Advanced Machine Learning Seminar 2

Exercise 1 Prove that the Bayes optimal predictor has the smallest error among all possible classifiers.

Proof Let $\mathcal{D}$ be a probability distribution over $X \times \{0,1\}$. The Bayes classifier is defined as:

$$f_{\mathcal{D}} \colon X \to \{0,1\},\ f_{\mathcal{D}} = \begin{cases} 1, & \overbrace{\underset{(x,y)\sim\mathcal{D}}{P}(y = 1 \mid x)}^{\eta(x)} \geq \frac{1}{2} \\ 0, & otherwise \end{cases}$$

Let $g \colon X \to \{0,1\}$ be a random classifier. We want to show that $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

$$L_{\mathcal{D}}(f_{\mathcal{D}}) = \underset{(x,y)\sim\mathcal{D}}{E}(l(f_{\mathcal{D}}(x,y))) = \underset{(x,y)\sim\mathcal{D}}{E}(\mathbb{1}_{[f_{\mathcal{D}}(x)\neq y]}) = \underset{(x,y)\sim\mathcal{D}}{P}(f_{\mathcal{D}}(x) \neq y)$$

$$l(f_{\mathcal{D}}(x,y)) = \text{0-1 loss} = \begin{cases} 1, f_{\mathcal{D}}(x) \neq y \\ 0, f_{\mathcal{D}}(x) = y \end{cases} = \mathbb{1}_{[f_{\mathcal{D}}(x)\neq y]}$$

$$\underset{(x,y)\sim\mathcal{D}}{E}(\mathbb{1}_{[f_{\mathcal{D}}(x)\neq y]}) = \underset{x\sim\mathcal{D}_x}{E}[\underbrace{\underset{y\sim\mathcal{D}_{y|x}}{E}[\mathbb{1}_{[f_{\mathcal{D}}(x)\neq y]} \mid x]}_{=\underset{y\sim\mathcal{D}_{y|x}}{P}(f_{\mathcal{D}}(x)\neq y|x)}]$$

$$\begin{aligned} \underset{y\sim\mathcal{D}_{y|x}}{P}(f_{\mathcal{D}}(x) \neq y \mid x) &= P(y=1 \mid x)\cdot \mathbb{1}_{[\eta(x)<\frac{1}{2}]} + P(y=0 \mid x)\cdot \mathbb{1}_{[\eta(x)\geq\frac{1}{2}]} \\ &= \eta(x)\cdot \mathbb{1}_{[\eta(x)<\frac{1}{2}]} + (1-\eta(x))\cdot \mathbb{1}_{[\eta(x)\geq\frac{1}{2}]} \\ &= \begin{cases} \eta(x), \eta(x) < \frac{1}{2} \\ 1-\eta(x), \eta(x) \geq \frac{1}{2} \end{cases} = min(\eta(x), 1-\eta(x)) \end{aligned}$$

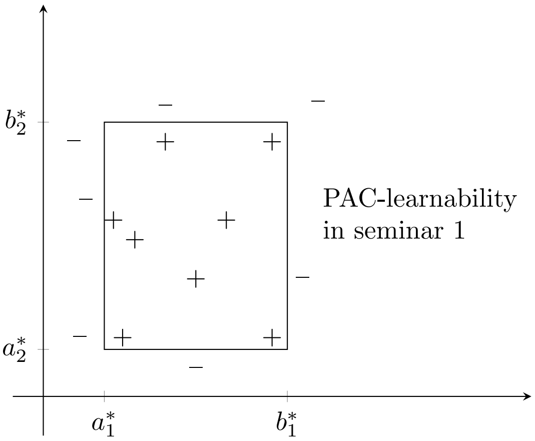
$$L_{\mathcal{D}}(g) = \underset{(x,y)\sim\mathcal{D}}{E}(l(g((x,y))) = \underset{(x,y)\sim\mathcal{D}}{E}(\mathbb{1}_{g(x)\neq y}) = \underset{x\sim\mathcal{D}_x}{E}[\underbrace{\underset{y\sim\mathcal{D}_{y|x}}{E}[\mathbb{1}_{g(x)\neq y} \mid x]}_{=\underset{y\sim\mathcal{D}_{y|x}}{P}(g(x)\neq y|x)}]\ (*)$$

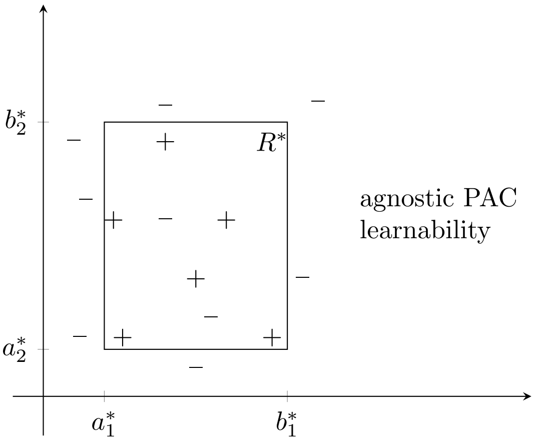
$$\begin{aligned} \underset{y\sim\mathcal{D}_{y|x}}{P}(g(x) \neq y \mid x) &= P(g(x)=0, y=1 \mid x) + P(g(x)=1, y=0 \mid x) \\ &= P(g(x)=0 \mid x)\cdot P(y=1 \mid x) + P(g(x)=1 \mid x)\cdot P(y=0 \mid x) \\ &= P(g(x)=0 \mid x)\cdot \underbrace{\eta(x)}_{\geq min(\eta(x),1-\eta(x))} + P(g(x)=1 \mid x)\cdot \underbrace{1-\eta(x)}_{\geq min(\eta(x),1-\eta(x))} \\ &\geq P(g(x)=0 \mid x)\cdot min(\eta(x), 1-\eta(x)) + P(g(x)=1 \mid x)\cdot min(\eta(x), 1-\eta(x)) \\ &\geq (P(g(x)=0 \mid x) + P(g(x)=1 \mid x))\cdot min(\eta(x), 1-\eta(x)) \\ &= min(\eta(x), 1-\eta(x)) = \underset{y\sim\mathcal{D}_{y|x}}{P}(f_{\mathcal{D}}(x) \neq y|x)\ (**) \end{aligned}$$

From (*) and (**) it follows that $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Exercise 2 *See the entire statement in the Seminar2.pdf.* Show that the algorithm returning the tightest rectangle containing positive points can still PAC-learn axis-aligned rectangles in the presence of this noise.

$$\mathcal{H}^2_{rec} = \left\{ \begin{array}{l} h_{(a_1,b_1,a_2,b_2)} : \mathbb{R}^2 \to \{0,1\}, a_1 \le b_1, a_2 \le b_2, \\ h_{(a_1,b_1,a_2,b_2)}(x) = \mathbb{1}_{[a_1,b_1]\times[a_2,b_2]}(x) = \left\{ \begin{array}{l} 1, x \in [a_1,b_1] \times [a_2,b_2] \\ 0, otherwise \end{array} \right. \end{array} \right\}$$





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

Consider the training set $S = \{(\underline{x_1}, y_1), (\underline{x_2}, y_2), \ldots, (\underline{x_m}, y_m)\}$, where the label y_i is given in the agnostic case by a distribution. We have that:

$$y_i = \left\{ \begin{array}{l} 0,\ if\ \underline{x_i} \notin [a_1^*, b_1^*] \times [a_2^*, b_2^*] \\ 0,\ with\ probability\ \eta\ if\ \underline{x_i} \in [a_1^*, b_1^*] \times [a_2^*, b_2^*] \\ 1,\ with\ probability\ 1-\eta\ if\ \underline{x_i} \in [a_1^*, b_1^*] \times [a_2^*, b_2^*] \end{array} \right.$$

We denote with $R^* = [a_1^*, b_1^*] \times [a_2^*, b_2^*]$ the rectangle determined by h^*.

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 $\mathcal{D}(R^*) \times (1-\eta)$.

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

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

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

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

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

We want to show that $\mathcal{H}^2_{rec}$ is agnostic PAC-learnable: there exists a function $m_{\mathcal{H}^2_{rec}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that for every $\epsilon > 0$, for every $\delta > 0$, for every distribution $\mathcal{D}$ over $Z = \mathbb{R}^2 \times \{0,1\}$, $\mathcal{D} = \mathcal{D}_X \times \mathcal{D}_Y$ when we run the learning algorithm A on a training set S consisting of $m \geq m_{\mathcal{H}^2_{rec}}(\epsilon, \delta)$ 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

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

In our case, the smallest achievable real error is

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

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

Case 1: if $\mathcal{D}_X(R^*) \leq \epsilon \Rightarrow h_S$ can only make errors on points inside R^*, so $\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D}}(h_S) \leq \epsilon) = 1$ ✓

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

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

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

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

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

So in this case we have $\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D}}(h_S) \leq \eta \cdot \mathcal{D}(R^*) + \epsilon) = 1$

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

In order to have $L_{\mathcal{D}}(h_S) > \eta \cdot \mathcal{D}(R^*) + \epsilon$, we need that R_S will not intersect at least one rectangle R_i.

We denote with F_i the event that R_S does not intersect the rectangle R_i based on the points sampled in S, so we have $F_i = \{S \sim \mathcal{D}^m \mid R_S \cap R_i = \emptyset\}$.

This leads to the following:

$$\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D}}(h_S) > \eta \cdot \mathcal{D}(R^*) + \epsilon) \leq \underset{S\sim\mathcal{D}^m}{P}(\overset{\text{at least one } F_i \text{ will happen}}{F1 \cup F2 \cup F3 \cup F4}) \leq \sum_{i=1}^{4} \underset{S\sim\mathcal{D}^m}{P}(F_i)$$

$$\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D}}(h_S) > \eta \cdot \mathcal{D}(R^*) + \epsilon) \leq \sum_{i=1}^{4} \underset{S\sim\mathcal{D}^m}{P}(F_i)$$

$\underset{S\sim\mathcal{D}^m}{P}(F_i) =$ the probability of sampling m points and none of them is a positive point in R_i

$$= \left(\underbrace{1 - \frac{\epsilon}{4}}_{\substack{\text{prob. of sampling}\\ \text{a point outside } R_i}} + \underbrace{\frac{\epsilon}{4} \cdot \eta}_{\substack{\text{prob of sampling}\\ \text{a point in } R_i\\ \text{but flipping its label}}} \right)^m = \Big(1 - \frac{\epsilon}{4}(1 - \eta)\Big)^m$$

$$1 - x \leq e^{-x}$$
$$1 - \frac{\epsilon}{4}(1 - \eta) \leq e^{-\frac{\epsilon}{4}(1-\eta)}$$

So:

$$\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D}}(h_S) > \eta \cdot \mathcal{D}(R^*) + \epsilon) \leq \underline{4 \cdot e^{-\frac{\epsilon}{4}(1-\eta)\cdot m}} < \delta$$

$$4e^{-\frac{\epsilon}{4}(1-\eta)\cdot m} < \delta$$

$$e^{-\frac{\epsilon}{4}(1-\eta)\cdot m} < \frac{\delta}{4} \,\Bigg|\, \log_e$$

$$m \cdot \left(-\frac{\epsilon}{4}\right)(1-\eta) < \log\frac{\delta}{4} \,\Bigg|\cdot \left(-\frac{4}{\epsilon}\right) \cdot \frac{1}{1-\eta}$$

$$m > -\frac{4}{\epsilon} \cdot \frac{1}{1-\eta} \cdot \log\frac{\delta}{4}$$

$$\boxed{m > \frac{4}{\epsilon} \cdot \frac{1}{1-\eta} \cdot \log\frac{4}{\delta}}$$

Exercise 3

$$\mathcal{C} = \mathcal{H}_{intervals} = \left\{ \begin{array}{l} h_{a,b} \colon \mathbb{R} \to \{0,1\}, a \leq b \\ h_{a,b}(x) = \mathbb{1}_{[a,b]}(x) = \left\{ \begin{array}{l} 1, x \in [a,b] \\ 0, otherwise \end{array} \right. \end{array} \right\}$$

Consider a training set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. 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 the training set S and outputs $h_S = A(S) =$ the tightest interval containing all the positive examples.

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

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

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

Let $\epsilon > 0, \delta > 0$ and $\mathcal{D}$ a distribution over R. We want no find how many number of training examples $m \geq m_H(\epsilon, \delta)$ do we need such that:

$$\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D},h^*}(h_S) > \epsilon) < \delta$$

Case 1: if $\mathcal{D}([a^*, b^*]) \leq \epsilon$ then $\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D},h^*}(h_S) > \epsilon) = 0$ ✓

Case 2: if $\mathcal{D}([a^*, b^*]) > \epsilon$

Build R_1 and R_2, $R_1 = [a_1^*, a_1]$, $R_2 = [b_1, b_1^*]$ such that $\mathcal{D}(R_1) = \mathcal{D}(R_2) = \frac{\epsilon}{2}$.

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

Else $\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D},h^*}(h_S) > \epsilon) \leq 2 \cdot \left(1 - \frac{\epsilon}{2}\right)^m \leq 2 \cdot e^{-\frac{\epsilon}{2}m} < \delta \Rightarrow \boxed{m > \frac{2}{\epsilon} \log \frac{2}{\delta}}$

Exercise 4 PAC-learning algorithm for the class $\mathcal{C}_2$ formed by unions of two closed intervals:

$$\mathcal{C}_2 = \left\{ \begin{array}{l} h_{(a,b,c,d)} \colon \mathbb{R} \to [0,1], \ h_{(a,b,c,d)} = \mathbb{1}_{[a,b]\cup[c,d]} \\ a \leq b \leq c \leq d, \ h_{(a,b,c,d)}(x) = \left\{ \begin{array}{l} 1, x \in [a,b] \cup [c,d] \\ 0, otherwise \end{array} \right. \end{array} \right\}$$

Consider $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where $y_i = h^*(x_i), h^* = h_{(a^*,b^*,c^*,d^*)}$ [1].

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 intervals. You can obtain one or two intervals (or no interval = there are no positive training examples). In the case that there are no positive examples take $a_S = b_S = c_S = d_S = z$, where z a random point such that $(z, 0)$ doesn't appear in the training set S.

iii) If you obtained just one interval, you can have $a_S = \min\limits_{\substack{(x_i,y_i) \\ y_i=1}} x_i$, $b_S = \max\limits_{\substack{(x_i,y_i) \\ y_i=1}} x_i$, $c_S = d_S = b_S$

If you obtained two intervals, then $a_S = \min\limits_{\substack{(x_i,y_i) \\ y_i=1}} x_i$, $d_S = \max\limits_{\substack{(x_i,y_i) \\ y_i=1}} x_i$, $a_S \leq b_S < c_S \leq d_S$

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 $m \geq m_{\mathcal{C}_2}(\epsilon, \delta)$ such that for $\epsilon > 0$, $\delta > 0$ and for every $\mathcal{D}$ distribution over $\mathbb{R}$ we have

$$\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D},h^*}(h_S) > \epsilon) < \delta$$

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

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

Case 1: If $\mathcal{D}([a^*, d^*]) \leq \epsilon$, then $\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D},h^*}(h_S) > \epsilon) = 0$.

Case 2 If $\mathcal{D}([a^*, d^*]) > \epsilon$

The types of error that h_S can make are:

[1] realizability assumption

- false negatives in $[a^*, b^*]$ and $[c^*, d^*]$
- false positive in (b^*, c^*) if sample S does not contain any points sampled from (b^*, c^*).

Denote L_{FP}, $L_{FN,1}$, $L_{FN,2}$ these type of errors, where:

$$\begin{aligned}
L_{FP}(h_S) &= \underset{x\sim\mathcal{D}}{P}(x \in [a_S, b_S] \cup [c_S, d_S] \setminus ([a^*, b^*] \cup [c^*, d^*])) \\
&= \underset{x\sim\mathcal{D}}{P}(x \in [b^*, c^*] \subseteq [a_S, b_S] \cup [c_S, d_S]) \\
L_{FN,1}(h_S) &= \underset{x\sim\mathcal{D}}{P}(x \in [a^*, b^*] \setminus [a_S, b_S]) \\
L_{FN,2}(h_S) &= \underset{x\sim\mathcal{D}}{P}(x \in [c^*, d^*] \setminus [c_S, d_S])
\end{aligned}$$

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

Define

$$\begin{aligned}
F_1 &= \left\{S \sim \mathcal{D}^m \mid L_{FP}(h_S) > \frac{\epsilon}{3}\right\} \\
F_2 &= \left\{S \sim \mathcal{D}^m \mid L_{FN,1}(h_S) > \frac{\epsilon}{3}\right\} \\
F_3 &= \left\{S \sim \mathcal{D}^m \mid L_{FN,2}(h_S) > \frac{\epsilon}{3}\right\}
\end{aligned}$$

So,

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

$$\begin{aligned}
P(F_1) &= \underset{S\sim\mathcal{D}^m}{P}\left(L_{FP}(h_S) > \frac{\epsilon}{3}\right) \\
&= \left(\text{this means that } \mathcal{D}([b^*, c^*]) > \frac{\epsilon}{3} \text{ and no point from } [b^*, c^*] \text{ is sampled in } S\right) \\
&\leq \left(1 - \frac{\epsilon}{3}\right)^m \leq e^{-\frac{\epsilon}{3}m} \\
P(F_2) &= \underset{S\sim\mathcal{D}^m}{P}\left(L_{FN,1}(h_S) > \frac{\epsilon}{3}\right)
\end{aligned}$$

Construct $R_1 = [a^*, a_0]$ and $R_2 = [b_0, b^*]$ such that $\mathcal{D}(R_1) = \mathcal{D}(R_2) = \frac{\epsilon}{6}$.

If $[a_S, b_S] \cap R_1 \neq \emptyset$ and $[a_S, b_S] \cap R_2 \neq \emptyset$, then the error made by h_S on $[a^*, b^*]$ is smaller than $\frac{\epsilon}{6} + \frac{\epsilon}{6} \geq \frac{\epsilon}{3}$.

So $L_{FN,1}(h_S) > \frac{\epsilon}{3} \Rightarrow [a_S, b_S] \cap R_1 = \emptyset$ or $[a_S, b_S] \cap R_2 = \emptyset$.

Define

$$\begin{aligned}
F_{21} &= \{S \sim \mathcal{D}^m \mid [a_S, b_S] \cap R_1 = \emptyset\} \\
F_{22} &= \{S \sim \mathcal{D}^m \mid [a_S, b_S] \cap R_2 = \emptyset\}
\end{aligned}$$

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

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

$$\underset{S\sim\mathcal{D}^m}{P}(L_{\mathcal{D},h^*}(h_S) > \epsilon) \leq e^{-\frac{\epsilon}{3}m} + 4 \cdot e^{-\frac{\epsilon}{6}m} \leq e^{-\frac{\epsilon}{6}m} + 4 \cdot e^{-\frac{\epsilon}{6}m} = 5 \cdot e^{-\frac{\epsilon}{6}m} < \delta$$

$$\begin{aligned}
&\Rightarrow e^{-\frac{\epsilon}{6}m} < \frac{\delta}{5} \;\Bigg|\; \cdot \log \\
&\Rightarrow -\frac{\epsilon}{6}m < \log\frac{\delta}{5} \;\Bigg|\; \cdot \left(-\frac{6}{\epsilon}\right) \\
&\boxed{m > \frac{6}{\epsilon} \cdot \log\frac{5}{\delta}}
\end{aligned}$$

In 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

$$\boxed{m \geq \frac{2(2p-1)}{\epsilon} \cdot \log \frac{p+2p-1}{\delta}}$$

$$\boxed{m \geq \frac{2(2p-1)}{\epsilon} \cdot \log \frac{3p-1}{\delta}}$$

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

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

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

$$\begin{aligned}\underset{x\sim\bar{D}_m}{P}(h(x) = f(x)) &= \left(x \text{ can be sampled from each } D_i\text{, with probability } \frac{1}{m}\right)\\ &= \frac{1}{m}\cdot \underset{x\sim\bar{D}_1}{P}[h(x)=f(x)] + \cdots + \frac{1}{m}\cdot \underset{x\sim\bar{D}_m}{P}[h(x)=f(x)]\\ &= \frac{1}{m}\sum_{i=1}^{m} \underset{x\sim\bar{D}_i}{P}[h(x)=f(x)] < 1-\epsilon\end{aligned}$$

Consider the training set S $= \{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m)) \mid \textit{where } x_i \sim D_i\}$ h consistent with S if $L_S(h) = 0$.

$$\begin{aligned}\underset{S\sim D_1\times D_2\times\cdots\times D_m}{P}[L_S(h)=0] &= \prod_{i=1}^{m} \underset{x_i\sim D_i}{P}[h(x_i)=f(x_i)]\\ &= \prod_{i=1}^{m} \underset{x\sim D_i}{P}[h(x)=f(x)]\\ &= \left[\left(\prod_{i=1}^{m} \underset{x\sim D_i}{P}[h(x)=f(x)]\right)^{\frac{1}{m}}\right]^{m}\\ &= \text{geometric mean } = (a_1 \cdot a_2 \cdot \cdots \cdot a_m)^{\frac{1}{m}}\\ &\quad \text{where } a_i = \underset{x\sim D_i}{P}[h(x)=f(x)] = \begin{array}{c}\text{probability that } h \text{ correctly}\\ \text{labels a point } x\sim D_i\end{array}\\ &\leq \text{arithmetic mean} = \left(\frac{a_1+a_2+\cdots+a_m}{m}\right)\\ &\leq \left[\frac{1}{m}\sum_{i=1}^{m} \underset{x\sim D_i}{P}[h(x)=f(x)]\right]^{m} < (1-\epsilon)^m \leq e^{-\epsilon m}\end{aligned}$$

There are at most $|\mathcal{H}|$ number of h hypotheses. So, we observe that

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