CSC 588: Machine learning theory **Spring 2021**

Lecture 4: Hoeffding's Inequality, Bernstein's Inequality

Lecturer: Chicheng Zhang *Scribe: Brian Toner*

1 Hoeffding's Inequality and its supporting lemmas

Theorem 1 (Hoeffding's Inequality). *Suppose that $Z_1, ..., Z_n$ are iid such that for each i, $Z_i \in [a, b], \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \mu = \mathbb{E}[Z_i]$. Then for all $\epsilon > 0$,*

$$\mathbb{P}(|\bar{Z} - \mu| > \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

n The converse is almost true (up to a constant scaling of σ).

2 Proof of Lemma 3

Lemma 3: If X is σ^2-SG, then $\forall t > 0$,

$$\mathbb{P}(|X - \mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Proof.

$$X : \forall\lambda, \quad \mathbb{E}\left[\exp\left(\lambda(x-\mu)\right)\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right)$$

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(X - \mu \leq -t) + \mathbb{P}(X - \mu \geq t)$$

The above equality is true because they are mutually exclusive events.

$$\begin{aligned}
\mathbb{P}(X - \mu \geq t) &= \mathbb{P}\left(\exp\left(\lambda(x-\mu)\right) \geq \exp(\lambda t)\right) \quad \forall\lambda > 0 \\
&\leq \frac{\mathbb{E}[\exp(\lambda(X-\mu))]}{\exp(\lambda t)} \\
&\leq \exp(-\lambda t)\exp\left(\frac{\sigma^2\lambda^2}{2}\right) \quad \text{By Markov's inequality} \\
&= \exp\left(-\lambda t + \frac{\sigma^2\lambda^2}{2}\right)
\end{aligned}$$

Now choose $\lambda > 0$ to minimize the bound

$$\sigma^2\lambda - t = 0 \Rightarrow \lambda = \frac{t}{\sigma^2}$$

$$\Rightarrow \exp\left(-\lambda t + \frac{\sigma^2\lambda^2}{2}\right) = \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

$$P(|X - \mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

□

3 Proof of Lemma 2

Lemma 2: If $X_1, ..., X_n$ are independent and for all i, X_i is σ_i^2-SG, then $\sum_{i=1}^n a_i X_i$ is $\sum_{i=1}^n a_i^2 \sigma_i^2$-SG $\forall a_1, ..., a_n$.

1. Show aX_1 is $a^2\sigma_i^2$-SG. Let $\mathbb{E}[X_1] = \mu_1, \mathbb{E}[aX_1] = a\mu_1$.

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda(aX_1 - a\mu_1))\right] &= \mathbb{E}\left[\exp(\lambda a(X_i - \mu_i))\right] \\ &\leq \exp\left(\frac{(\lambda a)^2\sigma_1^2}{2}\right) \\ &= \exp\left(\frac{\lambda^2(a^2\sigma_1^2)}{2}\right)\end{aligned}$$

   $\Rightarrow aX_1$ is $a^2\sigma_1^2$-SG

2. Show that is X_1, X_2 are independent, $X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$-SG. Let $\mathbb{E}[X_1] = \mu_1, \mathbb{E}[X_2] = \mu_2$

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda(X_1 + X_2 - \mu_1 - \mu_2))\right] &= \mathbb{E}\left[\exp(\lambda(X_1 - \mu_1))\exp(\lambda(X_2 - \mu_2))\right] \\ &= \mathbb{E}\left[\exp(\lambda(X_1 - \mu_1))\right]\mathbb{E}\left[\exp(\lambda(X_2 - \mu_2))\right] \text{ By Independence} \\ &\leq \exp\left(\frac{\lambda^2\sigma_1^2}{2}\right)\exp\left(\frac{\lambda^2\sigma_2^2}{2}\right) \\ &= \exp\left(\frac{\lambda^2(\sigma_1^2 + \sigma_2^2)}{2}\right)\end{aligned}$$

   So $X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$-SG

3. $\sum_{i=1}^n a_i X_i$ is $\sum_{i=1}^n a_i^2\sigma_i^2$-SG by 1. and 2.

4 Proof of Lemma 1

Lemma 1: If X takes valeus in $[a, b]$, then X is $\frac{(b-a)^2}{4}$ -sub gaussian (SG)

We want to show:

$$\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp\left(\frac{(b-a)^2\lambda^2}{8}\right)$$

For X supported on $[a, b]$. Let $\psi(\lambda) = \ln\left(\mathbb{E}[\exp(\lambda(X - \mu))]\right)$. It suffices to show that $\psi(\lambda) \leq \frac{(b-a)^2\lambda^2}{8}$. Note that $\psi(\lambda)$ is called the cumulant generating function of $X - \mu$. Let $0 \leq \xi \leq \lambda$ and begin by Taylor expanding ψ.

$$\psi(\lambda) = \psi(0) + \psi'(0)\lambda + \frac{\psi''(\xi)}{2}\lambda^2$$

$$\psi(0) = 0$$

Let $Y = (X - \mu)$

$$\psi'(\lambda) = \frac{\mathbb{E}\left[\frac{\partial}{\partial\lambda}e^{\lambda Y}\right]}{\mathbb{E}[e^{\lambda Y}]} = \frac{\mathbb{E}\left[Ye^{\lambda Y}\right]}{\mathbb{E}[e^{\lambda Y}]}$$

So $\psi'(0) = \mathbb{E}[Y] = 0$.

$$\begin{aligned}
\psi''(\lambda) &= \frac{\mathbb{E}\left[Y^2 e^{\lambda Y}\right]}{\mathbb{E}[e^{\lambda Y}]} - \left(\frac{\mathbb{E}\left[Y e^{\lambda Y}\right]}{\mathbb{E}[e^{\lambda Y}]}\right)^2 \\
&= \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \\
&= Var(Z) \\
&= \mathbb{E}[Z - \mathbb{E}[Z]]^2 \\
&\leq \mathbb{E}\left[Z - \left(\frac{a+b}{2} - \mu\right)\right]^2 \\
&\leq \left(\frac{a-b}{2}\right)^2 = \frac{(b-a)^2}{4}
\end{aligned}$$

For random variable Z with density:

$$\mathbb{P}_Z(y) = \frac{\mathbb{P}_Y(y) e^{\lambda y}}{\int_{\mathbb{R}} \mathbb{P}_Y(y) e^{\lambda y} dy}$$

5 Proof of Hoeffding's Inequality

Hoeffding's Inequality: Suppose that $Z_1, ..., Z_n$ are iid such that for each i, $Z_i \in [a, b], \bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i, \mu = \mathbb{E}[Z_i]$. Then for all $\epsilon > 0$,

$$\mathbb{P}(|\bar{Z} - \mu| > \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

X_i is $\frac{(b-a)^2}{4}$-SG. Therefore, $\frac{1}{n}\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \left(\frac{1}{n}\right)^2 \frac{(b-a)^2}{4} = \frac{(b-a)^2}{4n}$-SG. By Lemma 3, $\forall \epsilon$:

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \epsilon\right) &\leq 2\exp\left(\frac{\epsilon^2}{2\frac{(b-a)^2}{4n}}\right) \\
&= 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)
\end{aligned}$$

6 Bernstein's Inequality

Theorem 2. *Let $X_1, ..., X_n$ be iid Random variables, and $\forall i, |X_i - \mathbb{E}X_i| \leq R$. Let $\sigma^2 = \mathrm{Var}(X_i)$. Then $\forall \epsilon \geq 0$*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}R\epsilon}\right).$$

Note: in some cases, $\sigma^2 \ll (b-a)^2$ which would imply $\frac{1}{\sigma^2} \gg \frac{1}{(b-a)^2}$. Let us set a small value of ϵ so that

$$
\begin{aligned}
&2\exp\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}R\epsilon}\right) \leq \delta \\
&\Leftarrow n\epsilon^2 \geq (2\sigma^2 + \frac{1}{3}R\epsilon)\ln\left(\frac{2}{\delta}\right) \\
&\Leftarrow n\epsilon^2 \geq 4\sigma^2 \ln\left(\frac{2}{\delta}\right) \text{ and } n\epsilon \geq \frac{2}{3}R\epsilon \ln\left(\frac{2}{\delta}\right) \\
&\Leftarrow \epsilon \geq \sqrt{\frac{4\sigma^2 \ln\frac{2}{\delta}}{n}} \text{ and } \epsilon \geq \frac{4R\ln\frac{2}{\delta}}{3n}
\end{aligned}
$$

Choosing

$$\epsilon = \sqrt{\frac{4\sigma^2 \ln\frac{2}{\delta}}{n}} + \frac{4R\ln\frac{2}{\delta}}{3n}$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq \delta$$

This implies the following corollary:

Corollary 3. *Let $X_1, ..., X_n$ be iid Random variables, and $\forall i, |X_i - \mathbb{E}X_i| \leq R$. Let $\sigma^2 = Var(X_i)$. Then with probability $1-\delta$:*

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \leq \sqrt{\frac{4\sigma^2 \ln\frac{2}{\delta}}{n}} + \frac{4R\ln\frac{2}{\delta}}{3n}$$

7 Exercise

Let $\mathcal{D}$ be a distribution over (X, Y) where $X \sim$unif$([0,1])$ and

$$Y \mid (X = x) = \begin{cases} -1 & x \in [0, 0.5] \\ +1 & x \in [0.5, 1] \end{cases}$$

deterministically.

Algorithm: Memorization: Given $\mathcal{S}$, returns $\hat{h}$ such that

$$\hat{h}(x) = \begin{cases} y_i & x = x_i \text{ for some } i \\ +1 & \text{otherwise} \end{cases}$$

1. $\text{err}(\hat{h}, \mathcal{S}) = 0$
2. $\text{err}(\hat{h}, D) = \frac{1}{2}$
3. Is it correct that with probability $1-\delta$,

$$\left|\text{err}(\hat{h}, \mathcal{S}) - \text{err}(\hat{h}, \mathcal{D})\right| \leq \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}?$$

No. Hoeffding's does not apply because we have:

$$\mathrm{err}(\hat{h}, \mathcal{S}) = \frac{1}{m} \sum_{i=1}^{m} I(\hat{h}(x_i) \neq y_i)$$

but $I(\hat{h}(x_i) \neq y_i) \not\sim \text{Bernoulli}(\mathrm{err}(\hat{h}, \mathcal{D}))$.

Chicheng notes: Here $\hat{h}$ is selected *after* seeing the (x_i, y_i)'s, which can also make $I(\hat{h}(x_i) \neq y_i)$'s dependent. Note that if instead $\hat{h}$ is chosen before seeing the (x_i, y_i)'s (which makes $\hat{h}$ independent of the (x_i, y_i)'s), then conditioned on $\hat{h}$, $\sum_{i=1}^{m} I(\hat{h}(x_i) \neq y_i)$ does come from $\text{Binomial}(m, \mathrm{err}(\hat{h}, D))$.