterday weatherman from the oracular weatherman. If the oracular weatherman predicted a 100% chance of rain, it would *always* rain on such days. But the yesterday weatherman frequently predicts a 100% chance of rain and is wrong. In other words, the yesterday weatherman violates *prediction conditioned average consistency.* We'll bucket the weatherman's predictions into 100 buckets (i.e. by percentage points), and we'll say that a prediction p_t is in bucket i ($p_t \in B(i)$) if it is closer to $i/100$ than any other point $j/100$.

Definition 2 *Given a sequence of predictions and outcomes $(p_1, y_1, \ldots, p_T, y_T)$, let $n_T(i) = |\{t : p_t \in B(i)\}|$ be the number of rounds on which the prediction was in bucket i. The sequence satisfies ϵ-prediction conditioned average consistency for a bucket i if:*

$$\left|\frac{\sum_{t:p_t \in B(i)} y_t - p_t}{n_T(i)}\right| \leq \epsilon$$

In other words, conditioned on making a prediction of a $\approx i/100$ probability of rain, the weather forecaster should have been right — i.e. on the days on which he predicted a $\approx i/100$ probability of rain, it should have rained roughly a $i/100$ fraction of the time.

Thus suggests a stronger test: *calibration.* The idea is to ask for prediction conditioned average consistency for *all* 100 buckets i. But if we think about this a bit harder we realize that the oracular weatherman might not be able to satisfy this. Suppose there is only a single day for which $p_t^* \in B(30)$, and that as luck would have it, on that day it actually rained? A single stroke of bad luck (that happens 30% of the time!) would ruin conditional average consistency for $i = 30$. However, we can ask that this condition hold on average over the buckets i, weighted by their frequency:

Definition 3 *A prediction strategy satisfies ϵ-average calibration if for all sequences of outcomes, the sequence of predictions it generates $(p_1, y_1, \ldots, p_T, y_T)$ satisfies:*

$$\mathrm{E}\left[\sum_{i=1}^{100} \frac{n_T(i)}{T} \cdot \left|\frac{\sum_{t:p_t \in B(i)} y_t - p_t}{n_T(i)}\right|\right] = \frac{1}{T}\mathrm{E}\left[\sum_{i=1}^{100}\left|\sum_{t=1}^{T} \mathbb{1}[p_t \in B(i)](y_t - p_t)\right|\right] \le \epsilon$$

We say it satisfies average calibration if $\epsilon \to 0$ as $T \to \infty$

It will be more convenient to instead work with a "Euclidean" metric of calibration error:

$$L_T = \sum_{i=1}^{100}\left(\sum_{t=1}^{T} \mathbb{1}[p_t \in B(i)](y_t - p_t)\right)^2$$

You can confirm (this is the "Cauchy-Schwartz inequality") that the average calibration loss ϵ of a strategy is upper bounded by:

$$\epsilon \le \mathrm{E}\left[\frac{10}{T}\sqrt{L_T}\right] \le \frac{10}{T}\sqrt{\mathrm{E}[L_T]}$$

It turns out there is an algorithm that will let any weatherman pass the calibration test as well, even without any knowledge of weather. To design the algorithm, lets suppose our weatherman has already made predictions up through day $s-1$, and is considering what he should predict on day s. If he predicts $p_s \in B(i)$ and the outcome turns out to be y_s, then the increase in the loss function will be:

$$\begin{aligned}
\Delta_s(p_s, y_s) &= L_s - L_{s-1} \\
&= \left(\sum_{t=1}^{s} \mathbb{1}[p_t \in B(i)](y_t - p_t)\right)^2 - \left(\sum_{t=1}^{s-1} \mathbb{1}[p_t \in B(i)](y_t - p_t)\right)^2 \\
&= \left(V_{s-1}^i + (y_s - p_s)\right)^2 - \left(V_{s-1}^i\right)^2 \\
&\le 2V_{s-1}^i \cdot (y_s - p_s) + 1
\end{aligned}$$

where $V_{s-1}^i = \sum_{t=1}^{s-1} \mathbb{1}[p_t \in B(i)](y_t - p_t)$ is a fixed constant at the time that the weatherman must make her decision on day s. Observe that $|V_{s-1}^i| \le T$.

Now suppose we could show that the weatherman had a distribution over predictions that would guarantee that $\mathrm{E}[\Delta_s(p_s, y_s)] \le 2T/m + 1$ at every round. Then we would have that:

$$\mathrm{E}[L_T] = \sum_{t=1}^{T} \mathrm{E}[\Delta_t(p_t, y_t)] \le \frac{2T^2}{m} + T = O\left(\frac{T^2}{m} + T\right)$$

and our calibration loss would be bounded by $\epsilon \le \frac{10}{T}\sqrt{\mathrm{E}[L_T]} = O(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{T}})$. Hence if we chose $m = T$, we would have calibration loss on the order of $O(1/\sqrt{T})$, and therefore a predictions strategy satisfying average calibration.

So that's the plan. To understand how our algorithm should make predictions at round s Define a zero-sum game that has cost function taking value:

$$C_s(p, y) = 2V_{s-1}^i \cdot (y_s - p_s) + 1$$

for each $p \in B(i)$. The minimization player (The Learner) has action set $A_1 = \{1/m, 2/m, \ldots, 1\}$, and the maximization player (The Adversary) has action set $A_2 = \{0, 1\}$. We have by construction that $\Delta_s(p_s, y_s) \leq C_s(p_s, y_s)$, so we need to bound the value of this game. This is easier to do if the adversary moves first, because its much easier to make a prediction about something you already know! This corresponds the max min value of the game, in which the adversary first commits to a distribution $q \in \Delta\{0, 1\}$.

Note that once the adversary commits to a distribution q, this fixes $\mathrm{E}_{y\sim q}[y]$, which the Learner knows. So in this ordering of moves, the Learner is actually in the role of the oracular weatherman! To best respond, the Learner should set $p = \mathrm{E}[y]$, which would guarantee that $\mathrm{E}[C_s(p, y)] = 1$. The learner cannot necessarily quite do this (because his action set only contains multiples of $1/m$), but he can always find a p such that $|p - \mathrm{E}_q[y]| \leq 1/m$. Hence we have:

$$\max_{q\in\Delta A_2} \min_{p\in A_1} \mathrm{E}_{y\sim q}[C_s(p, y)] \leq 2 \cdot \frac{|V^i_{s-1}|}{m} + 1 \leq 2 \cdot \frac{T}{m} + 1$$

Applying the minimax theorem, we can conclude that the value of the game remains the same if the Learner moves first:

$$\min_{\hat{p}\in\Delta A_1} \max_{y\in A_2} \mathrm{E}_{p\sim\hat{p}}[C_s(p, y)] \leq 2 \cdot \frac{T}{m} + 1$$

In other words, at every round s, the learner has a distribution over predictions $\hat{p}_s$ that guarantees that *no matter what the label y_s is*:

$$\mathrm{E}_{p_s\sim\hat{p}_s}[\Delta_s(p_s, y_s)] \leq \max_{y\in A_2} \mathrm{E}_{p\sim\hat{p}}[C_s(p, y)] \leq 2 \cdot \frac{T}{m} + 1$$

Which is exactly what we have wanted. In other words, we have proven the following theorem:

Theorem 4 *There exists a prediction strategy that against an arbitrary adversarially chosen sequence of T outcomes satisfies ϵ-average calibration for $\epsilon = O(1/\sqrt{T})$*

What is that strategy? It simply plays the minmax equilibrium strategy for the Learner in the zero-sum game we derived above! We can always efficiently compute the equilibrium of a zero-sum game by writing it as a linear program which explicitely finds the distribution over actions for the learner that minimizes the maximum cost resulting from any action of the adversary:

Algorithm 1 Algorithm for Predicting at Round s

```
Let $\hat{p} \in \Delta[m]$ be the solution to the following linear program defined over variables $\hat{p}_1, \ldots, \hat{p}_T$:

    Minimize $\gamma$ such that:

    $\sum_{t=1}^{T} \hat{p}_t = 1, \quad \sum_{t=1}^{T} \hat{p}_t C_s\left(\frac{t}{T}, 0\right) \leq \gamma, \quad \sum_{t=1}^{T} \hat{p}_t C_s\left(\frac{t}{T}, 1\right) \leq \gamma$

Select $p_s = \frac{t}{T}$ with probability $\hat{p}_t$.
```

There are only 3 constraints, and T variables in this linear program, so solving it takes time polynomial in T. In fact, in this particular case, the minmax equilibrium strategy for the learner has a nice closed form (that you may work out on the homework) that can be sampled from in time independent of T, with no need to solve a linear program.

A couple of remarks are in order:

1. Here the minimax theorem gave us an existential proof of the existence of an algorithm! We only needed to reason about the (easy) problem of predicting something about a distribution we already

know, because the adversary (who in this thought experiment is forced to announce his strategy) told us. The minimax theorem tells us we can do just as well in the actual case, in which we must commit to an algorithm first, without knowledge of the adversary's plans.

2. This argument was rather generic to any linear (i.e. based on bounding sums or averages) test aimed at distinguishing the oracular weatherman from a fraud. This is because the minimax theorem literally is allowing us to analyze the Learner as if she is the oracular weatherman!

3. We are able to mimic the oracular weatherman even if the truth is that outcomes are chosen adversarially, without any probabilistic model at all. This should make you think critically about how much we can learn from empirical tests of probabilistic models.