last lecture:

1) How much data sufficient to learn from a distribution $\mathbb{P}$
when I know $\exists\, h^* \in \mathcal{H}$ $\underline{\underline{err_P(h^*) = 0}}$

$$m \geq \underline{\underline{\frac{1}{\varepsilon}}}\left(\underline{\underline{\ln(|H|)}} + \ln\left(\frac{1}{\delta}\right)\right)$$ you can learn.

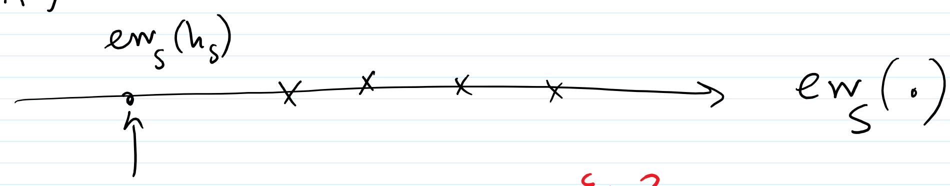
$\forall\, \underline{\underline{h_S}} \in \mathcal{H}$ $\quad err_P(h_S) < \varepsilon \quad$ w.p $1-\delta$

h_S : $err_S(h_S) = 0$

2) What if there is no $h^* \in \mathcal{H}$, s.t. $err_P(h^*) \approx 0$?

3) what if $|\mathcal{H}|$ is infinite?

Alg



$$err_P(h_S) - err_P(h^*) = \underbrace{err_P(h_S) - err_S(h_S)}_{\leq \varepsilon/2\,?} + \underbrace{\boxed{err_S(h_S) - err_S(h^*)}}_{\leq 0}$$
$$+ \underbrace{err_S(h^*) - err_P(h^*)}_{\leq \varepsilon/2\,?} \leq \varepsilon$$

If for all $h \in \mathcal{H}$ $\quad \underline{|err_S(h) - err_P(h)| \leq \varepsilon/2}$ then

$err_P(h_S) \leq err_P(h^*) + \varepsilon$.

$$\Pr\left[\exists\, h \in \mathcal{H},\ |err_S(h) - err_P(h)| > \varepsilon/2\right] \leq \delta \qquad m?$$

$$\Pr\left[\exists h\in\mathcal{H},\ \left|\mathrm{err}_S(h)-\mathrm{err}_P(h)\right| > \varepsilon/2\right] \le \delta \qquad m?$$

$$= \Pr\left[\left|\mathrm{err}_S(h_1)-\mathrm{err}_P(h_1)\right| > \tfrac{\varepsilon}{2} \ \vee \ \cdots \ \vee \left|\mathrm{err}_S(h_{|H|})-\mathrm{err}_P(h_{|H|})\right| > \varepsilon/2\right]$$

(Union bound)
$$\le \sum_{i=1}^{|H|} \underbrace{\Pr\left[\left|\mathrm{err}_S(h_i)-\mathrm{err}_P(h_i)\right| > \varepsilon/2\right]}_{\text{for a fixed } h \text{ that doesn't depend on}}$$

$$\le |H|\cdot 2\exp\left(-\frac{m\varepsilon^2}{2}\right)$$

lemma: for any i

$$\Pr\left[\left|\mathrm{err}_S(h_i)-\mathrm{err}_P(h_i)\right| > \frac{\varepsilon}{2}\right] \le 2\exp\left(-2m(\varepsilon)\right)$$

$$= 2\exp\left(\frac{-m\varepsilon^2}{2}\right).$$

Answer:

Hoeffding's inequality: $\quad p$: prob. (head)

$$\Pr\left[\left|\frac{K}{m}-p\right| > \varepsilon\right] \le 2\exp\left(-2m\varepsilon^2\right)$$

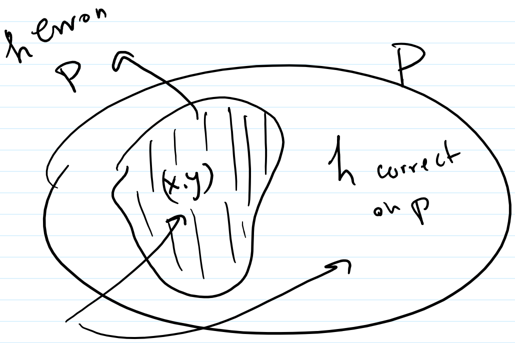
$K \to$ # heads, $m \to$ # tosses

p:

K: # mistakes h makes on S $\qquad \frac{K}{m} = \mathrm{err}_S(h)$

Recall: $x_1, \ldots, x_m$ i.i.d P
$y_1 \qquad y_m$

$$= \Pr\left[h(x_1)\ne y_1\right] = \Pr\left[h(x_2)\ne y_2\right] = \cdots = \Pr\left[h(x_m)\ne y_m\right] = \mathrm{err}_P(h)$$



$Pr[(////)] = err_P(h)$.

$Pr[[\quad]] = 1 - err_P(h)$.

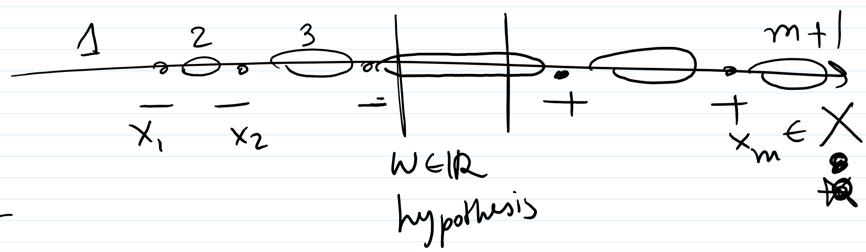
$$2|H| \cdot \exp\left(\frac{-m\varepsilon^2}{2}\right) \leq \delta$$

→ Try at home.
follow previous lecture.

$$m \geq \frac{2}{\varepsilon^2}\left(\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right)$$

$1/\varepsilon^2$

Growth function for 1-D threshold



$X = \mathbb{R}$

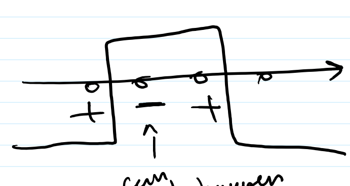
*

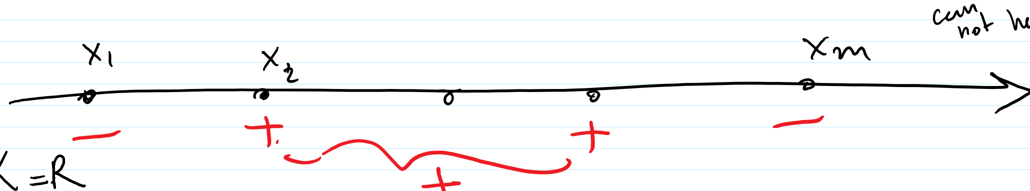
$x_1, x_2 \ldots x_m \in X$

Growth function for intervals

Intervals: $w \leq x \leq w'$ → $h_{w,w'}(x) = +1$

otherwise. $= -1$



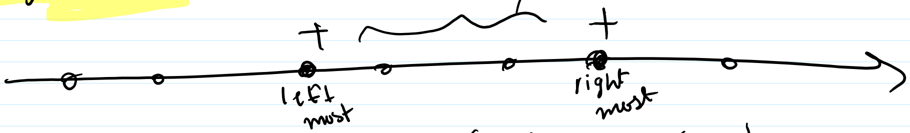


$X = R$

* Any interval will label some subset in the middle as +ve and

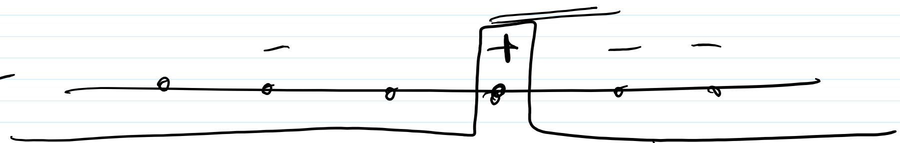
* Any interval will label some subset in the middle as +ve and others as −ve.

* Any labeling is uniquely determined by its left-most and its right most +ve points



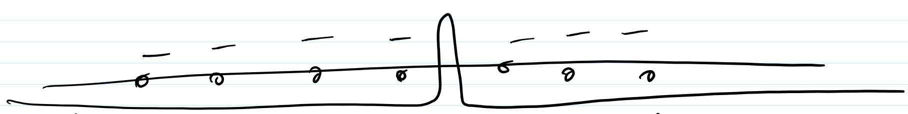
2 different points from m: $\binom{m}{2} = \frac{m(m-1)}{2}$

1



choose 1 point out of m: $\binom{m}{1} = m$

0



$\binom{m}{0} = 1$: choose 0 points out of m

labelings: $\binom{m}{0} + \binom{m}{1} + \binom{m}{2} = O(m^2) << 2^m$.