Advanced Machine Learning Seminar 6

Exercise 1 Fix $\epsilon \in \left(0, \frac{1}{2}\right)$. Let the training sample be denoted by m points in the plane with $\frac{m}{4}$ negative points all at coordinate $(+1, +1)$, another $\frac{m}{4}$ negative points all at coordinate $(-1, -1)$, $\frac{m(1+\epsilon)}{4}$ positive points all at coordinate $(-1, +1)$, $\frac{m(1-\epsilon)}{4}$ positive points all at coordinate $(+1, -1)$.

a. Describe the behavior of AdaBoost when run on this sample using boosting stumps for the first two rounds.

b. What is the error of the optimal classifier chosen at round 1 in the second round?

AdaBoost

- construct distribution $\mathbf{D}^{(t)}$ on $\{1, \dots, m\}$:
  - $\mathbf{D}^{(t)}(i) = 1/m$
  - given $\mathbf{D}^{(t)}$ and h_t: $D^{(t+1)}(i) = \dfrac{D^{(t)}(i) \times e^{-w_t h_t(x_i) y_i}}{Z_{t+1}}$

where Z_{t+1} normalization factor ($\mathbf{D}^{(t+1)}$ is a distribution): $Z_{t+1} = \sum_{i=1}^{m} D^{(t)}(i) \times e^{-w_t h_t(x_i) y_i}$

w_t is a weight: $w_t = \frac{1}{2} \ln\left(\frac{1}{\epsilon_t} - 1\right) > 0$ as the error $\epsilon_t < 0.5$

ϵ_t is the error of h_t on $\mathbf{D}^{(t)}$: $\epsilon_t = \Pr_{i \sim D^{(t)}}[h_t(x_i) \neq y_i] = \displaystyle\sum_{i=1}^{m} D^{(t)}(i) \times \mathbb{1}_{[h_t(x_i) \neq y_i]}$

If example $\mathbf{x}_i$ is correctly classified, then $h(\mathbf{x}_i) = \mathrm{y}_i$, so at the next iteration $t+1$ its importance (probability distribution) will be decreased to:

$$D^{(t+1)}(i) = \frac{D^{(t)}(i) \times e^{-w_t}}{Z_{t+1}} = \frac{D^{(t)}(i) \times e^{-\frac{1}{2}\ln\left(\frac{1}{\epsilon_t} - 1\right)}}{Z_{t+1}} = \frac{D^{(t)}(i) \times \left(\frac{1}{\epsilon_t} - 1\right)^{-\frac{1}{2}}}{Z_{t+1}} = \frac{D^{(t)}(i) \times \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}}{Z_{t+1}}$$

If example x_i is misclassified, then $h(\mathrm{x}_i) \neq \mathrm{y}_i$, so at the next iteration $t+1$ its importance (probability distribution) will be increased to:

$$D^{(t+1)}(i) = \frac{D^{(t)}(i) \times e^{w_t}}{Z_{t+1}} = \frac{D^{(t)}(i) \times e^{\frac{1}{2}\ln\left(\frac{1}{\epsilon_t} - 1\right)}}{Z_{t+1}} = \frac{D^{(t)}(i) \times \left(\frac{1}{\epsilon_t} - 1\right)^{\frac{1}{2}}}{Z_{t+1}} = \frac{D^{(t)}(i) \times \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}}{Z_{t+1}}$$

Solution.

$$\frac{m(1+\epsilon)}{4} + \frac{m(1-\epsilon)}{4} = \frac{m}{2} \text{ points with + label}$$
$$\frac{m}{4} + \frac{m}{4} = \frac{m}{2} \text{ points with } - \text{ label}$$

The probability distribution of the training point $(-1, 1)$ with label $+$ is $\dfrac{\frac{m(1+\epsilon)}{4}}{m} = \dfrac{1+\epsilon}{4}$. For point $(1, -1)$, we obtain $\dfrac{1-\epsilon}{4}$, for points $(1, 1)$ and $(-1, -1)$ with label $-$ we obtain $\frac{1}{4}$.

The initial problem with m points in the training sample is similar with the problem with 4 points with the corresponding probabilities.

$$S = \left\{ ((\underset{\substack{\downarrow \\ point}}{-1, +1}), \underset{\substack{\updownarrow \\ label}}{+1}), ((+1, -1), +1), ((+1, +1), -1), ((-1, -1), -1) \right\}$$

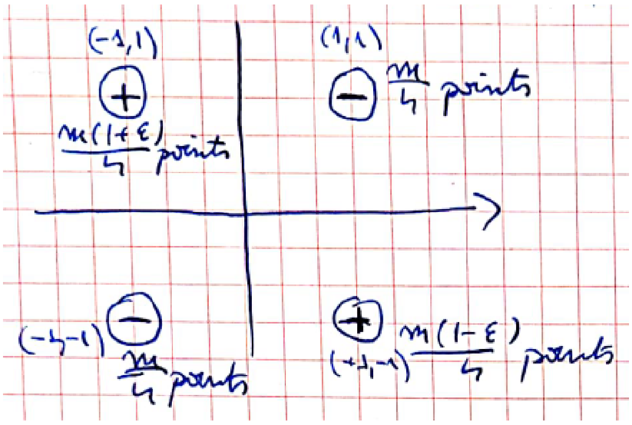

Figure 1: Representation of the m points in the plane.

$$D^{(1)}\colon \begin{pmatrix} (-1,1) & (1,-1) & (1,1) & (-1,-1) \\ \dfrac{1+\epsilon}{4} & \dfrac{1-\epsilon}{4} & \dfrac{1}{4} & \dfrac{1}{4} \end{pmatrix}$$

Base hypothesis class = decision stumps in $\mathbb{R}^2$.

$$\mathcal{H}^2_{\mathrm{DS}} = \left\{ h_{i,\theta,b}\colon \mathbb{R}^2 \to \{-1,1\},\, h_{i,\theta,b}(x_1,x_2) = \operatorname{sign}(\theta - x_i)\cdot b \quad \begin{array}{l} 1 \le i \le 2 \\ \theta \in \mathbb{R} \\ b \in \{+1,-1\} \end{array} \right\}$$

= pick a coordinate i (1 or 2), project the input $x = (x_1, x_2)$ on the i-th coordinate and obtain x_i if $x_i \le \theta$, label the example x_i with label b, else with label $-b$

For our problem, we can see that we can take a set of representation thresholds θ to be $\theta = \{-2, 0, 2\}$. So we have at most 12 base classifiers: $h_{1,-2,1}; h_{1,-2,-1}; \ldots; h_{2,2,-1}$

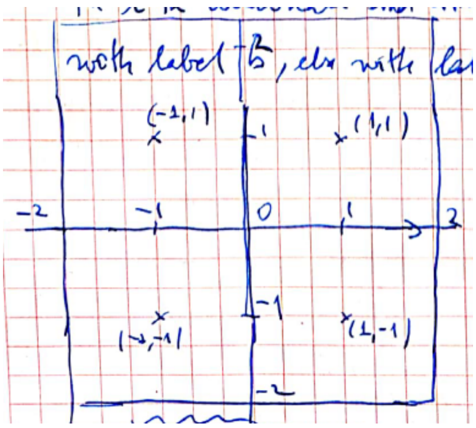

Figure 2: There are 12 base classifiers decision stumps in the plae for our problem: $h_{1,-2,1}; h_{1,-2,-1}; \ldots; h_{2,2,-1}$).

$h_{1,-2,+1} \to$ project on x_1, compare to -2, all points < -2 get label $+1$, all other get label -1

$h_{1,-2,-1} \to$ project on x_1, compare to -2, all points < -2 get label -1, all other get label $+1$

$h_{1,+2,+1} \to$ project on x_1, compare to $+2$, all points $< +2$ get label $+1$, all other get label -1

So we see that on our training set $h_{1,-2,-1}$ and $h_{1,+2,+1}$ will have the same behavior (all points will receive label $+1$).

If we analyze the behavior of all 12 base classifiers (decision stumps in $\mathbb{R}^2$), we will see that in the end there are only 6 unique base classifiers.

$$
\underset{h^1}{\begin{array}{c|c} + & + \\ \hline + & + \end{array}}
\qquad
\underset{h^2}{\begin{array}{c|c} - & - \\ \hline - & - \end{array}}
\qquad
\underset{h^3}{\begin{array}{c|c} + & - \\ \hline + & - \end{array}}
\qquad
\underset{h^4}{\begin{array}{c|c} - & + \\ \hline - & + \end{array}}
\qquad
\underset{h^5}{\begin{array}{c|c} - & - \\ \hline + & + \end{array}}
\qquad
\underset{h^6}{\begin{array}{c|c} + & + \\ \hline - & - \end{array}}
$$

So we have $B = \{h^1, h^2, h^3, h^4, h^5, h^6\}$.

<u>Round 1</u>

- distribution $D^{(1)}$: $\begin{pmatrix} (-1,1) & (1,-1) & (1,1) & (-1,-1) \\ \dfrac{1+\epsilon}{4} & \dfrac{1-\epsilon}{4} & \dfrac{1}{4} & \dfrac{1}{4} \end{pmatrix}$

- select the best classifier from $\mathcal{H}$, the one with minimum empirical risk

$$
\begin{aligned}
L_{D^{(1)}}(h^1) &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\
L_{D^{(1)}}(h^2) &= \frac{1+\epsilon}{4} + \frac{1-\epsilon}{4} = \frac{1}{2} \\
L_{D^{(1)}}(h^3) &= \frac{1}{4} + \frac{1-\epsilon}{4} = \frac{1}{2} - \frac{\epsilon}{4} \\
L_{D^{(1)}}(h^4) &= \frac{1+\epsilon}{4} + \frac{1}{4} = \frac{1}{2} + \frac{\epsilon}{4} \\
L_{D^{(1)}}(h^5) &= \frac{1+\epsilon}{4} + \frac{1}{4} = \frac{1}{2} + \frac{\epsilon}{4} \\
L_{D^{(1)}}(h^6) &= \frac{1}{4} + \frac{1-\epsilon}{4} = \frac{1}{2} - \frac{\epsilon}{4}
\end{aligned}
$$

So, the minimum achievable error is $\frac{1}{2} - \frac{\epsilon}{4}$ and it is attained by base classifiers h^3 and h^6. Let's choose h^3 as our weak classifier: $h^3 = h_{1,0,+1}$.

So, for $t = 1$ (round 1) we have $h_t = h^3 = h_{1,0,+1}$.

The error of the base classifier is $\epsilon_1 = \frac{1}{2} - \frac{\epsilon}{4}$.

$$
w_1 = \frac{1}{2}\ln\left(\frac{1}{\epsilon_1} - 1\right) = \frac{1}{2}\left(\ln\left(\frac{4}{2-\epsilon} - 1\right)\right) = \ln\left(\frac{2+\epsilon}{2-\epsilon}\right)^{\frac{1}{2}} = \ln\sqrt{\frac{2+\epsilon}{2-\epsilon}}
$$

Based on $D^{(1)}$ we will build $D^{(2)}$. Examples correctly classified at round 1 will have now the weight decreased, examples misclassified at round 1 will have their weight increased.

$$
\begin{aligned}
D^{(2)}((-1,+1)) &= \frac{1}{Z_2} D^{(1)}((-1,+1)) \cdot \sqrt{\frac{\epsilon_1}{1-\epsilon_1}} = \frac{1}{Z_2} \cdot \left(\frac{1+\epsilon}{4}\right) \cdot \sqrt{\frac{2-\epsilon}{2+\epsilon}} \quad \searrow \\
D^{(2)}((+1,-1)) &= \frac{1}{Z_2} \cdot \left(\frac{1-\epsilon}{4}\right) \cdot \sqrt{\frac{2+\epsilon}{2-\epsilon}} \quad \nearrow \\
D^{(2)}((+1,+1)) &= \frac{1}{Z_2} \cdot \frac{1}{4} \cdot \sqrt{\frac{2-\epsilon}{2+\epsilon}} \quad \searrow \\
D^{(2)}((-1,-1)) &= \frac{1}{Z_2} \cdot \frac{1}{4} \cdot \sqrt{\frac{2+\epsilon}{2-\epsilon}} \quad \nearrow
\end{aligned}
$$

We can find the value of Z_2 such that $D^{(2)}$ is a probability distribution, meaning that the sum of probability mass should be equal to 1.

$$
D^{(2)}((-1,+1)) + D^{(2)}((+1,-1)) + D^{(2)}((+1,+1)) + D^{(2)}((-1,-1)) = 1
$$

$$
\begin{aligned}
\Rightarrow Z_2 &= \frac{1+\epsilon}{4} \cdot \sqrt{\frac{2-\epsilon}{2+\epsilon}} + \frac{1-\epsilon}{4} \cdot \sqrt{\frac{2+\epsilon}{2-\epsilon}} + \frac{1}{4} \cdot \sqrt{\frac{2-\epsilon}{2-\epsilon}} + \frac{1}{4} \cdot \sqrt{\frac{2+\epsilon}{2+\epsilon}} \\
&= \frac{1}{4} \cdot \sqrt{\frac{2-\epsilon}{2+\epsilon}} \cdot \left((1+\epsilon) + (1-\epsilon) \cdot \frac{2+\epsilon}{2-\epsilon} + 1 + \frac{2+\epsilon}{2-\epsilon} \right) \\
&= \frac{1}{4} \cdot \sqrt{\frac{2-\epsilon}{2+\epsilon}} \cdot \frac{(1+\epsilon)\cdot(2-\epsilon) + (1-\epsilon)\cdot(2+\epsilon) + (2-\epsilon) + 2 + \epsilon}{2-\epsilon} \\
&= \frac{1}{4} \cdot \sqrt{\frac{2-\epsilon}{2+\epsilon}} \cdot \frac{2+\epsilon-\epsilon^2+2-\epsilon-\epsilon^2+2-\epsilon+2+\epsilon}{2-\epsilon} \\
&= \frac{1}{4} \cdot \sqrt{\frac{2-\epsilon}{2+\epsilon}} \cdot \frac{8-2\epsilon^2}{2-\epsilon} = \frac{1}{4} \cdot \sqrt{\frac{2-\epsilon}{2+\epsilon}} \cdot \frac{2(2-\epsilon)(2+\epsilon)}{2-\epsilon} \\
&= \frac{1}{2} \cdot \sqrt{(2-\epsilon)(2+\epsilon)}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow D^{(2)}((-1,+1)) &= \frac{1+\epsilon}{2(2+\epsilon)} \\
D^{(2)}((+1,-1)) &= \frac{1-\epsilon}{2(2-\epsilon)} \\
D^{(2)}((+1,+1)) &= \frac{1}{2(2+\epsilon)} \\
D^{(2)}((-1,-1)) &= \frac{1}{2(2-\epsilon)}
\end{aligned}
$$

What is the error of the base classifier $h^3 = h_{1,0,+1}$ selected at round 1 on $D^{(2)}$?

$$
\text{Loss}(h^3) = \frac{1}{2(2-\epsilon)} + \frac{1-\epsilon}{2(2-\epsilon)} = \frac{2-\epsilon}{2(2-\epsilon)} = \frac{1}{2}
$$

<u>Round 2</u>

- distribution $D^{(2)}$: $\begin{pmatrix} (-1,1) & (1,-1) & (1,1) & (-1,-1) \\ \frac{1+\epsilon}{2(2+\epsilon)} & \frac{1-\epsilon}{2(2-\epsilon)} & \frac{1}{2(2+\epsilon)} & \frac{1}{2(2-\epsilon)} \end{pmatrix}$
- select the best classifier from $\mathcal{H}$, the one with minimum empirical risk

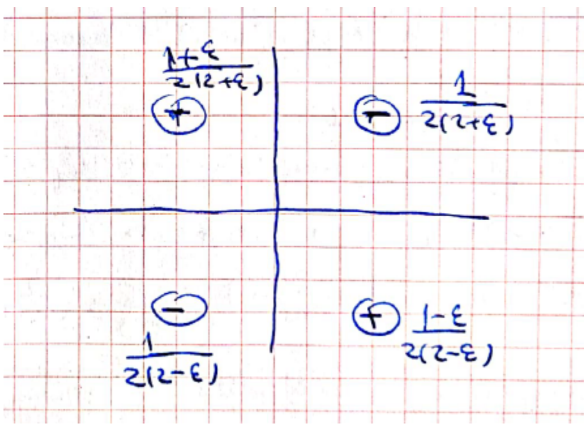

Figure 3: Updated distribution $D^{(2)}$ of samples after round 1 of AdaBoost.

$$
\begin{aligned}
L_{D^{(2)}}(h^1) &= \frac{1}{2(2-\epsilon)} + \frac{1}{2(2+\epsilon)} = \frac{2+\epsilon+2-\epsilon}{2(2-\epsilon)(2+\epsilon)} = \frac{2}{(2-\epsilon)(2+\epsilon)} = \frac{4}{2(2-\epsilon)(2+\epsilon)} \\
L_{D^{(2)}}(h^2) &= \frac{1+\epsilon}{2(2+\epsilon)} + \frac{1-\epsilon}{2(2-\epsilon)} = \frac{(1+\epsilon)\cdot(2-\epsilon)+(1-\epsilon)\cdot(2+\epsilon)}{2(2-\epsilon)(2+\epsilon)} = \frac{4-2\epsilon^2}{2(2-\epsilon)(2+\epsilon)} \\
L_{D^{(2)}}(h^3) &= \frac{1}{2} = \frac{4-\epsilon^2}{2(2-\epsilon)(2+\epsilon)} \\
L_{D^{(2)}}(h^4) &= \frac{1}{2} = \frac{4-\epsilon^2}{2(2-\epsilon)(2+\epsilon)} \\
L_{D^{(2)}}(h^5) &= \frac{1+\epsilon}{2(2+\epsilon)} + \frac{1}{2(2-\epsilon)} = \frac{(1+\epsilon)\cdot(2-\epsilon)+2+\epsilon}{2(2+\epsilon)(2-\epsilon)} = \frac{4+2\epsilon-\epsilon^2}{2(2+\epsilon)(2-\epsilon)} \\
L_{D^{(2)}}(h^6) &= \frac{1}{2(2+\epsilon)} + \frac{1-\epsilon}{2(2-\epsilon)} = \frac{(2-\epsilon)+(1-\epsilon)\cdot(2+\epsilon)}{2(2+\epsilon)(2-\epsilon)} = \frac{4-2\epsilon-\epsilon^2}{2(2+\epsilon)(2-\epsilon)}
\end{aligned}
$$

The smallest error is attained by h^6. This is the base classifier selected at the current round. So, for $t = 2$ (round 2) we have $h_2 = h^6 = h_{2,0,-1}$.

$$
\begin{aligned}
\epsilon_2 &= \frac{4-2\epsilon-\epsilon^2}{2(2+\epsilon)(2-\epsilon)} \\
w_2 &= \frac{1}{2}\ln\left(\frac{1}{\epsilon_2}-1\right) = \frac{1}{2}\ln\left(\frac{4-\epsilon^2+2\epsilon}{4-\epsilon^2-2\epsilon}\right)
\end{aligned}
$$

□