Assignment 2

Ana-Cristina Rogoz

June 21, 2020

Problem 1 Consider $\mathcal{H} = \{h_{\theta_1} : \mathbb{R} \to \{0,1\}, h_{\theta_1}(x) = \mathbf{1}_{[x \geq \theta_1]}(x) = \mathbf{1}_{[\theta_1, \infty)}(x), \theta_1 \in \mathbb{R}\} \cup \{h_{\theta_2} : \mathbb{R} \to \{0,1\}, h_{\theta_2}(x) = \mathbf{1}_{[x < \theta_2]}(x) = \mathbf{1}_{(-\infty, \theta_2)}(x), \theta_2 \in \mathbb{R}\}$.

a) Compute the shattering coefficient $\tau_{\mathcal{H}}(m)$ of the growth function for m ≥ 0.

b) Compare your result with the general upper bound for the growth functions.

c) Does there exist a hypothesis class $\mathcal{H}$ for which $\tau_{\mathcal{H}}(m)$ is equal to the general upper bound (over $\mathbb{R}$ or another domain X)? If your answer is yes please provide an example, if your answer is no please provide a justification.

Solution a) In the previous homework we've shown that the VC-dimension for the following problem is 2 (since we can obtain all the labels for a chosen subset C, with $|C| = 2$ using one of the hypothesis from either the first or the second subset. Thus, by finding that subset C, $|C| = 2$ that is shattered by $\mathcal{H}$, we know that VCdim($\mathcal{H}$) ≥ 2 (1).
Afterwards, we showed that for any subset C with $|C| = 3, C = x_1, x_2, x_3$, with $x_1 < x_2 < x_3$ there is no hypothesis in $\mathcal{H}$ that can obtain the following label: (0,1,0). If we would be using a hypothesis from the first set, that one would label with 1 all the x values between $[\theta_1, \infty)$, so if x_2 has label one it means that $x_2 \geq \theta_1$, and since $x_3 > x_2$, x_3 can't have label 0.
If we would be using a hypothesis from the second set, that one would label with 1 all the x values between $(-\infty, \theta_2)$, so if x_2 has label one it means that $x_2 < \theta_2$, and since $x_1 < x_2$, x_1 can't have label 0. Therefor, VCdim($\mathcal{H}$) < 3 (2)
From (1) and (2) $\Rightarrow$ $\boxed{VC - dim(\mathcal{H}) = 2}$
In general, we notice that this hypothesis class can output for a subset $|C| = m$, where $C = x_1, x_2, ...x_m, x_1 < x_2 < ... < x_m$ only the following type of labels: either (0, 0, 0, ..., 1,

1, 1) (sequence of zeros followed by sequence of ones – with an element of type h_{θ_1}, from the first set) or (1, 1, 1, ..., 0, 0, 0) (sequence of ones followed by sequence of zeros – with an element of type h_{θ_2}, from the second set). Now, we will count how many possibilities we have for each case:

- (0, 0, 0, ..., 1, 1, 1) – since we have m elements in total the length of the 0 labels can vary between 1 and (m-1):

length of 0 sequence	**length of 1 sequence**
1	(m-1)
2	(m-2)
...	...
(m-1)	1

  $\Rightarrow$ $\boxed{(m-1) functions}$ can be obtained by the first pattern (1)

- (1, 1, 1, ..., 0, 0, 0) – since we have m elements in total the length of the 1 labels can vary between 1 and (m-1):

length of 1 sequence	**length of 0 sequence**
1	(m-1)
2	(m-2)
...	...
(m-1)	1

  $\Rightarrow$ $\boxed{(m-1) functions}$ can be obtained by the second pattern (2)

- we can also have the two trivial cases where all the labels are (0, 0, ..., 0) or (1, 1, ..., 1) (3)

So in the end, the shattering coefficient $\tau_{\mathcal{H}}(m)$ which is the maximum number of different functions from a set C of size m to 0,1 that can be obtained by restricting H to C can be bounded by 2m (from (1), (2), (3)). □

b) From **Sauer's lemma**, we have that for a hypothesis class $\mathcal{H}$, with VC-dim($\mathcal{H}$) $\leq d$ (in this case = 2). Then for all m, we have that: $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} C_m^i$ In this case, $\boxed{\textbf{the general upper bound} \text{ is } C_m^0 + C_m^1 + C_m^2 = \frac{m^2+m+2}{2}}$

$\boxed{\textbf{The shatter coefficient of } \tau_{\mathcal{H}}(m) \text{ found previously is 2m}}$

$2m \leq \frac{m^2+m+2}{2}, \forall m \in \mathbb{N} \iff 0 \leq m^2 - 3m + 2, \forall m \in \mathbb{N}$. We can define the following function $f : \mathbb{N} \rightarrow \mathbb{R}, f(x) = m^2 - 3m + 2,\ f'(x) = 2m - 3 \Rightarrow f'(x) = 0 \iff x = 3/2$

x	0	1	3/2	2	...	∞
f(x)	2	0	-1/4	0	...	∞
f'(x)	-	-	0	+	+	+

Between [0, 1] f is decreasing but its values are still greater that 0 and from 2 forward f(x) is positive, meaning that the general upper bound is greater or equal to the shatter coefficient found for subpoint a). □

c) For this task I'll use the $\mathcal{H}_{thresholds} = \{h_\theta : \mathbb{R} \rightarrow \{0, 1\}, h_\theta(x) = \mathbf{1}_{[x<\theta]}(x) = \mathbf{1}_{(-\infty,\theta)}(x), \theta \in \mathbb{R}\}$ hypothesis class used in Lecture 5 (Slide 28).

From the lecture we know that the VC-dimension for $\mathcal{H}_{thresholds}$ is equal to 1. Moving on, we will compute the shattering coefficient $\tau_{\mathcal{H}}(m)$ for $\mathcal{H}_{thresholds}$ and then compare it to the general upper bound.

Shattering coefficient: For the $\mathcal{H}_{thresholds}$ we notice that the maximum number of different functions from a set C of size m to 0,1 that can be obtained by restricting $\mathcal{H}$ to C has the following pattern (1,1,1, ..., 0,0,0) – a sequence of ones followed by a sequence of zeros. Since our set C has m elements, lets say $x_1, x_2, ..., x_m$, where $x_1 < x_2 < ... < x_m$ we will be able to obtain all the label sets where there is a ones sequence followed by a zeros sequence (because if some x_i has label 0, it means that $x_i > \mathit{threshold}$ so no higher x_j with $x_j > x_i$ can have label 1 again). Thus, since we have m elements, the ones sequence can have a length between 0 and m.

length of 1 sequence	**length of 0 sequence**
0	m
1	(m-1)
...	...
(m)	0

$\Rightarrow \boxed{\tau_{\mathcal{H}}(m) = m + 1}$

General upper bound From Sauer's we get the general upper bound, which is the following $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} C_m^i$. Since in our case VC-dimension is $1 \Rightarrow$ d = 1, we can compute the sum from the right side:

$$\sum_{i=0}^{d} C_m^i = C_m^0 + C_m^1 = \frac{m!}{m! \cdot 0!} + \frac{m!}{(m-1)! \cdot 1!} = 1 + m$$

Thus, we notice that $\tau_{\mathcal{H}}(m)$ is equal to the general upper bound (m+1) $\forall m \in N$, so yes, there is a hypothesis class $\mathcal{H}$ for which $\tau_{\mathcal{H}}(m)$ is equal to the general upper bound. $\square$

Problem 2 Let Σ be a finite alphabet and let $\mathcal{X} = \Sigma^m$ be a sample space of all strings of length m over Σ. Let $\mathcal{H}$ be a hypothesis space over $\mathcal{X}$, where $\mathcal{H} = \{h_w : \Sigma^m \rightarrow \{0, 1\}, w \in \Sigma^*, 0 < |w| \leq m, s.t. h_w(x) = 1$ if w is a substring of x$\}$.

a) Give an upper bound (any upper bound that you can come up) of the VCdimension of $\mathcal{H}$ in terms of $|\Sigma|$ and m.

b) Give an efficient algorithm for finding a hypothesis h_w consistent with a training set in the realizable case. What is the complexity of your algorithm?

Solution a) In order to give an upper bound for the VC-dimension of $\mathcal{H}$, I'll use one of the properties given in Lecture 5(slide 41), the one that states the following: $\boxed{VCdim(\mathcal{H}) \leq log_2|\mathcal{H}|}$. Thus, we will move on and compute the $|\mathcal{H}|$. Since $\mathcal{H}$ includes all h_w, where $0 < |w| \leq m$, I'll count the number of functions for each possible length:

- $|w| = 1 \Rightarrow |\Sigma|$ functions
- $|w| = 2 \Rightarrow |\Sigma|^2$ functions
- ....
- $|w| = m \Rightarrow |\Sigma|^m$ functions

Leading therefore to a total of $|\mathcal{H}| = |\Sigma| + |\Sigma|^2 + ... + |\Sigma|^m$ functions $\leq m \cdot |\Sigma|^m$
$\Rightarrow log_2|\mathcal{H}| \leq log_2(m \cdot |\Sigma|^m) = log_2(m) + m \cdot log_2(|\Sigma|)$
Coming back to our initial inequality: $VCdim(\mathcal{H}) \leq log_2|\mathcal{H}| \leq log_2(m) + m \cdot log_2(|\Sigma|)$
$\Rightarrow \boxed{VCdim(\mathcal{H}) \leq log_2(m) + m \cdot log_2(|\Sigma|)}$ $\square$

b) From the previous subpoint a) we managed to find an upper bound for both the $|\mathcal{H}|$ and its VC-dimension. Because $\mathcal{H}$ is a finite hypothesis class, having at most $log_2(m)+m \cdot log_2(|\Sigma|)$ hypothesis, we can use Corollary 3.2. (from the "Understanding Machine Learning: From Theory to Algorithms" book) which states the following:

Corollary 3.2. Every finite hypothesis class is PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{log(|\mathcal{H}|/\delta)}{\epsilon}$**.**

$\Rightarrow$ We know that the sample complexity of learning a finite class is upper bounded by $\boxed{m_{\mathcal{H}}(\epsilon, \delta) = \frac{log(|\mathcal{H}|/\delta)}{\epsilon}}$ in the realizable case (1)

Also, from Theorem 6.7 (The Fundamental Theorem of Statistical Learning) we know that: $\boxed{\textbf{\textit{Any ERM rule is a successful PAC learner for }} \mathcal{H}}$ (2).

Assuming that the number of training examples is order of $m_{\mathcal{H}}(\epsilon, \delta) = \frac{log(|\mathcal{H}|/\delta)}{\epsilon}$, we will present an ERM rule algorithm over $\mathcal{H}$ which is guaranteed to (ϵ, δ)-learn $\mathcal{H}$:

- **Input data**: $S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $|x_i| = m, \forall i \in \overline{1, n}$ and $y_i \in \{0, 1\}, \forall i \in \overline{1, n}$,
  $possible\_results$ = {all the possible substrings with length between 1 and m} # *in this variable we will store all the hypothesis which are still valid for our training set*
- **Step 1**: $\forall i \in \overline{1, n}$, if $y_i == 1$ go to Step 2
- **Step 2**: Compute the set of all the possible substrings for sample i in Ss_i, where $|Ss_i| \leq \frac{m \cdot (m+1)}{2}$, then go to Step 3
- **Step 3**: $possible\_results = possible\_results \bigcap Ss_i$, then go to Step 1
- **Step 4**: $\forall i \in \overline{1, n}$, if $y_i == 0$ go to Step 5
- **Step 5**: Compute the set of all the possible substrings for sample i in Ss_i, where $|Ss_i| \leq \frac{m \cdot (m+1)}{2}$, then go to Step 6
- **Step 6**: $possible\_results = possible\_results \setminus Ss_i$, then go to Step 4
- **Output data**: Any element of $possible\_results$ set.

Time complexity analysis:

Step 1 – iterates through all n examples $\Rightarrow \mathcal{O}(\frac{log(|\mathcal{H}|/\delta)}{\epsilon})$ steps

Step 2 – computes all possible substrings for a x_i of length m $\Rightarrow \mathcal{O}(\frac{m\cdot(m+1)}{2})$ steps

Step 3 – computes the intersection between *possible_results* and $Ss_i \Rightarrow \mathcal{O}(|\mathcal{H}| \cdot \frac{m\cdot(m+1)}{2})$ steps

Thus, for the first part which only keeps the hypothesis which satisfy the positive examples we have at most $\boxed{\mathcal{O}(\frac{log(|\mathcal{H}|/\delta)}{\epsilon} \cdot (\frac{m\cdot(m+1)}{2} + |\mathcal{H}| \cdot \frac{m\cdot(m+1)}{2}))}$ steps (3)

Step 4 – iterates through all n examples $\Rightarrow \mathcal{O}(\frac{log(|\mathcal{H}|/\delta)}{\epsilon})$ steps

Step 5 – computes all possible substrings for a x_i of length m $\Rightarrow \mathcal{O}(\frac{m\cdot(m+1)}{2})$ steps

Step 6 – computes the difference between the *possible_results* set and Ss_i set $\Rightarrow \mathcal{O}(|\mathcal{H}|)$ steps

Thus, for the second part which eliminates the hypothesis based on the negative examples we have at most $\boxed{\mathcal{O}(\frac{log(|\mathcal{H}|/\delta)}{\epsilon} \cdot (\frac{m\cdot(m+1)}{2} + |\mathcal{H}|))}$ steps (4)

By adding up (3) and (4) $\Rightarrow \mathcal{O}(\frac{log(|\mathcal{H}|/\delta)}{\epsilon} \cdot (2 \cdot \frac{m\cdot(m+1)}{2} + |\mathcal{H}| \cdot (1 + \frac{m\cdot(m+1)}{2})))$ (we can eliminate the constant values) $\Rightarrow \mathcal{O}(\frac{log(|\mathcal{H}|/\delta)}{\epsilon} \cdot (\frac{m\cdot(m+1)}{2} + |\mathcal{H}| \cdot \frac{m\cdot(m+1)}{2}))$

From the previous subpoint a) we know an upper bound for $\mathcal{H} \leq m \cdot |\Sigma|^m$, so we will substitute it with its upper bound.

$\Rightarrow \mathcal{O}(\frac{log(m\cdot|\Sigma|^m/\delta)}{\epsilon} \cdot (\frac{m\cdot(m+1)}{2} + m \cdot |\Sigma|^m \cdot \frac{m\cdot(m+1)}{2})) \approx \mathcal{O}(\frac{1}{\epsilon} \cdot log(m \cdot |\Sigma|^m/\delta) \cdot (\frac{m\cdot(m+1)}{2} + m \cdot |\Sigma|^m \cdot \frac{m\cdot(m+1)}{2})) \approx \underbrace{\boxed{\mathcal{O}(\frac{1}{\epsilon} \cdot (log(m) + m \cdot log(|\Sigma|) - log(\delta)) \cdot (m^3 \cdot |\Sigma|^m))}}_{\text{algorithm time complexity}}$ □

Problem 3 Consider the boosting algorithm described (page 4) in the article "Rapid object detection using a boosted cascade of simple features", P. Viola and M. Jones, CVPR 2001. Consider that the number of positives is equal with the number of negative examples (l = m).

a) Prove that the distribution w_{t+1} obtained at round t + 1 based on the algorithm described in the article is the same with the distribution $D^{(t+1)}$ obtained based on the procedure described in lecture 11 (slides 10-12).

b) Prove that the two final classifiers (the one described in the article and the one described in the lecture) are equivalent.

c) Assume that at each iteration t of AdaBoost, the weak learner returns a hypothesis h_t for which the error ϵ_t satisfies $\epsilon_t \leq 1/2-\gamma, \gamma > 0$. What is the probability that the classifier h_t (selected as the best weak learner at iteration t) will be selected again at iteration t+1? Justify your answer.

Solution a) We will start by analyzing the two update steps, the one from the lecture and the one from the paper. We will consider that we have the following training set S = $\{(x_1, y_1), (x_2, y_2), .., (x_n, y_n)\}$

In the article (labels are 1 and 0):

- **Initial weights**: $w_{1,i} = 1/2 * no\_negative\_examples$ if $y_i = 0$ or $1/2 * no\_positive\_examples$ if $y_i = 1$. Since the number of positive examples and negative examples is equal in our case, $\boxed{w_{1,i} = \frac{1}{n}, \forall i \in \overline{1,n}}$
- **Update rule**: $w_{t+1,i} = w_{t,i} \cdot \beta_t^{1-e_i}$, where $e_i = 0$ if x_i correctly classified and 1 otherwise

$$\beta_t = \frac{\epsilon_t}{1-\epsilon_t} \Rightarrow \boxed{w_{t+1,i} = \begin{cases} w_{t,i} \cdot \frac{\epsilon_t}{1-\epsilon_t} & , x_i \text{ is labeled correctly i.e. } h_t(x_i) = y_i \\ w_{t,i} & , otherwise \end{cases}}$$

In the lecture (labels are -1 and 1):

- **Initial weights distribution**: $\boxed{D^{(1)}(i) = \frac{1}{n}, \forall i \in \overline{1,n}}$
- **Update rule**: $D^{(t+1)}(i) = \frac{D^{(t)}(i) \cdot e^{-w_t \cdot h_t(x_i) \cdot y_i}}{\sum_{j=1}^{n} D^{(t)}(j) \cdot e^{-w_t \cdot h_t(x_j) \cdot y_j}}$

  We know that $w_t = \frac{1}{2} \cdot ln(\frac{1}{\epsilon_t} - 1)$ and $\epsilon_t = \sum_{i=1}^{n} D^{(t)}(i) \times 1_{[h_t(x_i) \neq y_i]}$

  So after replacing w_t with its value and making some computations

$$\Rightarrow \boxed{D^{(t+1)}(i) = \begin{cases} \frac{D^{(t)}(i) \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}}{\sum_{j=1}^{n} D^{(t)}(j) \cdot e^{-w_t \cdot h_t(x_j) \cdot y_j}} & , x_i \text{ is labeled correctly i.e. } h_t(x_i) = y_i \\ \frac{D^{(t)}(i) \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}}{\sum_{j=1}^{n} D^{(t)}(j) \cdot e^{-w_t \cdot h_t(x_j) \cdot y_j}} & , otherwise \end{cases}}$$

Moving on, I'll reduce the denominator from the update step from the lecture to a simpler form and then show that $w_{t+1,i}$ and $D^{(t+1)}(i)$ are proportional (since $w_{t+1,i}$ will be normalized right at the start of the next step t+1, not at the time of the assignment)

$\sum_{j=1}^{n} D^{(t)}(j) \cdot e^{-w_t \cdot h_t(x_j) \cdot y_j} = \sum_{j=1}^{n} D^{(t)}(j) \cdot e^{-\frac{1}{2} ln(\frac{1}{\epsilon_t}-1) \cdot h_t(x_j) \cdot y_j}$ (we will split the sum in two parts, the one that sums the correctly classified examples and the one that sums the incorrect examples) $= \underbrace{\sum_i D^{(t)}(i) \cdot e^{-\frac{1}{2} ln(\frac{1}{\epsilon_t}-1) \cdot (-1)}}_{\text{incorrectly classified examples}} + \underbrace{\sum_j D^{(t)}(j) \cdot e^{-\frac{1}{2} ln(\frac{1}{\epsilon_t}-1) \cdot 1}}_{\text{correctly classified examples}}$

$$= \underbrace{\sum_i D^{(t)}(i) \cdot (\frac{1}{\epsilon_t} - 1)^{\frac{1}{2}}}_{\text{incorrectly classified examples}} + \underbrace{\sum_j D^{(t)}(j) \cdot (\frac{1}{\epsilon_t} - 1)^{\frac{-1}{2}}}_{\text{correctly classified examples}}$$

$$= (\tfrac{1}{\epsilon_t} - 1)^{\frac{1}{2}} \cdot \underbrace{\sum_i D^{(t)}(i)}_{\text{incorrectly classified examples}} + (\tfrac{1}{\epsilon_t} - 1)^{\frac{-1}{2}} \cdot \underbrace{\sum_j D^{(t)}(j)}_{\text{correctly classified examples}}$$

We notice that the sum of incorrectly classified examples is actually ϵ_t, so we can replace the first sum with it. Also, because $D^{(t)}$ is a distribution it means that the sum of all i's is 1, thus the sum of $D^{(t)}$ for correctly classified examples is $1 - \epsilon_t$

$\Rightarrow (\frac{1}{\epsilon_t} - 1)^{\frac{1}{2}} \cdot \epsilon_t + (\frac{1}{\epsilon_t} - 1)^{\frac{-1}{2}} \cdot (1 - \epsilon_t) = (\frac{1}{\epsilon_t} - 1)^{\frac{1}{2}} \cdot \epsilon_t + (\frac{1}{\epsilon_t} - 1)^{\frac{-1}{2}} - (\frac{1}{\epsilon_t} - 1)^{\frac{-1}{2}} \cdot \epsilon_t$

$= \epsilon_t[(\frac{1}{\epsilon_t} - 1)^{\frac{1}{2}}) - (\frac{1}{\epsilon_t} - 1)^{\frac{-1}{2}})] + (\frac{1}{\epsilon_t} - 1)^{\frac{-1}{2}} = \epsilon_t(\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} - \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}) + \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}$

$= \epsilon_t(\frac{1-\epsilon_t}{\sqrt{\epsilon_t} \cdot \sqrt{1-\epsilon_t}} - \frac{\epsilon_t}{\sqrt{\epsilon_t} \cdot \sqrt{1-\epsilon_t}}) + \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} = \epsilon_t(\frac{1-2\epsilon_t}{\sqrt{\epsilon_t} \cdot \sqrt{1-\epsilon_t}}) + \frac{\epsilon_t}{\sqrt{1-\epsilon_t} \cdot \sqrt{\epsilon_t}}$

$= \frac{2 \cdot \epsilon_t \cdot (1-\epsilon_t)}{\sqrt{1-\epsilon_t} \cdot \sqrt{\epsilon_t}} = \boxed{2 \cdot \sqrt{\epsilon_t} \cdot \sqrt{1 - \epsilon_t}}$

Coming back to the update rule from the lecture, we have:

$$D^{(t+1)}(i) = \begin{cases} \frac{D^{(t)}(i) \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}}{2 \cdot \sqrt{\epsilon_t} \cdot \sqrt{1-\epsilon_t}} & , x_i \text{ is labeled correctly i.e. } h_t(x_i) = y_i \\ \frac{D^{(t)}(i) \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}}{2 \cdot \sqrt{\epsilon_t} \cdot \sqrt{1-\epsilon_t}} & , otherwise \end{cases}$$

In order to show that the two distributions are the same, I'll prove that for both the correctly classified examples and the incorrectly classified ones, the $D^{(t+1)}(i)$ and $w_{t+1,i}$ are proportional (because we compare a normalized D distribution with an un-normalized w distribution – since the normalization in the paper is done at the beginning of the next step).

We will use mathematical induction in order to prove that if $D^{(t)}(i) = w_{t,i} \Rightarrow D^{(t+1)}(i) = w_{t+1,i}$.

The induction hypothesis $D^{(t)}(i) = w_{t,i}$:

- x_i **is wrongly classified**

According to the lecture: $\boxed{D^{(t+1)}(i) = D^{(t)}(i) \cdot \dfrac{1}{2 \cdot \epsilon_t}}$

According to our hypothesis $\boxed{D^{(t)}(i) = w_{t,i}}$

According to the article: $\boxed{w_{t+1,i} = w_{t,i}}$

$\Rightarrow \boxed{D^{(t+1)}(i) = w_{t+1,i} \cdot \dfrac{1}{2 \cdot \epsilon_t}}$

- x_i **is correctly classified**

According to the lecture: $\boxed{D^{(t+1)}(i) = D^{(t)}(i) \cdot \dfrac{1}{2 \cdot (1-\epsilon_t)}}$

According to our hypothesis $\boxed{D^{(t)}(i) = w_{t,i}}$

According to the article: $w_{t+1,i} = w_{t,i} \cdot \frac{\epsilon_t}{1-\epsilon_t} \Rightarrow \boxed{w_{t,i} = w_{t+1,i} \cdot \dfrac{1-\epsilon_t}{\epsilon_t}}$

$\Rightarrow \boxed{D^{(t+1)}(i) = w_{t+1,i} \cdot \dfrac{1}{2 \cdot \epsilon_t}}$

Thus, since in both cases we found out that $D^{(t+1)}(i) = w_{t+1,i} \cdot \frac{1}{2 \cdot \epsilon_t}$ (the difference between the two of them comes in because w_t is normalized at the beginning of the t+1 step and at the end of step t its values are not the ones of a distribution – they do not sum up to 1), we can draw the conclusion that these two distributions are equivalent. $\square$

b) **In the article** (labels 0 and 1), the final classifier is the following:

$$h(x) = \begin{cases} 1 & , \sum_{t=1}^{T} \alpha_t \cdot h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & , otherwise \end{cases} \text{, where } \alpha_t = ln(\tfrac{1}{\beta_t}) \text{ and } \beta_t = (\tfrac{\epsilon_t}{1-\epsilon_t})$$

By substitution for β_t and $\alpha_t \Rightarrow h(x) = \begin{cases} 1 & , \sum_{t=1}^{T} ln(\frac{1-\epsilon_t}{\epsilon_t}) \cdot h_t(x) - \frac{1}{2} \sum_{t=1}^{T} ln(\frac{1-\epsilon_t}{\epsilon_t}) \geq 0 \\ 0 & , otherwise \end{cases}$

$$\Rightarrow \boxed{h(x) = \begin{cases} 1 & , \sum_{t=1}^{T} ln(\frac{1-\epsilon_t}{\epsilon_t}) \cdot (h_t(x) - \frac{1}{2}) \geq 0 \\ 0 & , otherwise \end{cases}}$$

In the lecture (labels -1, 1), the final classifier is the following:

$$h(x) = sign(\textstyle\sum_{t=1}^{T} w_t \cdot h_t(x)) = \begin{cases} 1 & , \sum_{t=1}^{T} w_t \cdot h_t(x) \geq 0 \\ -1 & , otherwise \end{cases} \text{, where } w_t = \tfrac{1}{2} \cdot ln(\tfrac{1}{\epsilon_t} - 1)$$

If we substitute w_t in the final classifier we get the following:

$$\boxed{h(x) = \begin{cases} 1 & , \sum_{t=1}^{T} ln(\frac{1-\epsilon_t}{\epsilon_t}) \cdot \frac{1}{2} \cdot h_t(x) \geq 0 \\ -1 & , otherwise \end{cases}}$$

From the previous subpoint a), we know that the w_t distribution after normalization from the article and the $D^{(t)}$ distribution from the lecture are the same. Since they are the same, and they use the same training set they will produce the same ϵ_t minimum error for the same classifier h_t (by summing up the weights for the examples which are wrongly classified on the t-th feature).

Moving on, I'll name the final classifier from the article $h_1(x)$ and the one from the lecture $h_2(x)$. Now I want to show that the two sums from the two classifiers give the same results:

- $h_1(x) : \sum_{t=1}^{T} ln(\frac{1-\epsilon_t}{\epsilon_t}) \cdot (h_t(x) - \frac{1}{2})$
- $h_2(x) : \sum_{t=1}^{T} ln(\frac{1-\epsilon_t}{\epsilon_t}) \cdot \frac{1}{2} \cdot h_t(x)$

We know that the ln-parenthesis is the same in both sums, so we want to show that:

- $(h_t(x) - \frac{1}{2})$ from the article (which has labels 0, 1)
- $\frac{1}{2} \cdot h_t(x)$ from the lecture (which has labels -1, 1)

give the same results.

$h_t(x)$ label (article)	$(h_t(x) - \frac{1}{2})$ article result	$h_t(x)$ label (lecture)	$\frac{1}{2} \cdot h_t(x)$ lecture result
0	-0.5	-1	-0.5
1	0.5	1	0.5

$$\Rightarrow \boxed{(h_t(x) - \frac{1}{2})(\text{from the article, labels 0,1}) = \frac{1}{2} \cdot h_t(x)(\text{from the lecture, labels -1, 1})}$$

Thus, since we managed to bring the both final classifiers to a similar form and then to show that they give equal results for all cases, we can conclude that the two final classifiers are equivalent. $\square$

c) In order to see how the error will change from step t to (t+1), I'll compute it using the formula used in the lecture: $\epsilon_t = \sum_{i=1}^{n} D^{(t)}(i) \times 1_{[h_t(x_i) \neq y_i]}$.

Let's say that at step t, we found out h_t, the hypothesis with the smallest ϵ_t.

During step (t+1), the previous classifier h_t will have its error $= \epsilon_{t+1} = \sum_{i=1}^{n} D^{(t+1)}(i) \times 1_{[h_t(x_i) \neq y_i]} = \sum_{i=1}^{n} \frac{1}{2 \cdot \epsilon_t} \cdot D^{(t)}(i) \times 1_{[h_t(x_i) \neq y_i]} = \frac{1}{2 \cdot \epsilon_t} \cdot \sum_{i=1}^{n} D^{(t)}(i) \times 1_{[h_t(x_i) \neq y_i]} = \frac{1}{2 \cdot \epsilon_t} \cdot \epsilon_t = \frac{1}{2}$

Assuming the fact that at step t, there was another classifier let's call it h'_t with $\epsilon_t < \epsilon'_t \leq \frac{1}{2}$ and its complementary $h'_{t,comp}$ with error $1 - \epsilon'_t$ we will compute their error on the step (t+1):

- error for h'_t at (t+1) $= \epsilon'_{t+1} = \sum_{i=1}^{n} D^{(t+1)}(i) \times 1_{[h'_t(x_i) \neq y_i]} \geq \sum_{i=1}^{n} \frac{1}{2 \cdot \epsilon_t} \cdot D^{(t)}(i) \times 1_{[h'_t(x_i) \neq y_i]} = \frac{1}{2 \cdot \epsilon_t} \cdot \sum_{i=1}^{n} D^{(t)}(i) \times 1_{[h'_t(x_i) \neq y_i]} = \frac{1}{2 \cdot \epsilon_t} \cdot \epsilon'_t$ (we know that $\epsilon_t < \epsilon'_t$, so $\frac{\epsilon'_t}{\epsilon_t} \geq 1 \Rightarrow \epsilon'_{t+1} \geq \frac{1}{2}$

- error for $h'_{t,comp}$ at (t+1) $= \epsilon'_{t+1,comp} = \sum_{i=1}^{n} D^{(t+1)}(i) \times 1_{[h'_{t,comp}(x_i) \neq y_i]} \leq \sum_{i=1}^{n} \frac{1}{2 \cdot (1 - \epsilon_t)} \cdot D^{(t)}(i) \times 1_{[h'_{t,comp}(x_i) \neq y_i]} = \frac{1}{2 \cdot (1 - \epsilon_t)} \cdot \sum_{i=1}^{n} D^{(t)}(i) \times 1_{[h'_{t,comp}(x_i) \neq y_i]} = \frac{1}{2 \cdot (1 - \epsilon_t)} \cdot (1 - \epsilon'_t)$ (we know that $\epsilon_t < \epsilon'_t \Rightarrow (1 - \epsilon_t > 1 - \epsilon'_t)$, so $\frac{1 - \epsilon'_t}{1 - \epsilon_t} < 1 \Rightarrow \epsilon'_{t+1} < \epsilon_{t+1} = \frac{1}{2}$

Thus, if there exists another hypothesis at step t with $\epsilon < 1/2$, h_t will always be replaced at step t+1.