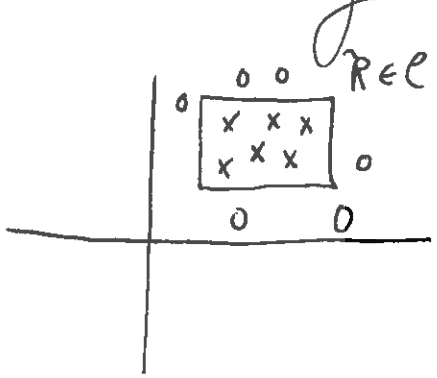
FoSML: Lecture 2

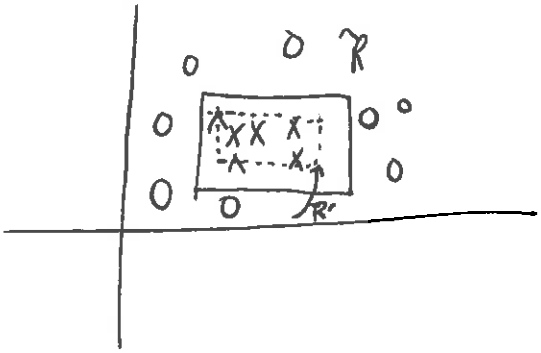
Remark: The concept class $\mathcal{C}$ is known to $\mathcal{A}$, but $c \in \mathcal{C}$ is not. The algorithm must work well (quantified by PAC definition) for all elements of $\mathcal{C}$.

ex (Axis-aligned rectangles). Let $\mathcal{C}$ be the concept class of all axis-aligned rectangles in $\mathbb{R}^2$. Note that $\mathcal{C}$ is uncountably infinite, but is in some sense quite simple.



o = 0
x = 1

In order to show $\mathcal{C}$ is PAC-learnable, we must consider an algorithm that estimates these rectangles "well" (ε-δ's...) given a sample. Consider the algorithm that simply fits the tightest possible (area-minimizing) rectangle to the data sample labeled 1:



Intuitively $R' \to R$ as the number of samples increases. We want to quantify "$R' \to R$" in terms of error of function estimation, and show that the convergence rate is sufficiently fast as a function of the number of samples n.

"R'" best estimator

Let R_S = [illegible], i.e. the error region, which is a function of the sample S. To prove PAC-learnability, we must analyze $\underset{S \sim D^m}{P}\left[R(R_S) > \varepsilon\right]$.

$$\Leftrightarrow \quad \exp(-m\varepsilon/4) \leq \delta/4$$

$$\Leftrightarrow \quad \frac{-m\varepsilon}{4} \leq \log\left(\frac{\delta}{4}\right)$$

$$\Leftrightarrow \quad \frac{m\varepsilon}{4} \geq \log\left(\frac{4}{\delta}\right)$$

$$\Leftarrow \quad m \geq \frac{4}{\varepsilon}\log\left(\frac{4}{\delta}\right).$$

So, we have shown that $m \geq \frac{4}{\varepsilon}\log\left(\frac{4}{\delta}\right)$

$$\rightarrow \mathbb{P}\left(R(\mathcal{R}_S) \leq \varepsilon\right)$$

$$= 1 - \mathbb{P}\left(R(\mathcal{R}_S) \geq \varepsilon\right)$$

$$\geq 1 - \delta.$$

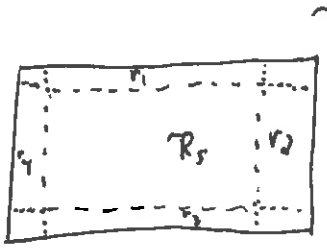
Note that $\underbrace{\frac{4}{\varepsilon}}_{\text{linear in } \frac{1}{\varepsilon}} \underbrace{\log\left(\frac{4}{\delta}\right)}_{\text{logarithmic in } \frac{1}{\delta}} \Rightarrow$ polylogarithmic $\Rightarrow$ PAC learnable.

- In the homework, you will see that other concept classes aren't quite so simple...

. Note that our scheme for axis-aligned rectangles was consistent, meaning in this case that there was no error on the training set.

. We now consider a generalization error estimate that holds for any finite concept

In this particular example, we can analyze this probability geometrically:



The r_i are chosen to have probability about exactly $\varepsilon/4$

(prob = area under a uniform distribution)

- In particular, due to the rectangle geometry, we may argue that

$$\mathbb{P}_{S\sim D}\left(R(R_S)>\varepsilon\right) \le \mathbb{P}_{S\sim D}\left[\bigcup_{i=1}^{4}\left\{R_S \cap r_i = \emptyset\right\}\right]$$

Indeed, $R(R_S)>\varepsilon$ $\Rightarrow$ at least one of these exterior rectangles isn't intersecting with R_S

- Recalling $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \le \sum_{i=1}^{\infty}\mathbb{P}(A_i)$ (union bound), we get

$$\mathbb{P}_{S\sim D}\left(R(R_S)>\varepsilon\right) \le \sum_{i=1}^{4}\mathbb{P}_{S\sim D}\left(\left\{R_S\cap r_i=\emptyset\right\}\right)$$

$$\le 4 \max_{i=1,\dots,4}\mathbb{P}_{S\sim D}\left(R_S\cap r_i=\emptyset\right)$$

$$\le 4\left(1-\frac{\varepsilon}{4}\right)^m$$

$$\le 4\exp\left(-m\varepsilon/4\right), \quad \text{where we note } 1-\varepsilon \le \exp(-\varepsilon)$$

- So, to ensure $4\exp(-m\varepsilon/4) \le \delta$, it suffices to solve for m:

$$4\exp(-m\varepsilon/4) \le \delta$$

Class and consistant algorithm.

Theorem (Learning bound, finite $\mathcal{H}$, consistant algorithm): Let $|\mathcal{H}| < \infty$, $h \in \mathcal{H}$ s.t. $h: \mathcal{X} \to \mathcal{Y}$. Let $\mathcal{A}$ be a consistant algorithm, i.e. for any target concept $c \in \mathcal{H}$, and for any i.i.d. sample S, the hypothesis h_S produced by $\mathcal{A}$ satisfies $\hat{R}_S(h_S) = 0$.

Then, $\forall \varepsilon, \delta > 0$, $m \geq \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)\right)$

$$\Rightarrow \mathbb{P}_{S \sim D}\left(R(h_S) \leq \varepsilon\right) \geq 1 - \delta.$$

Remark: It is equivalent to say $R(h_S) \leq \frac{1}{m}\left(\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)\right)$ with probability at least $1-\delta$.

Proof: Let $\mathcal{H}_\varepsilon = \{h \in \mathcal{H} \mid R(h) > \varepsilon\}$, for any $\varepsilon > 0$. Now, if we have m training samples and h is correct on all of them (i.e. h is consistant), this happens with probability

$$\mathbb{P}(\hat{R}_S(h) = 0) \leq (1-\varepsilon)^m.$$

($\hat{R}_S(h) = 0$: consistent, i.e. 0 empirical error; $\mathbb{P}(\text{wrong on one point}) = \varepsilon$.)

By a union bound, we see

$$\begin{aligned}\mathbb{P}(\exists\, h \in \mathcal{H}_\varepsilon \mid \hat{R}_S(h) = 0) &= \mathbb{P}\left(\{\hat{R}_S(h_1) = 0\} \cup \{\hat{R}_S(h_2) = 0\} \cup \dots \{\hat{R}_S(h_{|\mathcal{H}_\varepsilon|}) = 0\}\right) \\ &\leq \sum_{h \in \mathcal{H}_\varepsilon} \mathbb{P}\left(\hat{R}_S(h) = 0\right)\end{aligned}$$

$$\leq \sum_{h \in H_\varepsilon} (1-\varepsilon)^m$$

$$= |H_\varepsilon| (1-\varepsilon)^m$$

$$\leq |H| (1-\varepsilon)^m$$

$$\leq |H| e^{-m\varepsilon}.$$

Therefore, $\mathbb{P}_{S \sim D^m} \left(R(h_S) \leq \varepsilon \right)$

$$= 1 - \mathbb{P}\left(R(h_S) \geq \varepsilon \right)$$

$$\geq 1 - \delta$$

provided $|H| e^{-m\varepsilon} \leq \delta$

$$\Leftrightarrow \quad e^{-m\varepsilon} \leq \frac{\delta}{|H|}$$

$$\Leftrightarrow \quad -m\varepsilon \leq \log(\delta) - \log(|H|)$$

$$\Leftrightarrow \quad m\varepsilon \geq \log\left(\frac{1}{\delta}\right) + \log(|H|)$$

$$\Leftrightarrow \quad m \geq \frac{1}{\varepsilon}\left(\log\left(\frac{1}{\delta}\right) + \log(|H|) \right). \quad \blacksquare$$