Tyler Olivieri HW7

1) An axis aligned rectangle classifier in the plane is a classifier that assigns the value 1 to a point iff it is inside a certain rectangle. Formally, given real numbers $a_1 \leq b_1$, $a_2 \leq b_2$, define the classifier $h(a_1, b_1, a_2, b_2)$

$$h(a_1,b_1,a_2,b_2)(x_1,x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$

The class of all axis aligned rectangles in the plane is defined as $\mathcal{H}^2_{rec} = \{ h(a_1, b_1, a_2, b_2) : a_1 \leq b_1 \text{ and } a_2 \leq b_2 \}$

We Assume realizability assumption

1. Let A be the algorithm that returns the smallest rectangle enclosing all positive examples in the training set. Show that A is an ERM.

A is an ERM if

$$A = \underset{h(a_1,b_1,a_2,b_2) \in \mathcal{H}^2_{rec}}{\operatorname{argmin}} L_S(h(a_1,b_1,a_2,b_2))$$

Since A contains all positive examples, it will correctly classify all positive examples due to the definition of the classifier.

Due to the realizability assumption, we have a classifier / Algorithm A^* s.t. $L_S(A^*) = 0$

Thus, for a classifier A^* to have $L_S(A^*) = 0$ it must contain all positive examples and no negative examples. The assumption assumes existance of such an A^*.

Consider the case where $a_1, b_1, a_2,$ or b_2 of A realize a new boundary $a_1', b_1', a_2',$ or b_2' s.t.

$$a \le a_1', \quad b_1' \le b, \quad a_2 \le a_2', \quad b_2' \le b_2$$

in other words, the area of A is smaller. Since A is the tightest/smallest rectangle that contains all positive examples it must have a positive example on the boundary of the rectangle. ~~If~~ this wasn't the case, there would exist a smaller A.

Therefore, decreasing the area of A by translating a boundary would result in a new classifier A' that would misclassify the example that was on the boundary before translation

We also know $A \subseteq A^*$ (shown later), so we do not need to consider negative examples now being correctly classified.

$$\Rightarrow L_S(A) < L_S(A')$$

The other case is when the area of A gets larger by translating a boundary in the opposite direction. Since A contains all positive examples increasing the area of A will only allow the possibility that the new classifier A'', will incorrectly classify a negative example,

$$\Rightarrow L_S(A) \le L_S(A'')$$

In conclusion, $L_S(A) < L_S(A')$

$L_S(A) \leq L_S(A'')$

and the set of all rectangle classifiers

$$H^2_{rec} := \{ A, \{A'\}, \{A''\} \}$$

area more than A

the set of all classifiers w/ area less than A.

thus $A = \text{argmin}_{h(a_1, b_1, a_2, b_2) \in H^2_{rec}} L_S(h(a_1, b_1, a_2, b_2))$

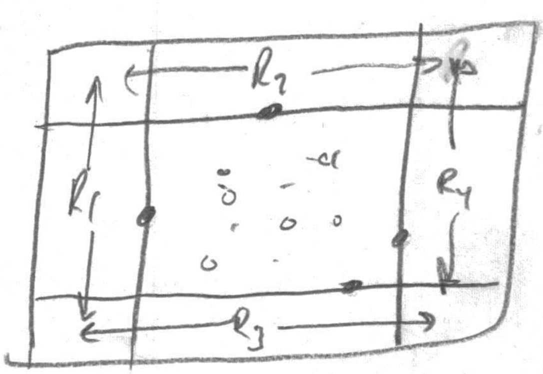
$\Rightarrow$ and A is an ERM.

2. Show that if A recieves a training set of size $\geq \frac{4\log(4/\delta)}{\varepsilon}$ then, w/ probability of at least $1-\delta$ it returns a hypothesis with error at most ε.

This will be done in steps following the outline of the problem.

if $R_1, \ldots, R_4$ have $\varepsilon/4$ mass each.

and samples lie in R_1 to R_4 $L_D(h) < \epsilon$



this is worst case scenario with error $= \epsilon$ because samples lie on boundaries of R_1, R_2, R_3, and R_4.

Show that $R(S) \subseteq R^*$

$R(S)$ is rectangle returned by A.

R^* is the rectangle that generates labels.

A realization of R^*, r^*, induces a realization $r(s)$ of $R(S)$. since $R(S)$ is defined by the smallest rectangle containing all positive instances, and R^* contains all positive instances as it is the true probability distribution,

$$\Rightarrow R(S) \subseteq R^*$$ (this proves $A \subseteq A^*$ used earlier)

Show that if S contains positive examples in all of the rectangles R_1, R_2, R_3, R_4, then the hypothesis returned by A has error at most ϵ.

Due to realizability assumption, we have a perfect classifying rectangle, $R(S)$. If a positive example appears in R_1 (as defined in the problem statement), then we would incorrectly classify at most $\epsilon/4$ data points under D.

Similarly for rectangles R_2, R_3, and R_4 contributing at most $\epsilon/4$ to $L_D(h)$. Therefore, due to geometry of the rectangles, the contribution to $L_D(h)$ by A is at most ϵ (as $R1, R2, R3, R4$ can overlap) and the error is at most the union of the area of these 4 rectangles.

For each $i \in \{1, \dots, 4\}$, upper bound the probability that S does not contain an example from R_i

$$\Pr\{ S \text{ does not contain an example from } R_i \}$$

$$= \Pr\{ \forall (x_i, y_i) \in S : x_i \notin R_i \} \leq \prod_{i=1}^{|S|} \left(1 - \frac{\varepsilon}{4}\right) = \left(1 - \frac{\varepsilon}{4}\right)^{|S|}$$

$\varepsilon/4$ probability $x_i \in R_i \underset{iid}{\Rightarrow} (1 - \varepsilon/4)$ probability $x_i \notin R_i$

Use the union bound to conclude the argument

$$\Pr\{ \forall (x_i, y_i) \in S : (x_i \notin R_1 \cup x_i \notin R_2 \cup x_i \notin R_3 \cup x_i \notin R_4) \}$$

by union bound

$$\leq \Pr\{\forall (x_i, y_i) \in S : x_i \notin R_i\} + \Pr\{\forall (x_i, y_i) \in S : x_i \in R_2\}$$

$$+ \Pr\{\forall (x_i, y_i) \in S : x_i \in R_3\} + \Pr\{\forall (x_i, y_i) \in S : x_i \in R_4\}$$

$$= 4\left(1 - \varepsilon/4\right)^{|S|} \leq 4\left(e^{-\varepsilon/4}\right)^{|S|} = 4e^{-\frac{\varepsilon|S|}{4}} \leq \delta$$

using $(1-x) \leq e^{-x}$

simplifying,

$$e^{-\frac{\varepsilon|S|}{4}} \leq \delta/4$$

$$\ln\left(e^{-\frac{\varepsilon|S|}{4}}\right) \leq \ln(\delta/4)$$

$$-\frac{\varepsilon|S|}{4} \leq \ln(\delta/4)$$

conclusion, $\Rightarrow$

$$|S| \geq (4/\varepsilon) \ln(\delta/4)$$

training set size must be greater than $(4/\varepsilon)\ln(\delta/4)$

Tyler Olivieri HW 7 #2

Let X be a discrete domain, and let

$H_{singleton} = \{h_z : z \in X\} \cup \{h^-\}$, where for each $z \in X$, h_z is the function defined by $h_z(x) = 1$ if $x = z$, $h_z(x) = 0$ if $x \neq z$

h^- is simply the all-negative hypothesis, namely, $\forall x \in X, h^-(x) = 0$.

The realizability assumption here implies that the true hypothesis f labels negatively all examples in the domain, perhaps except one.

1. Describe an algorithm that implements the ERM rule for learning $H_{singleton}$ in the realizable setup.

- Due to the realizability assumption, only one example can be positive (1). Thus, the algorithm that follows is ERM.

```
iterate over S:
    if y_i = 1:
        output h_z
        terminate
    else:
        continue

output h^-     (if no example has y_i = 1)
```

Now, since we know from the realizability assumption the true hypothesis f, there is two sceanios 1) all negative 2) (n-1) negative, 1 positive.

The algorithm handles both cases and gives $L_S(h) = 0$

So it has to be ERM.

2. Show that $\mathcal{H}_{singleton}$ is PAC learnable. Provide an upper bound on the sample complexity.

<u>Recall</u> PAC learnable if there exists a function $m_H : (0,1)^2 \longrightarrow \mathbb{N}$ and a learning algorithm w/ the following properties:

- For every $\epsilon, \delta \in (0,1)$, for every distribution D over X, and for every labelling function $f : X \rightarrow \{0,1\}$, if the realizability assumption holds wrt $\mathcal{H}, D, f$, when running the learning algorithm on $m \geq m_H(\epsilon, \delta)$ iid examples generated by D and labelled by f, the algorithm returns a hypothesis h s.t. w/ probability of at least $1-\delta$ (over the choice of the examples), $L_{(D,f)}(h) \leq \epsilon$

WTS $\Pr\{S \mid L_{(D,f)}(h_S) > \epsilon\} \leq \delta$

Previously I mentioned that there is two possibilies for f

1) All negative

2) All negative except 1 sample
   - however in a random same this f could generate all negative.

Consider the first f (all negative), then the algorithm designed in part (a) would return h^- which will always have $L_{(D,f)}(h_S^-) = 0$, thus for any ϵ, $L_{(D,f)}(h_S) > \epsilon$ never occurs and ~~the~~ $\Pr\{S \mid L_{(D,f)}(h_S) > \epsilon\} = 0$ and this is less than δ for any $\delta \in (0,1)$

If f is all negative (except for possibly one sample) the algorithm will output h_z or h_-. If it outputs h_z, than $L_{(D,f)}(h_z) = 0$ and for similar reasoning, $Pr\{S \mid L_{(D,f)}(h_z) > \epsilon\}$ can't occur because $L_{(D,f)}(h_z) = 0$ and $\epsilon > 0$, so $Pr\{S \mid L_{(D,f)}(h_z) > \epsilon\} \leq \delta$ for $\delta > 0$

If the algorithm outputs h^-, well, $L_{(D,f)}(h^-)$ is non-zero. let event E represent the event that there is a positive example. $\bar{E}$ when there is no positive example.

shown in book in 2.3.1 ($\{S|_x : L_{(D,f)}(h_S) > \epsilon\} \leq M$)

$$Pr\{S|_x : L_{(D,f)}(h_-) > \epsilon\} \leq Pr\{S|_x : L_S(h_-) = 0\} = Pr\{\bar{E}\}$$

by definition

when (*) $\epsilon < L_{(D,f)}(h_-) = Pr\{h_-(x) \neq f(x)\} = Pr\{E\}$

$$Pr\{\bar{E}\} = (1 - Pr\{E\})^m$$ (due to iid assumption with m samples)

since $\epsilon < Pr\{E\}$ (from (*))

$$Pr\{S | L_{(D,f)}(h_-) > \epsilon\} \leq Pr\{\bar{E}\} = Pr\{1 - Pr\{E\}\} \leq (1-\epsilon)^m \leq \delta$$

simplifying $(1-\epsilon)^m \leq \delta$

$$m \log(1-\epsilon) \leq \log \delta$$

$$m(-\log(1-\epsilon)) \geq -\log \delta$$

$$m \geq \frac{\log(1/\delta)}{\log(1/(1-\epsilon))}$$

thus if we have enough data, $Pr\{S | L_{(D,f)}(h_-) > \epsilon\} \leq \delta$ and we can conclude that this class, $H_{singleton}$, is PAC-learnable.

Tyler Olivieri HW 7 #3

Let $X = \mathbb{R}^2$, $Y = \{0,1\}$, and let $\mathcal{H}$ be the class of concentric circles in the plane, that is, $\mathcal{H} = \{h_r : r \in \mathbb{R}^+\}$ where $h_r(x) = \mathbb{1}\{\|x\| \le r\}$. Prove that $\mathcal{H}$ is PAC learnable, (assume realizability), and its sample complexity is bounded by

$$m_{\mathcal{H}}(\epsilon, \delta) \le \frac{\log(1/\delta)}{\epsilon}$$

With the realizability assumption for class $\mathcal{H} = \{h_r : r \in \mathbb{R}^+\}$ we can have a learning algorithm that outputs a circle w/ radius r_h that fits S s.t. $L_S(h) = 0$

r_h s.t. $L_S(h) = 0$ choose tightest circle (similar to first rectangle problem)

if we choose r_h to satisfy the above condition, than the $\Pr\{S : L_{(D,f)}(h) > \epsilon\} \le \Pr\{S : L_S(h) = 0\}$ because if $L_{(D,f)}(h) > \epsilon \Rightarrow$ area in between f and h $\ge \epsilon$

$$\le \Pr\{S : L_S(h) = 0\} = \Pr\{h(x_j) = f(x_j), \forall_j\}$$

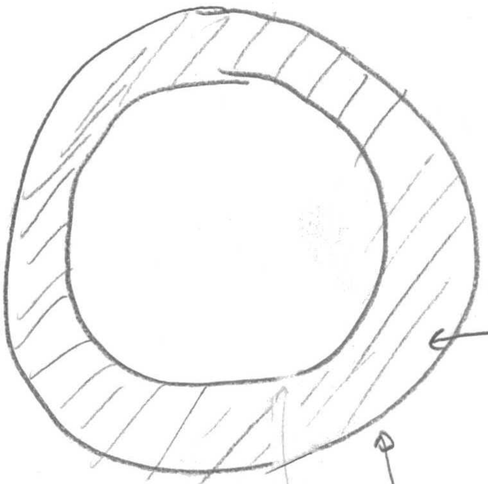
$$= \prod_{j=1}^{m} (1 - L_{(D,f)}(h)) \le \prod_{\substack{j=1 \\ L_{(D,f)}(h) > \epsilon}}^{m} (1-\epsilon) \le (1-\epsilon)^m \le e^{-m\epsilon} \le \delta$$

$$\log(e^{-m\epsilon}) \le \log \delta \Rightarrow -m\epsilon \le \log \delta$$

$$\Rightarrow m\epsilon \ge \log 1/\delta \Rightarrow m \ge \frac{\log(1/\delta)}{\epsilon}$$

and thus $m_{\mathcal{H}}(\epsilon,\delta) \le \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$, (the ceiling preserves bound while giving natural number)

for a given ϵ, δ it is possible to achieve $L_{(D,f)}(h) < \epsilon$ and if $m \ge \log(1/\delta)/\epsilon$



mass of $\overset{R}{\overset{\|}{\text{ring}}}$ $= \epsilon$ under D

true boundary f generating positive labels

$$\Pr\{x_i \notin R\} = (1-\epsilon)$$

$$\Pr\{(x_1, x_2, \dots, x_n) \notin R\} = (1-\epsilon)^m = \Pr\{S \notin R\}$$

is the probability $(x_1, x_2, \dots, x_n) \notin R$

$$\Pr\{S : L_{(D,f)}(h) > \epsilon\}$$

$$\leq \Pr\{S \notin R\} = (1-\epsilon)^m \leq e^{-m\epsilon} \leq \delta$$

If no sample point x_i is in R, our tightest circle (learned classifier) will have $L_D(h) > \epsilon$ because the mass of ring between learned classifier and boundary $f > \epsilon$.

$$e^{-m\epsilon} \leq \delta \Rightarrow \log(e^{-m\epsilon}) \leq \log(\delta) \Rightarrow -m\epsilon \leq \log(\delta)$$

$$\Rightarrow m\epsilon \geq \log(1/\delta) \Rightarrow m \geq \log(1/\delta)/\epsilon$$

$$m_H(\epsilon, \delta) \geq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$$

~~Pr.~~ Tyler Olivieri HW 7 Problem 4

Let X be a domain and let $D_1, D_2, \dots, D_m$ be a sequence of distributions over $\mathcal{X}$. Let H be a finite class of binary classifiers over $\mathcal{X}$ and let $f \in H$. Suppose we are getting a sample S of m examples, s.t. the instances are independent but are not identically distributed. the i^{th} instance is sampled from D_i and then y_i is set to be $f(x_i)$. Let $\bar{D}_m$ denote the average, that is,

$$\bar{D}_m = \frac{(D_1 + \dots + D_m)}{m}$$

fix an accuracy parameter $\epsilon \in (0,1)$. Show that

$$\mathbb{P}\{\exists h \in H \text{ s.t. } L_{(\bar{D}_m, f)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0\}$$
$$\le |H| e^{-\epsilon m}$$

$$L_{(\bar{D}_m, f)}(h) = \frac{1}{m} \sum_i L_{(D_i, f)}(h)$$ (from piazza) and book)

$$Pr\{(\exists h \in H \text{ s.t. } L_{(\bar{D}_m, f)}(h) > \epsilon) \text{ and } L_S(h) = 0\}$$

$$\le Pr\{ \cup_{h \in \mathcal{H}_B} (L_{(\bar{D}_m, f)}(h) > \epsilon \text{ and } L_S(h) = 0)\}$$

$\mathcal{H}_B$ is the set of "bad hypothesis" h that satisfy $L_{(\bar{D}_m, f)}(h) > \epsilon$ and $L_S(h) = 0$. This inequality is decribed in chapter 2 of the textbook. Note $\mathcal{H}_B \subseteq \mathcal{H}$

$$Pr\left\{ \cup_{h \in H_B} \left(L_{(\bar{D},f)}^{(h)} > \epsilon \text{ and } L_S(h) = 0 \right) \right\}$$

$$\leq \sum_{i=1}^{|H_B|} Pr\left\{ L_{(\bar{D},f)}^{(h_i)} > \epsilon \text{ and } L_S(h_j) = 0 \right\}$$ by union bound.

$$= \sum_{j=1}^{|H_B|} Pr\left\{ L_S(h_j) = 0 \right\} Pr\left\{ L_{(\bar{D},f)} > \epsilon \mid L_S(h_j) = 0 \right\}$$

by joint probability factorization.

$$\leq \sum_{j=1}^{|H_B|} \left(\prod_{i=1}^{m} \left(1 - L_{(D_i,f)}^{(h)} \right) \right) Pr\left\{ L_{(\bar{D},f)} > \epsilon \mid L_S(h_j) = 0 \right\}$$

because $Pr\left\{ L_S(h) = 0 \right\} = \prod_{i=1}^{m} \left(1 - L_{(D_i,f)}(h) \right)$

shown in book

By geometric-arithmetic mean inequality

$$\leq \sum_{i=1}^{|H_B|} \left(\frac{1}{m} \sum_{i=1}^{m} \left(1 - L_{(D_i,f)}^{(h)} \right) \right)^m Pr\left\{ L_{(\bar{D},f)} > \epsilon \mid L_S(h) = 0 \right\}$$

$$= \sum_{i=1}^{|H_B|} \left(1 - L_{(\bar{D},f)}(h) \right)^m Pr\left\{ L(\bar{D},f) > \epsilon \mid L_S(h) = 0 \right\}$$

$L_{(\bar{D},f)} > \epsilon$

$$\leq \sum_{i=1}^{|H_B|} (1-\epsilon)^m \underbrace{Pr\left\{ L(\bar{D},\epsilon) > \epsilon \mid L_S(h) = 0 \right\}}_{\text{this term is at most } 1}$$

$$\leq |H_B| (1-\epsilon)^m \underset{H_B \subseteq H}{\leq} |H| (1-\epsilon)^m \leq |H| e^{-\epsilon m}$$

which concludes the proof

Tyler Olivieri HW 7 #5

Show that for every probability distribution D, the Bayes optimal predictor f_D is optimal, in the sense that for every classifier g from X to $\{0,1\}$ $L_D(f_D) \leq L_D(g)$

$$L_D(h) = \mathbb{E}_{(x,y)\sim D}\left[1(h(x_i) \neq y_i)\right]$$ by definition

$$= \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[1(h(x_i) \neq y_i)\right]\right]$$ by law of iterative expectation

$$= \mathbb{E}_X\left[P(Y=1|X) \cdot 1(h(x) \neq 1) + P(Y=0|X) \cdot 1(h(x) \neq 0)\right]$$

for any prediction $h(x) = 0$ contributes $P(Y=1|x)$ to error

and $h(x)=1$ contributes $P(Y=0|x)$ to error, thus to

$$\min_h \mathbb{E}_X\left[P(Y=1|X) \cdot 1(h(x) \neq 1) + P(Y=0|x) \cdot 1(h(x) \neq 0)\right]$$

$$L_D(f_D) = \mathbb{E}_X\left[\min(P(Y=1|X), P(Y=0|X))\right]$$ rule to minimize above expression.

minimum possible error.

any other classifier $L_D(g)$ will have larger error.

we notice that the bayes optimum classifier is the decision rule $L_D(f_D)$, thus $L_D(f_D) =$ bayes optimum $\leq L_D(g) =$ any other

because $L_D(f_D)$ is the minimum possible loss w.r.t D