Lecture 6: VC DIMENSION, FUNDAMENTAL THEOREM

Instructor: Aditya Bhaskara *Scribe: Nikhil Kala*

CS 5966/6966: Theory of Machine Learning

January 27th, 2022

Abstract

The lecture covers learnability for finite hypothesis classes. Also introduces learnability for infinite classes using Growth Functions.

1 Chernoff Bounds

Suppose we have a probability distribution with mean μ and support [a,b] which is the range of the values that the distribution can take. Then if we take n iid samples $(X_1, X_2, ..X_n)$, the sample average is very close to the mean. This theorem tell us how fast the sample mean converges to the mean of the distribution.
We have:

$$Pr[|\frac{(X_1 + ... + X_n) - \mu}{n}| > \epsilon] \leq 2\exp^{-\frac{\epsilon^2 n}{(a-b)^2}}$$

So we can see as we increase the samples the probability drops exponentially, i.e, we converge quickly. Hence, this is also called "large deviation bound".

2 Learnability for Finite Classes

If $\mathbb{H}$ is finite hypothesis class,i.e, $\mathbb{H} = \{h_1, h_2...h_M\}$
Claim:
For any X and distribution D over it, a sample of size $O(\frac{1}{\epsilon^2} log \frac{|H|}{\delta})$ is representative with prob. at least $1-\delta$
Proof:
First we will look at a single $h \in \mathbb{H}$
Now the probability that absolute difference in the sample error and risk of h is greater than e can be viewed as an application of Chernoff bound,i.e,
| sample error(h) - risk(h) | $> \epsilon$
So to get the iid samples,
Let,

$$X_i \begin{cases} 1 & if\ h\ is\ incorrect\ on\ sample\ i \\ 0 & if\ h\ is\ correct\ on\ sample\ i \end{cases}$$

So now we can say using Chernoff bound, the probability that

| sample error(h) - risk(h) | $> \epsilon$ is bounded by:

$$2e^{-\epsilon^2 m} \leq \frac{\delta}{|\mathbb{H}|}$$

Now we can use Union bound to prove that for m samples:
Pr[| sample error(h) - risk(h) | $> \epsilon$] $< \delta$
To prove this we try to define bad samples, "S" for h, if
| sample error on S (h) - risk(h) | $> \epsilon$
then it is a bad sample.

Using the above bound we know that the probability that a random sample S is bad for some given h is $\leq \frac{\delta}{|H|}$,(this is true for all $h \in \mathbb{H}$)

This implies that the Probability that S is good for all $h \in \mathbb{H}$ is $\geq 1 - \delta$.
As the union of bad probabilities is $\frac{\delta}{|H|} * m$, as $|H| = m$, this comes out to be δ

Thus the claim is proved.
Note: This essentially means that for finite hypothesis class, $O(log(|H|))$ samples will be representative.

3 Learnability for Infinite Classes

Most functions on continuous domain are infinite so we need to take into consideration infinite classes as well. Since the classes are infinite then we would need infinite samples to learn, as using the above claim will be not defined.
To overcome that we can divide the hypothesis into finitely many classes.
The logic behind this is that, if the hypothesis in a class provide the same result then we can just take the class as just one hypothesis.
Example: Threshold functions on a line
Since the sign patterns are fixed for m points as they will be (+ + + +.., - + + +.., ..., ...- - - -) m+1 types of patterns we can reduce the infinite set of threshold functions to m+1 classes as the hypothesis in the classes will classify the points in the same pattern.

4 Growth Functions

From the learnability of infinite classes we reach the conclusion that, What matters isn't the number of distinct hypothesis, The number of ways in which the hypothesis classify the points of the domain matters. So to leverage that we have the growth function.
Growth function of a hypothesis class $\mathbb{H}$ over Domain X for input size m,

$$\tau_{\mathbb{H}}(m) = \underbrace{max}_{\substack{|S|=m \\ S \in X}} \{ \text{ Number of distinct ways in which hypothesis in } \mathbb{H} \text{ classify S.} \}$$

Note: Two classification are distinct if they differ even on one example, + + + is different from + + -
We can observe that for binary classification the growth function is $\leq 2^m$ as those are all the possible permutations.

5 Growth Function Examples

1. Growth function for Linear Threshold Functions(LTFs):

$$\tau_{\mathbb{H}}(m) = \underbrace{max}_{\substack{|S|=m \\ S \in X}} \{ \text{ Number of distinct ways in which hypothesis in } \mathbb{H} \text{ classify S.} \}$$

The distinct ways are:

$1 -> + + + + ... +$

$2 -> - + + + ... +$

$3 -> - - + + ... +$

$4 -> - - - + ... +$

.

.

.

$m+1 -> - - - - - ... -$

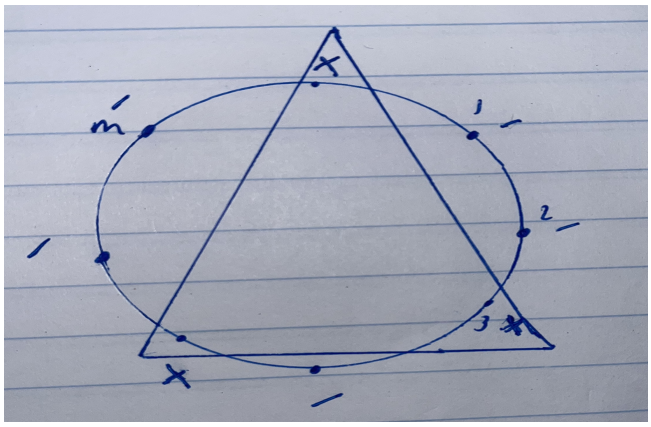
$$\tau_{\mathbb{H}}(m) = m + 1$$

2. Growth function for intervals on number line:
We can take interval as between any two spaces, so out of m+1 spaces we select 2, so $\binom{m+1}{2}$ which is $O(m^2)$

$$\tau_{\mathbb{H}}(m) = O(m^2)$$

3. Growth function for hypothesis class of convex polygons in real space:
The points inside the polygon are positive and outside are negative then to get all possible sign patterns, we can align the points equidistant on circumference of circle. Now we can take any points to construct a polygon, so we can get 2^m possible combinations.(As seen in fig. for 3 points inside the polygon)

$$\tau_{\mathbb{H}}(m) = O(2^m)$$



6 Learnability in terms of Growth Function

Theorem: Suppose $\tau_{\mathbb{H}}(m)$ is an upper bound on the total number of distinct "classifications" (or "sign patterns") possible for any sample of size m. Then for any X, D, if we take a sample S of size m, we have, with prob. 1-δ

$$\underbrace{sup}_{h \in H} |err(h, S) - err(h, D)| \leq \frac{4 + \sqrt{log \tau_{\mathbb{H}}(2m)}}{\delta \sqrt{2m}}$$

So to make this ϵ-representative , so

$$\frac{4 + \sqrt{log \tau_{\mathbb{H}}(2m)}}{\delta \sqrt{2m}} = \epsilon$$

Evaluating this for $\tau_{\mathbb{H}}(m) = m + 1$
We get,

$$m = \frac{4 log \frac{1}{\epsilon \delta}^2}{\epsilon^2 \delta^2}$$