14 Mar 2022 — Chernoff Bound Applications

Announcements:

1. Next quiz is a week from Weds., i.e. March 23.
2. Practice problems to be distributed by Thurs.
3. Topics: Gaussians and Chernoff bounds.
4. Lectures from last week have been typeset; see "Lectures" page on course website.

Applications of Chernoff Bounds

Suppose we can draw i.i.d. samples from $[0, M]$
call them $Y_1, Y_2, \ldots$
and we want an estimate $\hat{Y}$ s.t.

$$\Pr\left(|\hat{Y} - \mathbb{E}Y| > \varepsilon\right) < \delta$$

"probably approximately correct" (PAC)

means $\mathbb{E}[Y_i]$ for any i

If we just use $\hat{Y} = \frac{1}{N}(Y_1 + \cdots + Y_N)$ how large must N be to guarantee this?

"Sample complexity of (ε, δ)-PAC estimation"

First set $X_i = Y_i / M$. $\hat{X} = \hat{Y}/M$.

$$|\hat{X} - \mathbb{E}X| > \varepsilon/M \iff |\hat{Y} - \mathbb{E}Y| > \varepsilon.$$

$= \mathbb{E}X_i$

$$\iff \left|X_1 + \cdots + X_N - \mathbb{E}(X_1 + \cdots + X_N)\right| > \frac{\varepsilon N}{M}.$$

Aside: Another inequality called Hoeffding bound, with almost same proof, says

$$\Pr\left(\left| X_1 + \dots + X_N - E(X_1 + \dots + X_N) \right| > \lambda \right)$$

$$< 2 \exp\left(- \frac{2\lambda^2}{N} \right)$$

assuming X_i is $[0,1]$-valued

We're applying Hoeffding with $\lambda = \frac{\varepsilon N}{M}$ and we want

$$2 \exp\left(- \frac{2\lambda^2}{N} \right) < \delta$$

$$\iff \quad -\frac{2\lambda^2}{N} < \ln\left(\frac{\delta}{2}\right)$$

(multiply by $-\frac{1}{2}$)

$$\iff \quad \frac{\lambda^2}{N} > \frac{1}{2} \ln\left(\frac{2}{\delta}\right)$$

$$\iff \quad \frac{\varepsilon^2 N}{M^2} > \frac{1}{2} \ln\left(\frac{2}{\delta}\right)$$

$$\iff \quad N > \frac{M^2}{2\varepsilon^2} \ln\left(\frac{2}{\delta}\right)$$

"relative range" squared

log of inverse failure probability

To become 10 times more certain of success, average in another $1.2\, M^2/\varepsilon^2$ samples.

Empirical Risk Minimization

We have data samples $z_1, \ldots, z_N$.
hypotheses $h_1, \ldots, h_M$.
loss function $L(h,z) \in [0,1]$

$L(h,z)$ specifies how badly hypothesis h explains data point z.

Example: $z_i = (\phi, y)$ ϕ = feature vector
y = label

h: a function from feature vectors to labels

$$L(h,z) = \begin{cases} \emptyset & \text{if } y = h(\phi) \\ 1 & \text{if } y \neq h(\phi). \end{cases}$$

ERM is the algorithm that outputs hypothesis $h_{ERM} \in \{h_1, \ldots, h_M\}$ that minimizes

$$\frac{1}{N} \sum_{j=1}^{N} L(h, z_j).$$

Goal. Analyze generalization error of ERM.

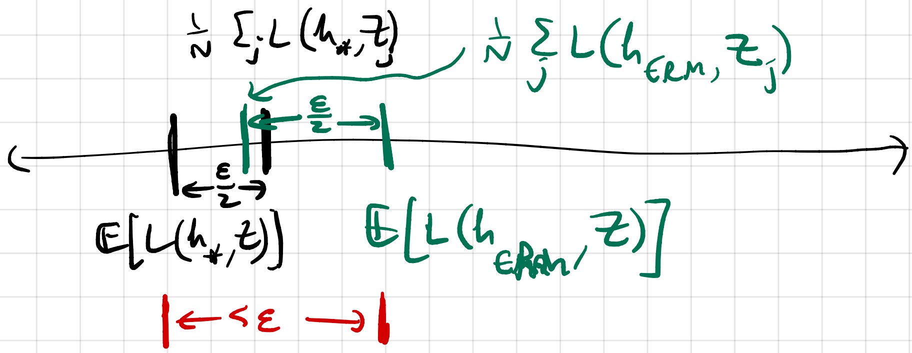
If $z_1, \ldots, z_N$ are i.i.d. samples from a distribution on z's, and z is one more sample (unseen during training) how large must N be so that with probability at least $1-\delta$,

$$(*) \qquad \mathbb{E}[L(h_{ERM}, Z)] \le \varepsilon + \min_{i \in [m]} \mathbb{E}[L(h_i, Z)].$$

Let h_* be the minimizer on the RHS above.

One way to guarantee that inequality $(*)$ happens is to insist that $\forall i \in [m]$,

$$(**) \qquad \left| \frac{1}{N} \sum_{j=1}^{N} L(h_i, Z_j) - \mathbb{E}[L(h_i, Z)] \right| < \frac{\varepsilon}{2}.$$



How large must N be so that $(**)$ holds with probability $\ge 1-\delta$ for any specific i?

$$\frac{\cancel{M}^2}{2\cancel{\varepsilon}^2} \ln\left(\frac{2}{\delta}\right) = \frac{2}{\varepsilon^2} \ln\left(\frac{2}{\delta}\right).$$

1 $\varepsilon/2$

Drive failure probability down to $\frac{\delta}{m}$ for each i by using $\boxed{\frac{2}{\varepsilon^2} \ln\left(\frac{2M}{\delta}\right)}$ samples.

$$\text{Now, } \Pr(\exists i \text{ violating } (**)) \le \mathbb{E}[\# i \text{ for which } (**) \text{ fails to hold}]$$

$$\le \sum_{i=1}^{M} \Pr((**) \text{ violated for } i)$$

$$\le M \cdot (\delta/M) = \delta$$