Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Varun Kumar Khare
Instructor: Purushottam Kar
Date: January 20, 2018

LECTURE 6

Infinite Function Classes

1 Introduction

For the context of the lecture, $\mathcal{X}$ is our feature space and $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is our function class. In the previous lectures we proved the bounds of point wise convergence and uniform convergence for function classes with finite number of possible functions. However, things break down immediately when the the domain of $\mathcal{X}$ becomes uncountable ($[a, b]\, a, b \in \mathbb{R}$) or equivalently when the size of function class ($\mathcal{F}$) reaches ∞. This was seen in the case for real regression in the last lecture. We had already proved the point wise convergence for linear regression and had stated the uniform convergence bound. We will be giving the proof of uniform convergence bound for infinitely large function classes like linear regression in this lecture and will also be informally introducing the concept of covering numbers to calculate these bounds.

2 Uniform Convergence

Till now we didn't give the actual definition of Uniform convergence and were just using intuitive understanding to get bounds. In order to get the bounds for uniform convergence the function class must exhibit such convergence or else we cannot use point wise convergence bounds to evaluate uniform convergence. This was justified by the example given in Lecture 4. Note that the bound that comes up for finite function class size is still not wrong. It just gives a trivial result. Lets formally define uniform convergence for better understanding.

Definition 6.1. A function class $\mathcal{F} \in \mathbb{R}^{\mathcal{X}}$ exhibits uniform convergence w.r.t a particular loss function (l) if for all distributions $\mathcal{D}$ over ($\mathcal{X} \times \mathbb{R}$), we have,

1. $\lim\limits_{n\to\infty} \mathop{\mathbb{E}}\limits_{\mathcal{S}\sim\mathcal{D}^n, f\in\mathcal{F}} \left[sup\left|er^l_{\mathcal{D}}\left[f\right] - er^l_{\mathcal{S}}\left[f\right]\right|\right] = 0$ or equivalently,

2. $\forall \epsilon > 0 \lim\limits_{n\to\infty} \mathop{\mathbb{P}}\limits_{\mathcal{S}\sim\mathcal{D}^n, f\in\mathcal{F}} \left[\left|er^l_{\mathcal{D}}\left[f\right] - er^l_{\mathcal{S}}\left[f\right]\right| > \epsilon\right] = 0$

Definition 6.2. A function class $\mathcal{F}$ that exhibits uniform convergence is called Uniform Glivenko-Cantelli Class (UGC).

Exercise 6.1. Prove statement 1 and 2 of `Definition 6.1` are equivalent i.e. $1 \Leftrightarrow 2$.
Hint:

1. Assume $\forall \mathbf{x}, \forall y, \forall f \in \mathcal{F}\ \ |l(f(\mathbf{x}), y)| \leq B$ and Apply Markov's Inequality

2. Use the definition of convergence of a sequence.

$$\lim_{n\to\infty} \mathcal{S}_n = 0 \Rightarrow \forall \xi > 0\ \exists n_\xi > 0\ st\ \forall n_q > n_\xi\ we\ have\ \mathcal{S}_n < \xi\ where\ n_q, n_\xi \in \mathbb{N}$$

We also utilized the upper bound of a variance of a random variable $\mathcal{X} \in [a, b]$ in the previous lecture. Just to ensure everyone is on the same page we will now give the proof of that bound.

Lets first see a lemma,

Lemma 6.1. $\mathrm{Var}\left[\mathcal{X}\right] \leq \mathbb{E}\left[(x-\nu)^2\right]$

Proof. This lemma can be proved easily by using linearity of expectations. Assume $\mathbb{E}\left[\mathcal{X}\right] = \mu$, then,

$$\begin{aligned}\mathrm{Var}\left[\mathcal{X}\right] - \mathbb{E}\left[(x-\nu)^2\right] &= \mathbb{E}\left[(x-\mu)^2\right] - \mathbb{E}\left[(x-\nu)^2\right] = \mathbb{E}\left[(x-\mu)^2 - (x-\nu)^2\right] \\ &= \mathbb{E}\left[\left((x-\mu)-(x-\nu)\right)(2x-\mu-\nu)\right] = (\nu-\mu)\,\mathbb{E}\left[(2x-\mu-\nu)\right] \\ &= (\nu-\mu)\left(2\mathbb{E}\left[x\right]-\mu-\nu\right) = (\nu-\mu)(\mu-\nu) \leq 0\end{aligned}$$

You can equivalently prove

$$\mathbb{E}\left[(x-\nu)^2\right] = \mathrm{Var}\left[\mathcal{X}\right] + \left(\mathbb{E}\left[x-\nu\right]\right)^2$$

□

Claim 6.2. For a random variable $\mathcal{X} \in [a, b]$, $\mathrm{Var}\left[\mathcal{X}\right] \leq \frac{(b-a)^2}{4}$ where $a, b \in \mathbb{R}$

Proof. Take $\nu = \frac{(b+a)}{2}$ in `lemma 6.2` which on applying bounds of $\mathcal{X}$ gives,

$$\mathrm{Var}\left[\mathcal{X}\right] \leq \mathbb{E}\left[(x-\nu)^2\right] \leq \mathbb{E}\left[\left(b - \frac{b+a}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}$$

hence `Claim 6.2` is justified. □

3 Method of Covering Numbers

Covering numbers are an important aspect of VC-dimension, Rademacher complexity and uniform entropy. Informally, covering numbers are defined as the number of spherical balls of a given size needed to completely cover a given space, with possible overlaps. In the context of identifying optimal function $f^* \in \mathcal{F}$, each ball refers to a cluster of functions that behave similarly on EVERY dataset $\mathcal{S} \sim \mathcal{D}^n$. We define two functions to behave similarly if the output of both of the functions lie within the ϵ range of the other. i.e.

$$f_1 \sim f_2 \Leftrightarrow f_1(x) \in f_2(x) + [-\epsilon, \epsilon]\, \forall x \in \mathcal{X} \;\; for\ some\ \epsilon \geq 0$$

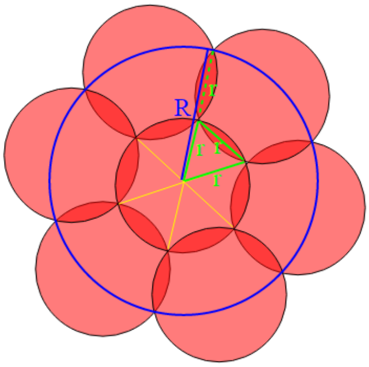

Figure 1: [1] showing covering numbers or overlapping clusters. Smaller balls completely fill out the bigger sphere

3.1 Covering numbers as Student Representatives

Assume we have a huge class of students and we have to decide an optimal question paper to objectively evaluate the true performance of every student. This task is similar to deciding an optimal dataset size (n) for $\mathcal{S} \sim \mathcal{D}^n$ to get empirical risk as the evaluator of the true performance of each function. Hence the task of getting uniform convergence is analogous to deciding optimal question paper for all the students. As the number of students (size of the function class $\mathcal{F}$) increases, the uniform convergence results become trivial. However, if we are able to group the students into clusters and apply the uniform convergence results on the set of clusters, we would be able to get a non-trivial bound. This is the intuition behind covering numbers. We identify a student representative for each cluster of similar students and ensure uniform convergence for the set of student representatives.

First, we focus on certain assumptions and the conclusions that can be drawn from these assumptions

1. Loss function should be L-lipchitz $\forall y \in \mathcal{Y}\ where\ \mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$.
   - L-lipchitz functions are defined as
   $$\left|l\left(\hat{y_1}, y\right) - l\left(\hat{y_2}, y\right)\right| \leq L * \left|\hat{y_1} - \hat{y_2}\right|$$
   This is required to maintain continuity.
   - It hence ensures that the models are not penalized too much if they give similar predictions $\hat{y}$.
   - W.r.t students, the marking system shouldn't be too erratic for example either 0 or 100 marks. In such a situation even though the students have similar knowledge their performance can vary a lot.
   - Eg: squared loss is a 4B-Lipchitz function
2. Assume two functions f_1, f_2 and analogously students S_1, S_2 are ϵ similar. i.e.
$$f_1, f_2 \in \mathcal{F}\ st\ \forall x \in \mathcal{X},\ \left|f_1\left(x\right) - f_2\left(x\right)\right| \leq \epsilon$$
$$\Rightarrow \forall (x, y) \in \mathcal{X} \times \mathcal{Y},\ \left|l_{sq}\left(f_1\left(x\right), y\right) - l_{sq}\left(f_2\left(x\right), y\right)\right| \leq 4B\epsilon$$

3. $\forall \mathcal{S} \in (\mathcal{X} \times \mathcal{Y})$,

$$\begin{aligned}
\left|er^l_{\mathcal{D}}[f_2] - er^l_{\mathcal{S}}[f_2]\right| &\leq \left|er^l_{\mathcal{D}}[f_2] - er^l_{\mathcal{D}}[f_1]\right| + \left|er^l_{\mathcal{D}}[f_1] - er^l_{\mathcal{S}}[f_2]\right| \\
\left|er^l_{\mathcal{D}}[f_2] - er^l_{\mathcal{D}}[f_1]\right| &\leq 4B\epsilon \\
\left|er^l_{\mathcal{D}}[f_1] - er^l_{\mathcal{S}}[f_2]\right| &\leq \left|er^l_{\mathcal{D}}[f_1] - er^l_{\mathcal{S}}[f_1]\right| + \left|er^l_{\mathcal{S}}[f_2] - er^l_{\mathcal{S}}[f_1]\right| \\
\left|er^l_{\mathcal{S}}[f_2] - er^l_{\mathcal{S}}[f_1]\right| &\leq 4B\epsilon \\
\Rightarrow \left|er^l_{\mathcal{D}}[f_2] - er^l_{\mathcal{S}}[f_2]\right| &\leq 8B\epsilon + \left|er^l_{\mathcal{D}}[f_1] - er^l_{\mathcal{S}}[f_1]\right|
\end{aligned}$$

$$\Rightarrow \left|er^l_{\mathcal{D}}[f_2] - er^l_{\mathcal{S}}[f_2]\right| - \left|er^l_{\mathcal{D}}[f_1] - er^l_{\mathcal{S}}[f_1]\right| \leq 8B\epsilon$$

This statement justifies that if two students are ϵ similar on every datapoint then the overall difference in their l-risk and empirical risk is $8B\epsilon$ similar. So if one of the two functions satisfies point-wise convergence condition with certain probability, the other function also satisfies point wise convergence (for different tolerance) simultaneously.

4. Assume the entire dataset or the set of all students (possibly infinite) can be partitioned into k finite overlapping clusters $(c_1, c_2 ... c_k)$ such that all points in a cluster are ϵ similar to atleast one common point. Let $\mathcal{C}$ be the set of these clusters. Then,

$$\mathcal{C} = \{c_i | i \in [k]\}$$

3.2 Uniform Convergence for the Clusters

Lets identify cluster representatives (student representatives $\hat{S}_i$) for each cluster $c_i \in \mathcal{C}$ as,

$$\hat{S}_i \in c_i \ st \ \hat{S}_i \sim S_i \forall S_i \in c_i$$

`Note:` Similarity here refers to ϵ similarity of functions defined above.
Now, by point wise convergence for each of these student representatives, we have,

$$\forall i \in [k]\,,\ \mathbb{P}\left[\left|er^l_{\mathcal{D}}\left[\hat{S}_i\right] - er^l_{\mathcal{S}}\left[\hat{S}_i\right]\right| > \xi\right] \leq \delta$$

where δ is given as (assuming n = size of dataset or $\mathcal{S} \sim \mathcal{D}^n, \mathcal{Y} \in [-B, B]$)

$$\delta = 2\exp\frac{-n\xi^2}{2B^4}$$

Taking union of all these events we get,

$$\mathbb{P}\left[\bigcup_{i\in[k]}\left\{\left|er^l_{\mathcal{D}}\left[\hat{S}_i\right] - er^l_{\mathcal{S}}\left[\hat{S}_i\right]\right| > \xi\right\}\right] \leq \sum_{i\in[k]}\mathbb{P}\left[\left|er^l_{\mathcal{D}}\left[\hat{S}_i\right] - er^l_{\mathcal{S}}\left[\hat{S}_i\right]\right| > \xi\right] \leq k\delta$$

Hence on taking compliment we get,

$$\mathbb{P}\left[\bigcap_{i\in[k]}\left\{\left|er^l_{\mathcal{D}}\left[\hat{S}_i\right] - er^l_{\mathcal{S}}\left[\hat{S}_i\right]\right| \leq \xi\right\}\right] \leq 1 - k\delta$$

We will see in the next lecture that the value of k for real bounded regression turns out to be $\left(1 + \frac{2B}{\xi}\right)^d$. We already proved that it is sufficient to prove convergence for one of the similar students/functions to get convergence for the other one. Hence, when all the student representatives satisfy point-wise convergence, every ϵ co-related student satisfies the convergence. So we get the bound for uniform convergence of real regression (an infinite function class) as

$$\mathbb{P}\left[\bigcap_{i\in[k]}\left\{\left|er^l_{\mathcal{D}}\left[\hat{S}_i\right]-er^l_{\mathcal{S}}\left[\hat{S}_i\right]\right|\leq\xi\right\}\right]\leq 1-\exp\left(\frac{-n\xi^2}{2B^2}+d\log\left(1+\frac{2B}{\xi}\right)\right)$$

References

Covering numbers. https://i.stack.imgur.com/TXTGM.png. stack-overflow.