Homework No. 5

ALEXANDER SHENDER 328626114, SNIR HORDAN 205689581

Question 1.

1. Prove that when running AdaBoost, the distribution is updated such that the error of the chosen weak classifier h_t, w.r.t the updated distribution $D_i^{(t+1)}$, is exactly $\frac{1}{2}$.

   That is, prove that $\sum_i D_i^{(t+1)} \cdot \mathbf{1}_{h_t(x_i)\neq y_i} = \frac{1}{2}$.

   Hint: You can fill the missing steps in the following derivation:

   $$\sum_i D_i^{(t+1)} \cdot \mathbf{1}_{h_t(x_i)\neq y_i} = \cdots = \frac{\epsilon_t}{\epsilon_t+(1-\epsilon_t)\exp\{-2w_t\}} = \cdots = \frac{1}{2}.$$

We start by indeed from writing the expression for the updated error value with the updated data weights (distribution) for the last chosen weak classifier:

$$\mathrm{E}_{t+1} = \sum_i D_i^{t+1} \cdot 1_{h_t(x_i)\neq y_i}$$

By using:

$$D_i^{t+1} = D_i^t \cdot \frac{\exp\left(-w_t y_i h_t(x_i)\right)}{\sum_j D_j^t \exp\left(-w_t y_j h_t(x_j)\right)} = D_i^t \cdot \frac{\exp\left(-w_t y_i h_t(x_i)\right)}{Z_t}$$

Putting back:

$$\mathrm{E}_{t+1} = \sum_i D_i^{t+1} \cdot 1_{h_t(x_i)\neq y_i} = \sum_i \frac{D_i^t \exp\left(-w_t y_i h_t(x_i)\right)}{Z_t} \cdot 1_{h_t(x_i)\neq y_i}$$

Z_t is a normalization factor, so we can put it outside of the sum:

$$\mathrm{E}_{t+1} = \frac{\sum_i D_i^t \exp\left(-w_t y_i h_t(x_i)\right)}{Z_t} \cdot 1_{h_t(x_i)\neq y_i}$$

We divide into 2 cases:

$$\begin{cases} E : h_t(x_i) = y_i \\ C : h_t(x_i) \neq y_i \end{cases}$$

For each case,

$$\begin{cases} E : 1_{h_t(x_i)\neq y_i} = 0 \,;\; y_i h_t(x_i) = 1 \\ C : 1_{h_t(x_i)\neq y_i} = 1 \,;\; y_i h_t(x_i) = -1 \end{cases}$$

So we get:

$$\mathrm{E}_{t+1} = \frac{\sum_{i\in E} D_i^t \exp\left(-w_t y_i h_t(x_i)\right)}{Z_t} \cdot 0 + \frac{\sum_{i\in C} D_i^t \exp\left(-w_t y_i h_t(x_i)\right)}{Z_t} \cdot 1 = \frac{\sum_{i\in C} D_i^t \exp(w_t)}{Z_t}$$

The numerator contains the expression for the error value, since we have isolated for case $\{E :\ h_t(x_i) = y_i\}$:

$$\mathrm{E}_t = \sum_{i\in C} D_i^t$$

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t \exp(w_t)}{Z_t}$$

Opening the denominator using same 2 cases:

$$Z_t = \sum_{i\in E} D_i^t \exp(-w_t) + \sum_{i\in C} D_i^t \exp(w_t)$$

Putting back:

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t \exp(w_t)}{\sum_{i\in E} D_i^t \exp(-w_t) + \sum_{i\in C} D_i^t \exp(w_t)} \quad ; \ / \exp(w_t)$$

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t}{\sum_{i\in E} D_i^t \exp(-2w_t) + \sum_{i\in C} D_i^t} = \frac{\mathrm{E}_t}{\sum_{i\in E} D_i^t \exp(-2w_t) + \mathrm{E}_t}$$

Taking $\exp(-2w_t)$ out of the sum:

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t}{\exp(-2w_t) \sum_{i\in E} D_i^t + \mathrm{E}_t}$$

The sum over weights of the correct predictions E, is 1-sum over incorrect predictions, since they sum to 1:

$$\sum_{i\in E} D_i^t = 1 - \sum_{i\in C} D_i^t = 1 - E_t$$

So we get:

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t}{\exp(-2w_t)\,(1 - E_t) + \mathrm{E}_t}$$

Using the expression for the weight of the weak classifier:

$$w_t = \frac{1}{2} \log\left(\frac{1}{\mathrm{E}_t} - 1\right)$$

$$\exp(-2w_t) = \exp\left(-log\left(\frac{1}{\mathrm{E}_t} - 1\right)\right) = \exp\left(\log\left(\left(\frac{1}{\mathrm{E}_t} - 1\right)^{-1}\right)\right) = \left(\frac{1}{\mathrm{E}_t} - 1\right)^{-1} = \left(\frac{1 - \mathrm{E}_t}{\mathrm{E}_t}\right)^{-1}$$
$$= \frac{\mathrm{E}_t}{1 - \mathrm{E}_t}$$

Putting everything back:

$$\mathrm{E}_{t+1} = \frac{\mathrm{E}_t}{\frac{\mathrm{E}_t}{1 - \mathrm{E}_t}(1 - E_t) + \mathrm{E}_t} = \frac{\mathrm{E}_t}{\mathrm{E}_t + \mathrm{E}_t} = \frac{1}{2}$$

Question 2

2.a.
Let $\alpha \in \mathbb{R}_{>0}$
Notice σ is positive-homoegeneous, i.e. $\sigma(\alpha x) = \max\{0, \alpha x\} = \alpha \max\{0, x\} = \alpha\sigma(x)$
Denote $\alpha\theta = \left(\alpha W^{(1)}, ..., \alpha W^{(L)}\right)$.
Let h be defined on θ as in question and $\tilde{h}$ defined on $\alpha\theta$ as in question.

Claim. $\tilde{h}_L(x) = \alpha^L h_L(x)$

Proof. Proof by induction on L.
Base : $\tilde{h}_1(x) = \sigma\left(\alpha {W^{(1)}}^T x\right) \underbrace{=}_{\alpha\, positive\, homogeneous} \alpha\sigma\left({W^{(1)}}^T x\right) = \alpha h_1(x)$
Assume for n<L, $\tilde{h}_n(x) = \alpha^n h_n(x)$
Induction step:
By definition, $\tilde{h_L}(x) = \sigma\left(\alpha {W^{(L)}}^T \tilde{h}_{L-1}(x)\right) \underbrace{=}_{induction\, hypothesis} \sigma\left(\alpha {W^{(L)}}^T \alpha^{L-1} h_{L-1}(x)\right) =$

$\underbrace{=}_{\alpha\, positive-homogeneous} \alpha^L \sigma\left({W^{(L)}}^T h_{L-1}(x)\right) = \alpha^L h_L(x)$ ☐

By definition, $F_{\alpha\theta}(x) = {W^{(L)}}^T \tilde{h}_{L-1}(x) = {W^{(L)}}^T \alpha^{L-1} h_{L-1}(x) = \alpha^{L-1} F_\theta(x)$
Then $C = \alpha^{L-1}$.

2.b.

$$c_{y_i} = \frac{\exp\left((F_{\alpha\theta}(x))_i\right)}{\sum\limits_{j=1}^{k} \exp\left((F_{\alpha\theta}(x))_j\right)} = \frac{\exp\left(\alpha^L (F_\theta(x))_i\right)}{\sum\limits_{j=1}^{k} \exp\left(\alpha^L (F_\theta(x))_j\right)} \underset{\alpha\to 0}{\to} \frac{1}{k}$$

Then the distribution is uniform.

2.c.
<u>Case 1</u> $(F_\theta(x))_i > (F_\theta(x))_j, \forall j \in \{m_1, ..., m_{k-s}\}$ and $(F_\theta(x))_i = (F_\theta(x))_j, \forall j \in \{m_{k-s+1}, ..., m_k\}$, $(\{m_1, ..., m_k\} = \{1, ..., k\})$

$$c_{y_i} = \frac{\exp\left((F_{\alpha\theta}(x))_i\right)}{\sum\limits_{j=1}^{k} \exp\left((F_{\alpha\theta}(x))_j\right)} = \frac{\exp\left(\alpha^L (F_\theta(x))_i\right)}{\sum\limits_{j=1}^{k} \exp\left(\alpha^L (F_\theta(x))_j\right)} = \frac{1}{\sum\limits_{j=1}^{k} \exp\left(\alpha^L \left((F_\theta(x))_j - (F_\theta(x))_i\right)\right)} = ...$$

$$... = \frac{1}{s + \underbrace{\sum\limits_{j\in\{m_1,...,m_{k-s}\}} \exp\left(\alpha^L \left((F_\theta(x))_j - (F_\theta(x))_i\right)\right)}_{\underset{\alpha\to\infty}{\to} 0}} \underset{\alpha\to\infty}{\to} \frac{1}{s}$$

($s \geq 1$ because exists at least one maximum value then limit is well-defined)

<u>Case 2</u> $\exists 1 \leq j \leq k \; s.t. \quad (F_\theta(x))_i < (F_\theta(x))_j$

$$c_{y_i} = \frac{1}{1 + \underbrace{\sum\limits_{j\neq i} \exp\left(\alpha^L \left((F_\theta(x))_j - (F_\theta(x))_i\right)\right)}_{\underset{\alpha\to\infty}{\to} \infty}} \underset{\alpha\to\infty}{\to} 0$$

Then distribution is uniform on the $s \leq k$ entries with maximum value and zero otherwise.

i.e. $\left(0, ..., \underbrace{\frac{1}{s}}_{m}, ..., 0\right)$ where $m \in argmax_{1 \leq i \leq k} \left(\left(F_{\theta}\left(x\right)\right)_{i}\right)$.