ORIE 6745: Causality and Learning for Intelligent Decision Making September 6, 2017

Lecture 5

Lecturer: Nathan Kallus *Scribe: Jessie Gao*

1 Recap

Last lecture, we talked about generalizing bounds for learning to decide fully-observed counterfactuals.

1.1 Generalization Bounds

This is my understanding of what we're doing here: bounding the $D(\Pi)$ we defined here can give us both: 1) Goal-1: The error term between estimated policy risk and true policy risk - so we can estimate the policy risk when the error term can shrink to 0 as number of data points increases, 2)Goal-2: $D(\Pi)$ also can bound the ERM (empirical risk minimization) so we can find the best-in-class policy, given the policy class we define, when the error can shrink to 0 as number of data points increased.

The **true policy risk** of a policy π is $R(\pi) = APO(\pi) = \mathbb{E}[Y(\pi(x))]$

If for each unit/example in out data $i = 1, 2, ..., n$, we observe every single counterfactual outcome $Y_i(z)\ \forall z \in Z$ (normally we don't know what would have happened to the unit under an alternative treatment, but only know the historically applied one of the potential treatments), then we can compute the **Empirical Policy Risk** $\hat{R}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\pi(x_i))$

Goal-1: We generalized bounds for $D(\pi)$, which is the error of using Empirical Policy Risk $\hat{R}_n(\pi)$ as the true policy risk$R(\pi)$ so that we can prove when it's scientific to use Empirical Policy Risk as estimator (when the error shrink to 0 as n increases to a large number).

Goal-2: The task is to find the best-in-class policy $\pi^\star$ in the **policy class** $\Pi \subseteq [\mathcal{X} \mapsto \mathcal{Z}]$: $\pi^\star(x) = \text{argmin}_{\pi\in\Pi}\, \mathbb{E}[Y(\pi(x))]$.

First, we focused on producing the bound in the form of this $D(\pi)$, which is defined as the worst case error (the biggest error for all policies in the policy class Π)

With probability $\geq 1-\delta$,
$D(\Pi) = \textbf{sup}_{\pi\in\Pi}\left|\hat{R}_n(\pi) - R(\pi)\right|$
$\leq const\sqrt{\frac{2\log(2/\delta)}{n}}$(**confidence term**) + (**complexity of** Π)

The *sup* here means the above is the same with the following:
$\forall\pi \in \Pi$, $R_n(\pi) \leq \hat{R}_n(\pi) + const\sqrt{\frac{2\log(2/\delta)}{n}}$ + (complexity of Π) (We only want to know upper-bound)

(Proof Recap: Radamacher, Lipschitz, McDiarmid's inequality)

$D(\pi)$ is upper-bounded by the confidence term plus the complexity of Π, that is, with high probability, $D(\pi)$ is upper-bounded in terms of the complexity of Π. Intuitively, the confidence term is the randomness caused by sampling, and the effect of the complexity of the policy class here can be understood by the fact that if

the policy class Π is too rich, it can cause the *sup* to be really big.

- A simple example of the interpretation of the complexity of Π: if the image of Π, Z is univariate and we have bounded and Lipschitz outcomes $Y_i(.)$, then the complexity of Π is a constant times the Rademacher complexity:

  Complexity of Π is:
  const $\mathscr{R}(\Pi) =$ const $\frac{1}{2^n}\sum_{\sigma\in\{-1,+1\}^n}\sup_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^n\sigma_i\pi(x_i) =$ const $\mathbb{E}_\sigma\sup_{\pi\in\Pi}\frac{1}{n}\sum_{i=1}^n\sigma_i\pi(x_i)$

Goal-1 achieved.

Later We hope to find the best-in-class policy $\pi^\star$ in the **policy class** $\Pi \subseteq [\mathcal{X}\mapsto\mathcal{Z}]$ by $\pi^\star(x) = \text{argmin}_{\pi\in\Pi}\mathbb{E}[Y(\pi(x))]$. However, we in fact only know the Empirical Policy Risk $\hat{R}_n(\pi_n)$ and cannot minimize $\mathbb{E}[Y(\pi(x))]$ directly.

So we can only do **Empirical Risk Minimization (ERM)**: $\hat{\pi_n} = \text{argmin}_{\pi\in\Pi}\hat{R}_n(\pi)$. And later we will see that since $D(\Pi)$ also gives bound to ERM, we can use $\hat{\pi_n}$ as our estimated best-in-class policy.

$$\begin{aligned}
R(\hat{\pi}_n) &\le \hat{R}_n(\hat{\pi}_n) + D(\Pi) && (D(\Pi) \text{ is the biggest absolute difference between } \hat{R}_n \text{ and } R_n \text{ by definition})\\
&\le \hat{R}_n(\pi^*) + D(\Pi) && (\hat{\pi}_n \text{ minimizes } \hat{R}_n)\\
&\le R(\pi^*) + 2D(\Pi) && (D(\Pi) \text{ is the biggest absolute difference between } \hat{R}_n \text{ and } R_n \text{ by definition})
\end{aligned}$$

Goal-2 achieved.

1.2 Two Important Examples

1. Continuous univariate decision: $Z = \mathbb{R}$, (e.g., regression or dosing for particular drug)

   For example, We look at $\Pi_M = \{X \mapsto \beta^T X : ||\beta||_2 \le M\}$, which has the linear decision rule with the linear part bounded by norm.

   In the last lecture we have proved that the difference between the true and the estimated is bounded by the confidence term plus the complexity of Π: with probability $\ge 1-\delta$, $\forall\pi\in\Pi$ $R_n(\pi) \le \hat{R_n}(\pi) + const\sqrt{\frac{2\log(2/\delta)}{n}} + const\frac{M}{\sqrt{n}}$. The term regarding the complexity of Π in this case is bounded by the size of the linear term and can go down to 0 with n going to infinity, so we are good.

   (Proof recap: The outcome $Y_i(z) = (L_i - z)^2$ is 2c-Lipschitz and bounded and $z\in\mathbb{R}$ is univariate. The complexity is interpreted using Radamacher Complexity.)

   Goal-1 achieved.

   The next step is to find the best-in-class policy.

   Note that by the definition of Π here, we would have $\Pi_M \subseteq \Pi_{M'}$, if $M \le M'$, and $\Pi_\infty = \cup_M \Pi_M$. The reasons are that Π_M is all the possible rules with the linear decision rule smaller than M' and is surely subset of $\Pi_{M'}$, and Π_∞ is all the possible rules (the biggest policy class including all M). As we increase M, the policy class will expand. We hope to find the smallest policy class for which we can

find the π to minimize the $\hat{R_n}(\Pi)$. So we will use some regularized approach like **SRM (Structural Risk Minimization)** to derive the best π, i.e., β.

$argmin_{\pi \in \Pi_\infty}\ min_{\pi \in \Pi_M}$ $(risk\ bound)$
The inside min means we find the min of best policy with the liner term bound being M. The outside $argmin$ is to find the smallest policy class trying all M.

$= argmin_\beta \hat{R}_n(\Pi(\beta)) + \lambda||\beta||_2^2$
$= argmin_\beta \hat{R}_n(x \mapsto \beta^T x) + \lambda||\beta||_2^2$, for some λ

Goal-2 achieved.

2. Discrete decision: $Z \in \{+1, -1\}$ (e.g, classification or give/don't give a treatment)

   We have proved from last lecture that:
   For binary decision policies and bounded outcomes (according to last class, this can also be extended to other cases),
   $R_n(\pi) \leq \hat{R}_n(\pi) + const\sqrt{\frac{2\log(2/\delta)}{n}} + const\frac{VC(\Pi)}{\sqrt{n}}$.

   The complexity of Π can be (upper) bounded by the VC Dimension term. VC dimension is the cardinality of the largest set of points that the algorithm (Π) can shatter. We have defined the concept of shattering by notation in the previous lecture. Here I found another interpretation that might help us understand the intuition of it:

   > *Definition:* a **dichotomy** of a set S is a partition of S into two disjoint subsets.
   >
   > *Definition:* a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

   It mean that:

   - $v \leq VC(\Pi) \Leftrightarrow \exists$ point set of size v that Π can shatter
   - $v > VC(\Pi) \Leftrightarrow \forall$ point sets of size v, Π cannot shatter it

   For example, Linear Classifiers on $X = \mathbb{R}^d$ have VC Dimension $d+1$, which means for $d = 2$, we have:

   - $\exists$ point set of size 3 that Π can shatter: For all arrangements of points as long as the three points are not on a straight line with the point in the middle being the only point of a class different from the other two points (we are able to find at least one instance that it can shatter so we are good).
   - $\forall$ point sets of size 4, Π cannot shatter it: for all possible arrangements of the 4 points, we cannot shatter

   (Proof recap: Massart's Lemma and Sauer's lemma)

2 Controlled Experiment

2.1 Motivation

In the previous lecture, we worked on situation where counterfactual results are fully-observed. However, in reality, we don't know counterfactuals but only know factual results-that is, we don't know $Y_i(z)\ \forall z \in Z$ but only know $Y_i = Y_i(z)$ for some single historically applied treatment z. For example, we only know what happened to a person who took aspirin but what would have happened if that person never took aspirin. That is the Fundamental Problem of Casual Inference.

To deal with this, we will create statistical homogeneity through randomization. Randomization means there is no alternative explanations for an observed association other than a casual effect.

2.2 Definition

Controlled experiment is also known as clinical/random trials, A/B testing.

A controlled experiment is a setting where the investigator has complete control over the assignment and application of treatments to the units.

- For example, we might want to study the effect of a job training program for helping disadvantaged workers on their wages. Given a cohort of of n such workers, a controlled experiment may assign some of these to receive the full benefits of the program and others to do this for themselves,and finally record each subject's wages in the year following the program.

Example of not a controlled experiment:

1. Subjects could choose for themselves whether to participate (two-way noncompliance like we asked them to show up but they didn't, or we didn't ask and they showed up; or one-way noncompliance like only people given the opportunities could choose reject). So we didn't have the treat side and control side being clear. We later will talk talk about instrumental variables to deal with this.
2. We just observe some people who did and didn't take a program and their wages. For instance, sometimes, we don't even have a chance for a controlled trial, for example, due to moral issue. Does smoke cause cancer? (We cannot either force people to smoke or not to smoke.) Does lead paint develop mental disability? Does gender = cis- women cause lower wages? (The treatment is not even a sex change operation but what would happen if the person was born with a different gender.)

When possible, controlled trials are the GOLD STANDARD for establishing causal relationship because the assumptions necessary are very minimal.

We will focus on two treatments $\mathcal{Z} = \{0, 1\}$. Recall our formal framework of potential outcomes. Random variables $Y_i(0), Y_i(1)$ corresponding to outcome one would observe under each assignment to treatment. Individual Treatment Effect: $TE_i = Y_i(1) - Y_i(0)$.

But we never observe both $Y_i(0), Y_i(1)$. We also never observe TE_i. The statistical solution is to focus on aggregate effect across many units. Sample average treatment effect $SATE = \frac{1}{n}\sum_i TE_i = \frac{1}{n}\sum_i (Y_i(1) - Y_i(0))$. If units are drawn from a population $PATE = E(TE) = E(Y(1) - Y(0)) = ESATE$. PATE is the generic effect over all units - expectation of a random person drawn from the population. ESATE is the expectation of the effect over different samples - expectation of mean of the n persons drawn from the population.

Controlled Experiment Setting: We choose (design) assignment $Z_{1:n} \in \{0,1\}^n$. After applying treatments, we observe $Y_{1:n} \in R^n$. The key assumptions for controlled experiment: $Y_i = Y_i(Z_i)$.

1. Non-interference: if $Z_{1:n}$, $Z_{1:n}\prime$ have $Z_i = Z_i'$, we observe the same outcome for i under either assignment.
2. Consistency: what we observe corresponds to the potential outcome of the treatment assigned.

One estimate of the effect is the mean difference estimator of effect: $\hat{Y} = \frac{\sum_{i:z_{i=1}} Y_i}{n_i} - \frac{\sum_{i:z_{i=0}} Y_i}{n_0}$, $n_z = \sum_{zi=z} 1$