Seminar 2

① The Bayes optimal predictor has the smallest error among all possible classifiers

Let D be a probability distribution over $X \times \{0,1\}$.

The Bayes classifier is defined as: $\eta(x)$

$f_D : X \to \{0,1\}$, $f_D(x) = \begin{cases} 1, & P_{(x,y)\in D}(y=1|x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$

Let $g : X \to \{0,1\}$ be a random classifier. We want to show that $L_D(f_D) \leq L_D(g)$.

$L_D(f_D) = E_{(x,y)\sim D}(\ell(f_D(x,y))) = E_{(x,y)\sim D}(\mathbb{1}_{[f_D(x)\neq y]}) = P_{(x,y)\sim D}(f_D(x)\neq y)$

$\ell(f_D(x,y)) = 0\text{-}1 \text{ loss} = \begin{cases} 1, & f_D(x)\neq y \\ 0, & f_D(x)=y \end{cases} = \mathbb{1}_{[f_D(x)\neq y]}$

$E_{(x,y)\sim D}(\mathbb{1}_{[f_D(x)\neq y]}) = E_{x\sim D_x}\left[\underbrace{E_{y\sim D_{y|x}}[\mathbb{1}_{[f_D(x)\neq y]}|x]}\right]$

$= P_{y\sim D_{y|x}}(f_D(x)\neq y|x)$

$P_{y\sim D_{y|x}}(f_D(x)\neq y|x) = P(y=1|x)\cdot\mathbb{1}_{[\eta(x)<\frac{1}{2}]} + P(y=0|x)\cdot\mathbb{1}_{[\eta(x)\geq\frac{1}{2}]}$

$= \eta(x)\cdot\mathbb{1}_{[\eta(x)<\frac{1}{2}]} + (1-\eta(x))\cdot\mathbb{1}_{[\eta(x)\geq\frac{1}{2}]}$

$= \begin{cases} \eta(x), & \eta(x)<\frac{1}{2} \\ 1-\eta(x), & \eta(x)\geq\frac{1}{2} \end{cases} = \min(\eta(x), 1-\eta(x))$

$L_D(g) = E_{(x,y)\sim D}(\ell(g(x,y))) = E_{(x,y)\sim D}(\mathbb{1}_{g(x)\neq y}) = E_{x\sim D_x}[\underbrace{E_{y\sim D_{y|x}}[\mathbb{1}_{[g(x)\neq y]}|x]}]$

$= P_{y\sim D_{y|x}}(g(x)\neq y|x)$

$P_{y\sim D_{y|x}}(g(x)\neq y|x) = P(g(x)=0, y=1|x) + P(g(x)=1, y=0|x)$

$= P(g(x)=0|x)\cdot P(y=1|x) + P(g(x)=1|x)\cdot P(y=0|x)$

$= P(g(x)=0|x)\cdot\underbrace{\eta(x)}_{\geq\min(\eta(x),1-\eta(x))} + P(g(x)=1|x)\cdot\underbrace{(1-\eta(x))}_{\geq\min(\eta(x),1-\eta(x))}$

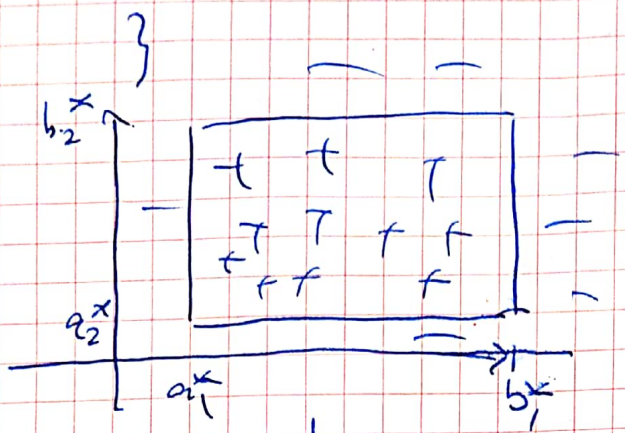
$\geq P(g(x)=0|x)\cdot\min(\eta(x),1-\eta(x)) + P(g(x)=1|x)\cdot\min(\eta(x),1-\eta(x))$

$\geq (P(g(x)=0|x) + P(g(x)=1|x))\cdot\min(\eta(x),1-\eta(x))$

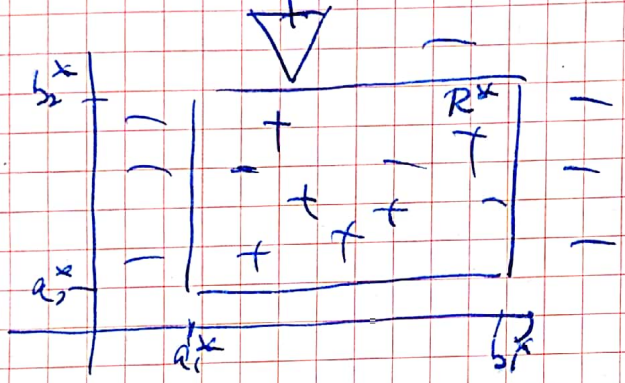
$= \min(\eta(x),(1-\eta(x)) = P_{y\sim D_{y|x}}(f_D(x)\neq y|x)$

(2) $\mathcal{H}_{rec}^2 = \{ h_{(a_1,b_1,a_2,b_2)} : \mathbb{R}^2 \to \{0,1\},\ a_1 \le b_1,\ a_2 \le b_2 \}$

$h_{(a_1,b_1,a_2,b_2)}(x) = \mathbb{1}_{[a_1,b_1]\times[a_2,b_2]}(x) = \begin{cases} 1, & x \in [a_1,b_1]\times[a_2,b_2] \\ 0, & \text{otherwise} \end{cases}$



PAC-learnability in seminar 1



agnostic PAC learnability

Positive labels are flipped with probability $0 < \eta < \frac{1}{2}$.

$S = \{ (x_1, y_1), (x_2, y_2), \dots (x_m, y_m) \}$

$y_i = \begin{cases} 0, & \text{if } x_i \notin [a_1^*, b_1^*] \times [a_2^*, b_2^*] \\ 0, & \text{with prob. } \eta \text{ if } x_0 \in [a_1^*, b_1^*] \times [a_2^*, b_2^*] \\ 1, & \text{with prob. } 1-\eta \text{ if } x_0 \in [a_1^*, b_1^*] \times [a_2^*, b_2^*] \end{cases}$

$R^* = [a_1^*, b_1^*] \times [a_2^*, b_2^*]$

The chance to get a training point labeled as positive is to sample a point from R^* and the label is not flipped so the chance is

$$D(R^*) \cdot (1-\eta)$$

A – the learning algorithm that returns the tightest rectangle containing positive points.

$h_S = A(S)$, $h_S = h_{(a_{1S}, b_{1S}, a_{2S}, b_{2S})}$

$a_{1S} = \min_{(x_i,1)\in S} x_{i1}$ $\quad a_{2S} = \min_{(x_i,1)\in S} x_{i2}$

$b_{1S} = \max_{(x_i,1)\in S} x_{i1}$ $\quad b_{2S} = \max_{(x_i,1)\in S} x_{i2}$

(if S doesn't contain positive samples then return let $z = (z_1, z_2)$ and $h_S = h_{(z_1, z_1, z_2, z_2)}$, where z is a random point in $\mathbb{R}^2$

We want to show that $\mathcal{H}^2_{rec}$ is agnostic PAC-learnable; there exist a function $m_{\mathcal{H}^2_{rec}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that:

for every $\varepsilon > 0$, for every $\delta > 0$, for every distribution $\mathcal{D}$ over $X = \mathbb{R}^2 \times \{0,1\}$, $\mathcal{D} = \mathcal{D}_X \times \mathcal{D}_Y$

when we run the learning algorithm A on a training set S consisting of $m \geq m_{\mathcal{H}^2_{rec}}(\varepsilon, \delta)$ examples sampled i.i.d from $\mathcal{D}$ the algorithm A returns a hypothesis $h_S = A(S)$ from $\mathcal{H}^2_{rec}$ such that

$$\underset{S \sim \mathcal{D}^m}{P}\left(L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \right) \geq 1 - \delta$$

In our case the smallest achievable real error is

$$\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = L_{\mathcal{D}}(h^*) = \eta \cdot \mathcal{D}_X(R^*)$$

Consider $\varepsilon > 0$, $\delta > 0$ and $\mathcal{D}_X$ a distribution over $\mathbb{R}^2$.

Case 1: if $\mathcal{D}_X(R^*) \leq \varepsilon \Rightarrow h_S$ can only make error on points inside R^*

so $\underset{S \sim \mathcal{D}^m}{P}\left(L_{\mathcal{D}}(h_S) \leq \varepsilon \right) = 1$ ✓.

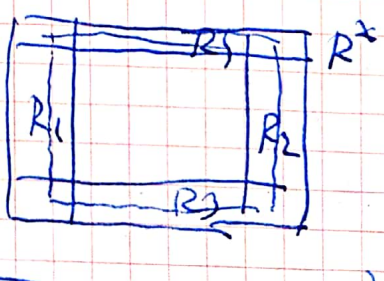
Case 2: if $\mathcal{D}_X(R^*) > \varepsilon$

Construct rectangles R_1, R_2, R_3, R_4 (like in similar 1) such that $\mathcal{D}_b(R_i) = \frac{\varepsilon}{4}$

i) if $h_S = A(S)$ intersects all R_i, $i = \overline{1,4}$ then h_S will make errors on:

R_S, because of flipping $\to \mathcal{D}(R_S) = \eta$

$R^* \setminus R$, $\mathcal{D}_X(R^* \setminus R) < \varepsilon$



So in this case we have:

$$P_{S\sim D^m}\left(L_D(h_S) \le \eta \cdot D(R^*) + \varepsilon\right) = 1.$$

ii) if $h_S = A(S)$ doesn't intersect a rectangle R_i

Denote $F_i = \{S \mid S \sim D_X^m$ s.t R_S – the rectangle learned by $A(S)$ doesn't intersect $R_i\}$

$$P_{S\sim D^m}\left(L_D(h_S) > \eta \cdot D(R^*) + \varepsilon\right) \le \sum_{i=1}^{4} D_X^m(F_i)$$

$D_X^m(F_i)$ = the probability of sampling m points and none of them is a positive point in R_i

$$= \left(1 - \frac{\varepsilon}{4} + \frac{\varepsilon}{4}\cdot\eta\right)^m = \left(1 - \frac{\varepsilon}{4}(1-\eta)\right)^m$$

($1 - \frac{\varepsilon}{4}$: probability of sampling a point outside R_i; $\frac{\varepsilon}{4}\cdot\eta$: prob of sampling a point in R_i but flipping its label)

$$1 - x \le e^{-x}$$

$$1 - \frac{\varepsilon}{4}(1-\eta) \le e^{-\frac{\varepsilon}{4}(1-\eta)}$$

So: $P_{S\sim D^m}\left(L_D(h_S) > \eta \cdot D(R^*) + \varepsilon\right) \le 4 \cdot e^{-\frac{\varepsilon}{4}(1-\eta)m} \le \delta$

$\Leftarrow$

$$4 e^{-\frac{\varepsilon}{4}(1-\eta)m} \le \delta$$

$$e^{-\frac{\varepsilon}{4}(1-\eta)m} \le \frac{\delta}{4} \quad / \log$$

$$m \cdot \left(-\frac{\varepsilon}{4}(1-\eta)\right) \le \log\frac{\delta}{4} \quad / \cdot\left(-\frac{4}{\varepsilon}\right)\cdot\frac{1}{1-\eta}$$

$$m \ge -\frac{4}{\varepsilon}\cdot\frac{1}{1-\eta}\cdot\log\frac{\delta}{4}$$

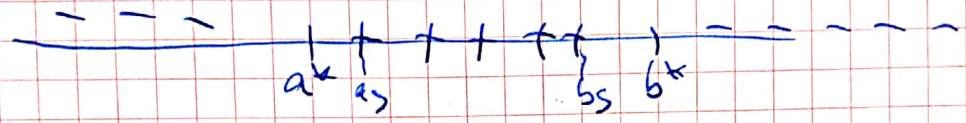
$$\boxed{m \ge \frac{4}{\varepsilon}\frac{1}{1-\eta}\cdot\log\frac{4}{\delta}}$$

(3) $\mathcal{C} = \mathcal{H}_{intervals} = \{ h_{a,b} : \mathbb{R} \to \{0,1\},\ a \le b \}$

$h_{a,b}(x) = \mathbb{1}_{[a,b]}(x) = \begin{cases} 1, & x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$

Consider a training set $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$

We are in the realizability case, $\exists h^* = h_{a^*,b^*} \in \mathcal{H}_{intervals}$ that labels the examples, $y_i = h^*(x_i)$



We want to show that $\mathcal{H}_{intervals}$ is PAC-learnable

Consider A the learning algorithm that gets sample S and outputs

$h_S = A(S)$ = the tightest interval containing all the positive examples.

$h_S = h_{a_S, b_S}$, $a_S = \min_{(x_i,1) \in S} x_i$, $b_S = \max_{(x_i,1) \in S} x_i$; $R_S = [a_S, b_S]$

If there is no $(x_i, 1) \in S$ (S doesn't contain positive examples) then $a_S = b_S = 2$ a random point such that $(2,0) \notin S$.

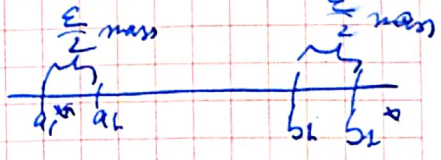
From construction we see that $L_S(h_S) = 0$.

Let $\varepsilon > 0$, $\delta > 0$ and D a distribution over $\mathbb{R}$. What should be $m \ge m_{\mathcal{H}}(\varepsilon, \delta)$ such that $P_{S \sim D^m}(L_{D,h^*}(h_S) > \varepsilon) < \delta$

Case 1: if $D([a^*, b^*]) \le \varepsilon$ then $P_{S \sim D^m}(L_{D,h^*}(h_S) > \varepsilon) = 0$ ✓

Case 2: if $D([a^*, b^*]) > \varepsilon$

Build R_1 and R_2, $R_1 = [a_1^*, a_1]$, $R_2 = [b_1, b_1^*]$



such that $D(R_1) = D(R_2) = \varepsilon/2$.

If $R_S \cap R_1 \neq \emptyset$ and $R_S \cap R_2 \neq \emptyset$ then $P_{S \sim D^m}(L_{D,h^*}(h_S) > \varepsilon) = 0$ ✓

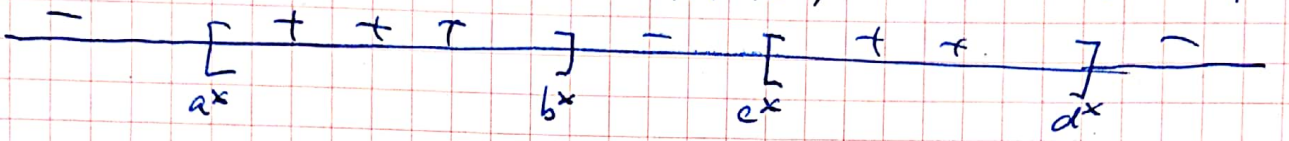
Else $P_{S \sim D^m}(L_{D,h^*}(h_S) > \varepsilon) \le 2 \cdot (1 - \frac{\varepsilon}{2})^m \le 2 \cdot e^{-\frac{\varepsilon}{2} m} < \delta$

$\Rightarrow \boxed{m > \frac{2}{\varepsilon} \log \frac{2}{\delta}}$

(4) PAC-learning algorithm for the class $\mathcal{C}_2$ formed by unions of two closed intervals: $\mathcal{C}_2 = \{ h_{(a,b,c,d)} : \mathbb{R} \to [0,1],\ h_{(a,b,c,d)} = \mathbb{1}_{[a,b]\cup[c,d]}$
$a \le b \le c \le d,\ h_{(a,b,c,d)}(x) = \begin{cases} 1, & x \in [a,b]\cup[c,d] \\ 0, & \text{otherwise} \end{cases} \}$

Consider $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$,

where $y_i = h^*(x_i)$, $h^* = h_{(a^*,b^*,c^*,d^*)}$ – realizability assumption



Consider the following learning algorithm A that takes input S:

i) Sort S in ascending order of x_i

ii) Go over the sorted training examples and take the intervals where consecutive training examples labeled as positive start and end the interval. You can obtain one or two intervals.

iii) If you obtained just one interval you can have $a_S = \min_{(x_i, y_i),\ y_i = 1} x_i$,

$b_S = \max_{y_i = 1} x_i$, $c_S = d_S = b_S$.

If you obtained two intervals then $a_S = \min_{y_i = 1} x_i$ $\quad a_S \le b_S < c_S \le d_S$

$d_S = \max_{y_i = 1} x_i$

Return $h_S = h_{(a_S, b_S, c_S, d_S)} = \mathbb{1}_{[a_S, b_S]\cup[c_S, d_S]}$

We need to find $m \ge m_{\mathcal{C}_2}(\varepsilon, \delta)$ such that for $\varepsilon > 0$, $\delta > 0$ and for every $\mathcal{D}$ distribution over $\mathbb{R}$ we have that

$$P_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}, h^*}(h_S) > \varepsilon \right) < \delta$$

Let $\varepsilon > 0$, $\delta > 0$ and let $\mathcal{D}$ be a distribution over $\mathbb{R}$.

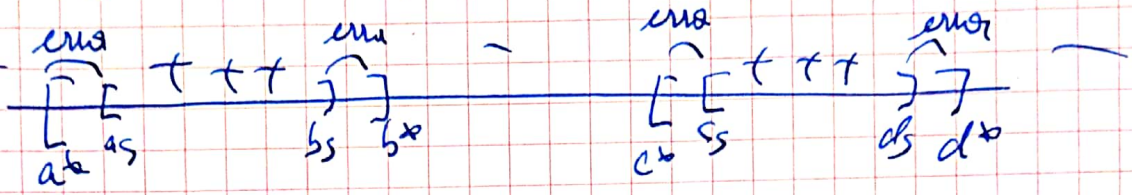
The region where h_S can make error is always $\subseteq [a^*, d^*]$

Case 1 So, if $\mathcal{D}([a^*, d^*]) \le \varepsilon$ then $P_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}, h^*}(h_S) > \varepsilon \right) = 0$.

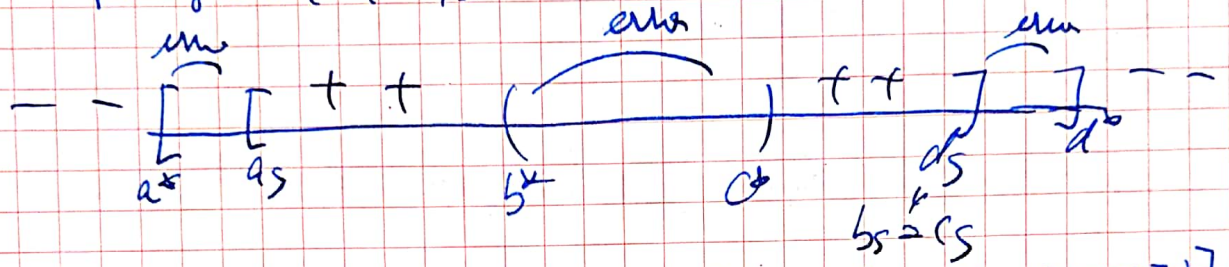
Case 2 | $\mathcal{D}([a^*, b^*]) > \varepsilon$

Three types of error that h_S can make are:

- false negatives in $[a^*, b^*)$ and $[c^*, d^*]$



- false positive in (b^*, c^*) if sample S does not contain any point sampled from (b^*, c^*).



Denote $L_{FP}(h_S) = \underset{x\sim D}{P}[x \in [a_S, b_S] \setminus ([a^*, b^*] \cup [c^*, d^*])]$

$= \underset{x\sim D}{P}[x \in (b^*, c^*) \subseteq [a_S, b_S] \cup [c_S, d_S]]$

$L_{FN,1}(h_S) = \underset{x\sim D}{P}(x \in [a^*, b^*] \setminus [a_S, b_S])$

$L_{FN,2}(h_S) = \underset{x\sim D}{P}(x \in [c^*, d^*] \setminus [c_S, d_S])$

So, if we want to have $L_{D,h^*}(h_S) > \varepsilon$ then one of the number $L_{FP}(h_S)$, $L_{FN,1}(h_S)$, $L_{FN,2}(h_S)$ must be $\geq \frac{\varepsilon}{3}$.

Define $F_1 = \{S \sim D^m \mid L_{FP}(h_S) > \frac{\varepsilon}{3}\}$

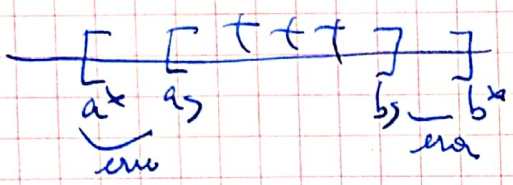
$F_2 = \{S \sim D^m \mid L_{FN,1}(h_S) > \frac{\varepsilon}{3}\}$

$F_3 = \{S \sim D^m \mid L_{FN,2}(h_S) > \frac{\varepsilon}{3}\}$

So, $\underset{S\sim D^m}{P}(L_{D,h^*}(h_S) > \varepsilon) \leq \underset{S\sim D^m}{P}(F_1 \cup F_2 \cup F_3) \leq \sum_{i=1}^{3} P(F_i)$

$$P(F_1) = \underset{S\sim D^m}{P}\left(L_{FP}(h_S) > \frac{\varepsilon}{3}\right) = (\text{this means that } D([b^*, c^*]) > \frac{\varepsilon}{3}$$
and no point from $[b^*, c^*]$ is sampled in $S) \le (1-\frac{\varepsilon}{3})^m \le e^{-\frac{\varepsilon}{3}m}$

$$P(F_2) = \underset{S\sim D^m}{P}\left(L_{FN,L}(h_S) > \frac{\varepsilon}{3}\right)$$



Construct $R_1 = [a^*, a_0)$ and $R_2 = [b_0, b^*)$ such that
$D(R_1) = D(R_2) = \frac{\varepsilon}{6}$.
if $[a_S, b_S) \cap R_1 \neq \phi$ and $[a_S, b_S] \cap R_2 \neq \phi$ then the error made by h_S on $[a^*, b^*)$ is smaller than $\frac{\varepsilon}{6} + \frac{\varepsilon}{6} = \frac{\varepsilon}{3}$.

So $L_{FN,L}(h_S) > \frac{\varepsilon}{3} \Rightarrow [a_S, b_S] \cap R_1 = \phi$ or $[a_S, b_S] \cap R_2 = \phi$.

define $F_{21} = \{S\sim D^m \mid [a_S, b_S] \cap R_1 = \phi\}$
$F_{22} = \{S\sim D^m \mid [a_S, b_S] \cap R_2 = \phi\}$

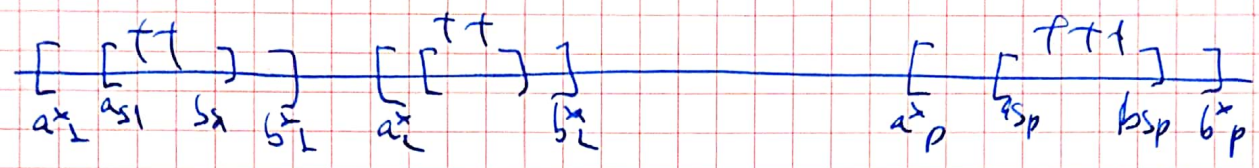
$$P(F_2) \le P(F_{21} \cup F_{22}) \le P(F_{21}) + P(F_{22}) = 2\cdot\left(1-\frac{\varepsilon}{6}\right)^m$$
$$\le 2\cdot e^{-\frac{\varepsilon}{6}m}$$

In the same way we can prove that
$$P(F_3) \le 2\cdot e^{-\frac{\varepsilon}{6}m}.$$

So, we obtain that
$$\underset{S\sim D^m}{P}\left(L_{D,h^*}(h_S) > \varepsilon\right) \le e^{-\frac{\varepsilon}{3}m} + 4\cdot e^{-\frac{\varepsilon}{6}m} \le e^{-\frac{\varepsilon}{6}m} + 4e^{-\frac{\varepsilon}{6}m}$$
$$= 5e^{-\frac{\varepsilon}{6}m} < \delta.$$
$$\Rightarrow e^{-\frac{\varepsilon}{6}m} < \frac{\delta}{5} \quad \Big| \log \Rightarrow -\frac{\varepsilon}{6}m < \log\frac{\delta}{5} \quad \Big| \cdot\left(-\frac{6}{\varepsilon}\right)$$
$$\boxed{m > \frac{6}{\varepsilon}\cdot\log\frac{5}{\delta}}$$

On the general case, for $\mathcal{C}_p$ = reunion of p intervals the proof is similar, the only differences are that:

- there are $(p-1)$ regions of false positives
- $2p$ regions of false negatives



So we have

$$\left| m \geq \frac{2(2p-1)}{\varepsilon} \times \log \frac{2+2p-1}{\delta} \right|$$

$$\left| m \geq \frac{2(2p-1)}{\varepsilon} \times \log \frac{3p-1}{\delta} \right|$$

time complexity → given by sorting $S = O(m \log m)$

5) Let $h \in \mathcal{H}$, with $L_{(\bar{D}_m, f)}(h) > \varepsilon \Rightarrow$

$$\underset{x \sim \bar{D}_m}{P}(h(x) \neq f(x)) > \varepsilon \iff \underset{x \sim \bar{D}_m}{P}(h(x) = f(x)) = 1 - \underset{x \sim \bar{D}_m}{P}(h(x) \neq f(x)) < 1 - \varepsilon$$

$\underset{x \sim \bar{D}_m}{P}(h(x) = f(x)) =$ (~~input the sample~~ x can be sampled from each D_i, with probability $\frac{1}{m}$) $=$

$$= \frac{1}{m} \cdot \underset{x \sim D_1}{P}[h(x) = f(x)] + \dots + \frac{1}{m} \cdot \underset{x \sim D_m}{P}[h(x) = f(x)]$$

$$= \frac{1}{m} \sum_{i=1}^{m} \underset{x \sim D_i}{P}[h(x) = f(x)] < 1 - \varepsilon.$$

$$S = \{(x_1, f(x_1)), (x_2, f(x_2)), \dots (x_m, f(x_m)) \mid \text{where } x_i \sim D_i\}$$

h consistent with S if $L_S(h) = 0$

$$\underset{S \sim D_1 \times D_2 \times \dots \times D_m}{P}[L_S(h) = 0] = \prod_{i=1}^{m} \underset{x_i \in D_i}{P}[h(x_i) = f(x_i)]$$

$$= \prod_{i=1}^{m} \underset{x \sim D_i}{P}[h(x) = f(x)]$$

$$= \left[\left(\prod_{i=1}^{m} \underset{x \sim D_i}{P}[h(x) = f(x)] \right)^{\frac{1}{m}} \right]^m$$

$= \text{geometric mean} = (a_1 \cdot a_2 \cdot \dots \cdot a_m)^{\frac{1}{m}}$

where $a_i = \underset{x \sim D_i}{P}[h(x) = f(x)]$ = probability that h correctly labels a point $x \sim D_i$

$$\leq \text{arithmetic mean} = \left(\frac{a_1 + a_2 + \dots + a_m}{m} \right)$$

$$\leq \left[\frac{1}{m} \sum_{i=1}^{m} \underset{x \sim D_i}{P}[h(x) = f(x)] \right]^m < (1 - \varepsilon)^m \leq e^{-\varepsilon m}$$

There are at most $|\mathcal{H}|$ member of h hypotheses.

So, we obtain that

$$P[\exists h \in \mathcal{H} \text{ s.t } L_{(\bar{D}_m, f)}(h) > \varepsilon \text{ and } L_{(S, f)}(h) = 0] \leq |\mathcal{H}| \cdot e^{-\varepsilon m}$$