Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Anupama Rajoriya
Instructor: Purushottam Kar
Date: January 17, 2018

LECTURE
5

The Hoeffding Inequality

1 Introduction

In previous lecture we derived the pointwise and uniform convergence for Bernoulli outcomes using Chernoff's inequality. In this lecture we will be looking at an example of a function class for which pointwise convergence hold but uniform convergence does not hold. Then we will move to real regression case and will introduce Hoeffding's inequality. The inequality enable us to derive the bounds for real regression problem.

2 Binary Classification

For binary classification problem with sample space $\mathcal{S} \in \mathcal{D}^n$ and finite hypothesis class $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$,

$$f_i : \mathcal{X} \to \mathcal{Y} = \{-1, 1\}$$

pointwise as well as uniform convergence hold, i.e.,$\forall f \in \mathcal{F}$

$$\mathbb{P}\left[\left|er_{\mathcal{D}}^{0-1}[f] - er_{\mathcal{S}}^{0-1}[f]\right| > \epsilon\right] \leq 2\exp\left(\frac{-n\epsilon^2}{3er_{\mathcal{D}}^{0-1}[f]}\right) \text{ if } |\mathcal{S}| = n, \;\; \mathcal{S} \sim \mathcal{D}^n \tag{1}$$

Or loosely speaking,

$$\mathbb{P}\left[\left|er_{\mathcal{D}}^{0-1}[f] - er_{\mathcal{S}}^{0-1}[f]\right| > \epsilon\right] \leq 2\exp\left(\frac{-n\epsilon^2}{3}\right) \text{ if } |\mathcal{S}| = n, \;\; \mathcal{S} \sim \mathcal{D}^n \tag{2}$$

The bound showing pointwise convergence directly comes from the Chernoff's inequality. The bound depends only on the type of loss function being used and does not depend on the type and size of output space $\mathcal{F}$ (the term $er_{\mathcal{S}}^{0-1}[f]$ is deterministic and does not depend on f). On the other hand, the bound showing uniform convergence says,

$$\mathbb{P}\left[\exists f \in \mathcal{F}, \left|er_{\mathcal{D}}^{0-1}[f] - er_{\mathcal{S}}^{0-1}[f]\right| > \epsilon\right] \leq 2\exp\left(\frac{-n\epsilon^2}{3p}\right)m \leq 2\exp\left(\frac{-n\epsilon^2}{3}\right)m \tag{3}$$

where $p = \max_{f \in \mathcal{F}} er_{\mathcal{S}}^{0-1}[f] \in [0, 1]$ and $m = |\mathcal{F}|$. The probability that there exist a function $f \in \mathcal{F}$ such that the deviation $\left|er_{D}^{0-1} - er_{S}^{0-1}\right| > \epsilon$ is proportional to the size of function space. Also, as the size grows calculating p becomes NP-hard for $0-1$ loss function.

Exercise 5.1. For the binary classification problem above with sample space $\mathcal{S} \in \mathcal{D}^n$ and finite hypothesis class $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$,

$$f_i : \mathcal{X} \to \mathcal{Y} = \{-1, 1\}$$

show that uniform convergence (given by (3)) implies pointwise convergence (given by (2)) but the converse is not true.

Now, let's take another example and study its convergence properties. Let the feature space $\mathcal{X} = \{0,1\}^d$ set of d-pixel images is mapped to a hypothesis space $\mathcal{F} = \{-1,1\}^{\mathcal{X}} = \{f : \mathcal{X} \to \{-1,1\}\}$ in a realizable setting such that $y = f^*(X)$ where $f^* \in \mathcal{F}$ and $X \sim \mathcal{D}(\mathcal{X})$. One can readily observe that for 0-1 loss function, pointwise convergence holds for this case as well because pointwise convergence does not require any other condition. Now the question arises, whether uniform convergence hold for this case or not.

In particular, uniform convergence does not hold for this function space. This can be explained as follows. Let n training samples are taken from the sample space of 2^d points. For this noiseless setting, we can get zero training error, $er_{\mathcal{S}}^{0-1}[\tilde{f}_s] = 0$ for some function $\tilde{f}(\cdot)$ but it will surely perform bad over the entire sample space. The overall test error will be, $er_{\mathcal{D}}^{0-1}[\tilde{f}] = \frac{\text{misclassified points}}{\text{size of sample space}} = \frac{2^d - n}{2^d}$ or the accuracy is $\frac{n}{2^d}$ which is very poor and can be achieved using any naive algorithm. Also the result obtained for example above given by equation (3) gives trivial result because of infinitely large function space($m = 2^{2^d}$). Thus the probability of deviation taking large value is not assured to take small values for all $f \in \mathcal{F}$.

Another common problem with infinitely large function spaces is in the calculation of $p = \max_{f \in \mathcal{F}} er_{\mathcal{D}}^{0-1}[f]$ which becomes NP-hard as dimension $|\mathcal{F}|$ grows. In the subsequent section we are going to discuss real regression problem where the output space is real valued instead of Boolean.

3 Real Regression

Let the sample space be $\mathcal{X} \subseteq \mathbb{R}^d$, output space and loss function space be $\mathcal{Y} = \mathbb{R}$, $\hat{\mathcal{Y}} = \mathbb{R}$ respectively. The hypothesis space is given by $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. Squared loss function,

$$\ell_{\text{sq}}(\hat{y}, y) = (y - \hat{y})^2 . \tag{4}$$

For this case we cannot use Chernoff's inequality as it demands for the outcome to be Boolean. So, here we introduce another inequality which can be seen as the extension of Chernoff's inequality for non-Bernoulli cases.

3.1 Hoeffding's Inequality

Theorem 5.1. (Hoeffding's Inequality) Let $X_1, X_2, \ldots, X_n$ be i.i.d. copies of random variable X s.t., $\mathbb{E}X = \mu$ and $X \in [a,b]$ a.s., $a, b \in \mathbb{R}$. then for $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and for every $t > 0$,

$$\mathbb{P}\left[\left|\bar{X}_n - \mu\right| > \epsilon\right] \le \exp\left(\frac{-n\epsilon^2}{2(b-a)^2}\right). \tag{5}$$

One can observe that this is more general inequality and the Chernoff's inequality is just a special case of Hoeffding's inequality. However, note that the variants of the Chernoff bound which are in terms of KL-divergence can not be derived from the Hoeffding's inequality.

Proof. We will just be giving the proof for a loose right tail bound using Cramèr-Chernoff algorithm and the left tail bound can be proved in a similar manner. We will proceed the same way as we did for Chernoff's bound proof.

$$\begin{aligned}\mathbb{P}\left[\bar{X}_n - \mu > \epsilon\right] &= \mathbb{P}\left[s\left(\bar{X}_n - \mu\right) > s\epsilon\right] \qquad\qquad \text{for some } s \geq 0\\ &= \mathbb{P}\left[\exp\left(s\left(\bar{X}_n - \mu\right)\right) > \exp\left(s\epsilon\right)\right]\end{aligned}$$

Using Markov's inequality for the random variable $\exp\left(s\left(\bar{X}_n - \mu\right)\right)$ we get,

$$\mathbb{P}\left[\bar{X}_n - \mu > \epsilon\right] \leq \frac{\mathbb{E}\left[\exp\left(s\left(\bar{X}_n - \mu\right)\right)\right]}{\exp\left(s\epsilon\right)}. \tag{6}$$

Let us define new random variables $\bar{Y}_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{n(b-a)}$ and $Y_i = \frac{X_i - \mu}{b-a}$. The new random variable has following properties, $\bar{Y}_n \in [-1, 1]$ a.s. and $\mathbb{E}\left[\bar{Y}_n\right] = 0$. The bound for moment generating function of random variable $\bar{Y}_n$ will be derived which will in turn bound $\mathbb{P}\left[\bar{X}_n - \mu > \epsilon\right]$.

$$\begin{aligned}\mathbb{E}\left[\exp\left(t\bar{Y}_n\right)\right] &= \left(\mathbb{E}\left[\exp\left(\frac{tY_i}{n}\right)\right]\right)^n = \left(\mathbb{E}\left[\exp\left(\frac{tY}{n}\right)\right]\right)^n\\ \mathbb{E}\left[\exp\left(t'Y\right)\right] &= \mathbb{E}\left[1 + t'Y + \frac{t'^2Y^2}{2!} + \frac{t'^3Y^3}{3!} + \cdots\right]\\ &= \mathbb{E}\left[1 + t'Y + \frac{t'^2Y^2}{2!}\left(1 + \frac{t'Y2!}{3!} + \frac{t'^2Y^2 2!}{4!} + \cdots\right)\right]\end{aligned}$$

where $t' = \frac{t}{n}$. As $Y \in [-1, 1]$ we have $|Y^n| \leq 1$ for all $n \in \mathbb{Z}_+$. Using this,

$$\begin{aligned}\exp\left(t'Y\right) &\leq 1 + t'Y + \frac{t'^2Y^2}{2!}\left(1 + \frac{t'2!}{3!} + \frac{t'^2 2!}{4!} + \cdots\right)\\ &\overset{(a)}{\leq} 1 + t'Y + \frac{t'^2Y^2}{2!}\exp\left(t'\right)\end{aligned}$$

where the inequality (a) comes using the property of factorials, $k! = (k-2)!2!$. Taking expectation to get moment generating function of Y we get,

$$\mathbb{E}\left[e^{t'Y}\right] \leq 1 + \frac{t'^2\text{Var}\left[Y\right]}{2}\exp\left(t'\right)$$

Here we will use a basic property of a bounded random variable given below.

Fact- For any random variable $Z \in [a, b]$, $\text{Var}\left[Z\right] \leq \frac{(b-a)^2}{4}$.

Using above fact we can lower bound MGF of Y as,

$$\begin{aligned}\mathbb{E}\left[e^{t'Y}\right] &\leq 1 + \frac{t'^2\text{Var}\left[Y\right]}{2}\exp\left(t'\right)\\ &\leq 1 + \frac{t'^2}{2}\exp\left(t'\right)\\ &\overset{(b)}{\leq} \exp\left(2t'^2\right) \qquad (7)\\ \mathbb{E}\left[e^{t\bar{Y}_n}\right] &= \left(\mathbb{E}\left[\exp\left(\frac{tY}{n}\right)\right]\right)^n \leq \left(\exp\left(\frac{2t^2}{n^2}\right)\right)^n \qquad (8)\end{aligned}$$

where inequality (b) came from the property $1 + X^2 \exp(X) \le \exp(2X^2)$. Replacing $\bar{Y}_n$ by $\bar{Y}_n = \frac{(\bar{X}_n - \mu)}{(b-a)}$ in equation (7),

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{t\left(\bar{X}_n - \mu\right)}{(b-a)}\right)\right] &\le \left(\exp\left(\frac{2t^2}{n^2}\right)\right)^n \\ \mathbb{E}\left[\exp\left(s\left(\bar{X}_n - \mu\right)\right)\right] &\le \left(\exp\left(2s^2(b-a)^2\right)\right)^n\end{aligned}$$

In last step we substitute t by $s = \frac{t}{n(b-a)}$.

Lemma 5.2. (Hoeffding's Lemma) For a random variable $X \in [a, b]$ a.s., then

$$\mathbb{E}\left[e^{sX}\right] \le \exp\left(\frac{s^2(b-a)^2}{8}\right). \tag{9}$$

Using this bound over $\mathbb{E}\left[\exp\left(s\left(\bar{X}_n - \mu\right)\right)\right]$ in equation (6) we obtain,

$$\begin{aligned}\mathbb{P}\left[\bar{X}_n - \mu > \epsilon\right] &\le \left(\frac{\mathbb{E}\left[\exp\left(s\left(\bar{X}_n - \mu\right)\right)\right]}{\exp(s\epsilon)}\right)^n \\ &\le \left(\exp\left(2s^2(b-a)^2 - s\epsilon\right)\right)^n.\end{aligned}$$

Now maximizing R.H.S. w.r.t. s we get the right tail bound as,

$$\mathbb{P}\left[\bar{X}_n - \mu > \epsilon\right] \le \exp\left(\frac{-n\epsilon^2}{8(b-a)^2}\right). \tag{10}$$

Similarly, the left tail bound can also be derived. Note that the proof given here results in a looser bound than (5). □

Exercise 5.2. To prove the Hoeffding inequality we used the fact that for any random variable $Z \in [a, b]$, $\mathrm{Var}[Z] \le \frac{(b-a)^2}{4}$ holds. Find a random variable for which the equality $\mathrm{Var}[Z] = \frac{(b-a)^2}{4}$ is achieved.

To apply bound to the real regression problem, the loss function and output need to be bounded. The loss function can be chosen such that it is bounded but the labels of the data are not in our hand. To get the bounded outcomes let us put some constraints over the sample space and define it as follows:

$$\begin{gathered}\mathbf{X} \subseteq \mathcal{B}(0,1) = \left\{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 \le 1\right\} \\ \mathcal{F} : \{f : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle, \|\mathbf{w}\|_2 \le B\}\end{gathered}$$

Now the function output is bounded by the scaler B, $|f(\mathbf{x})| = |\hat{y}| = |\langle \mathbf{w}, \mathbf{x} \rangle| \le \|\mathbf{w}\|_2 \|\mathbf{x}\|_2 \le B$. Also the squared loss function $\ell_{sq}(\hat{y}, y)$ is also bounded: $0 \le \ell_{sq}(\hat{y}, y) \le 4B^2$.

3.2 Pointwise Convergence for Real Regression

Hoeffding's inequality can be readily used to infer that the pointwise convergence hold here, i.e. $\forall f \in \mathcal{F}$,

$$\mathbb{P}\left[\left(\left|er^{sq}_{\mathcal{S}}[f] - er^{sq}_{\mathcal{D}}[f]\right|\right) > \epsilon\right] \le 2\exp\left(\frac{-n\epsilon^2}{32B^4}\right). \tag{11}$$

3.3 Uniform Convergence for Real Regression

Although we have seen an example of infinite function space that does not hold uniform convergence, fortunately we know that for this case uniform convergence hold. to establish this we are going to show following two useful results in subsequent lectures.

1. Uniform convergence depend on the dimensionality of the feature space d as follows:

$$\mathbb{P}\left[\left(\exists f \in \mathcal{F}, \left|er_{\mathcal{S}}^{sq}[f] - er_{\mathcal{D}}^{sq}[f]\right|\right) > \epsilon\right] \leq 2\exp\left(\frac{-n\epsilon^2}{32B^4} + 4d\log\frac{B}{\epsilon}\right) \tag{12}$$

   For above inequality to be true as n grows large such that $n \geq \left(\frac{64B^4}{\epsilon^2}d\log\frac{B}{\epsilon}\right)$ uniform convergence hold and the convergence does not depend on the dimensionality,

$$\mathbb{P}\left[\left(\exists f \in \mathcal{F}, \left|er_{\mathcal{S}}^{sq}[f] - er_{\mathcal{D}}^{sq}[f]\right|\right) > \epsilon\right] \leq 2\exp\left(\frac{-n\epsilon^2}{32B^4}\right). \tag{13}$$

   This is same as saying that the number of data points required in real regression should be atleast of order same as dimensionality of the feature space to get tolerable error.

2. Radamacher Analysis:

   We can get rid of the dependency over dimensionality and just need to retain the bound,

$$\mathbb{P}\left[\left(\exists f \in \mathcal{F}, \left|er_{\mathcal{S}}^{sq}[f] - er_{\mathcal{D}}^{sq}[f]\right|\right) > \epsilon\right] \leq 2\exp\left(\frac{-n\epsilon^2}{32B^4} + B\right). \tag{14}$$

   The probability of error is proportional to B which says that the weights need to take small values. It do not care about the values these weights are taking but require $\|\mathbf{w}\| \leq B$, which is still achievable.