Jonathan Nusantara jan 26 5

Problem 1. (10 points). Suppose AdaBoost is run on n training examples, and suppose on each round that the weighted training error ε_t of the tth weak hypothesis is at most $\frac{1}{2} - \gamma$, for some number $\gamma > 0$. Show that after $T > \frac{\ln n}{2\gamma^2}$ rounds of AdaBoost the final combined classifier has zero training error!

$$\varepsilon_t \leq \frac{1}{2} - \gamma$$

We know that in AdaBoost we will keep updating p_t for $i = 1 \dots n$

$$p_{t+1}(i) = \frac{p_t(i) \exp(-\alpha_t \cdot h_t(\bar{x}_i) \cdot y_i)}{\sum_{j=1}^{n} p_t(j) \exp(-\alpha_t \cdot h_t(\bar{x}_j) \cdot y_j)} \longrightarrow Z_t$$

$$p_t(i) = \frac{p_{t-1}(i) \exp(-\alpha_{t-1} \cdot h_{t-1}(\bar{x}_i) \cdot y_i)}{Z_{t-1}}$$

By induction:

$$p_{t+1}(i) = \frac{1}{Z_1 \cdot Z_2 \dots Z_t} \cdot \underbrace{p_1(i)}_{1/n \leftarrow \text{initialized probability at } t=1} \cdot \exp\left(-y_i \left(\alpha_1 h_1(\bar{x}_i) + \alpha_2 h_2(\bar{x}_i) + \dots + \alpha_t h_t(\bar{x}_i)\right)\right)$$

since probabilities add to 1:

$$Z_1 \cdot Z_2 \dots Z_T = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i(\alpha_1 h_1(\bar{x}_i) + \dots + \alpha_T h_T(\bar{x}_i))\right)$$

In lecture, we also proof that for h^* is the best h ← learner:

$$\mathbb{I}\{h_T^*(\bar{x}_i) \neq y_i\} \leq \exp\left(-y_i \underbrace{(\alpha_1 h_1(\bar{x}_i) + \dots + \alpha_T h_T(\bar{x}_i))}_{h_T^*(\bar{x}_i)}\right)$$

So, $\text{error}(h_T^*) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{h_T^*(\bar{x}_i) \neq y_i\}$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i(\alpha_1 h_1(\bar{x}_i) + \dots + \alpha_T h_T(\bar{x}_i))\right)$$

$$= Z_1 \cdot Z_2 \cdot \dots \cdot Z_T$$

Since, $\varepsilon_t = \sum_{i=1}^{n} p_t(i)\, \mathbb{I}\{h_t(\bar{x}_i) \neq y_i\}$,

$$Z_t = \sum_{i=1}^{n} p_t(\bar{x}_i) \cdot \exp(-\alpha_i h_t(\bar{x}_i) \cdot y_i)$$

$$= \varepsilon_t \cdot \exp(\alpha_t) + (1 - \varepsilon_t) \cdot \exp(-\alpha_t)$$

Because $error(h^*_T) \leq z_1 \cdot z_2 \cdot \ldots z_T$, let's make z small by choosing α

$$\frac{d}{d\alpha}\left(\varepsilon_t e^{\alpha} + (1-\varepsilon_t) e^{-\alpha}\right) = \varepsilon_t e^{\alpha} - (1-\varepsilon_t) e^{-\alpha} = 0$$

$$\log(\varepsilon_t e^{\alpha}) = \log((1-\varepsilon_t) e^{-\alpha})$$

$$\log(\varepsilon_t) + \log(e^{\alpha}) = \log(1-\varepsilon_t) + \log(e^{-\alpha})$$

$$\log(\varepsilon_t) + \alpha = \log(1-\varepsilon_t) - \alpha$$

$$2\alpha = \log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

$$\alpha_t = \frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

$$z_t = \varepsilon_t \cdot \exp(\alpha_t) + (1-\varepsilon_t) \cdot \exp(-\alpha_t)$$

$$= 2\sqrt{\varepsilon_t(1-\varepsilon_t)} \quad \leftarrow \quad \varepsilon_t \leq \frac{1}{2} - \gamma$$

$$\leq 2\sqrt{\left(\frac{1}{2}-\gamma\right)\left(\frac{1}{2}+\gamma\right)}$$

$$= \frac{1}{2}\left(1-(2\gamma)^2\right)^{1/2} \leq \exp(-2\gamma^2)$$

because we know if $x \leq e^x$ for all $x \in \mathbb{R}$

So, $z_t \leq \exp(-2\gamma^2)$

$$err(h_T^*) = z_1 \cdot \ldots \cdot z_T \leq \exp(-2\gamma^2 T)$$

If $T > \frac{\ln(n)}{2\gamma^2}$:

$$err(h_T^*) < \exp\left(-2\gamma^2 \cdot \frac{\ln(n)}{2\gamma^2}\right)$$

$$err(h_T^*) < \exp(\ln n)$$

$$err(h_T^*) < \frac{1}{n} \Rightarrow 0$$

Less than $\frac{1}{n}$ error means 0, because $\frac{1}{n}$ error means having one error out of the n training examples, and less than that means no errors.

Problem 2. (10 points). Recall bagging. Starting from a training set S of size n, we created m bootstrap training sets $S_1, \ldots, S_m$, each of size n each by sampling with replacement from S.

1. For a bootstrap sample S_i, what is the expected fraction of the training set that does not appear at all in S_i? As $n \to \infty$, what does this fraction approach?

2. Let $m > 2 \ln n$, and $n \to \infty$. Show that the expected number of training examples in S that appear in at least one S_i is more than $n - 1$.

1. Probability of selected $= \frac{1}{n}$

—1— not selected $= 1 - \frac{1}{n}$

← P of not selected in n trials $= \boxed{\left(1 - \frac{1}{n}\right)^n}$

in a single S_i

As $n \to \infty$,

$\lim_{n \to \infty} \left(1 - \frac{1}{n}\right)^n \to \frac{1}{e} = e^{-1} \approx \boxed{0.368}$

2. P(selected at least once) $= 1 -$ P(not selected at all)

in a single S_i ← P(selected at least once) $= 1 - \left(1 - \frac{1}{n}\right)^n$

P(selected in at least on S_i) $= 1 - \left(\left(1 - \frac{1}{n}\right)^n\right)^m$

Let $m > 2 \ln(n)$,

P(selected in at least on S_i) $> 1 - \left(\left(1 - \frac{1}{n}\right)^n\right)^{2 \ln(n)}$

Expected value = Probability $*$ n , $m > 2 \ln(n)$

$> n\left(1 - \left(\left(1 - \frac{1}{n}\right)^n\right)^{2 \ln(n)}\right)$

$> n - n\left(\left(1 - \frac{1}{n}\right)^n\right)^{2 \ln(n)}$

$> n - n\left(e^{-1}\right)^{2 \ln(n)}$ ← as $n \to \infty$ $\left(1 - \frac{1}{n}\right)^n \to e^{-1}$ proven in ①

$> n - n\left(e^{-2 \ln(n)}\right)$

$> n - n\left(\frac{1}{n^2}\right)$

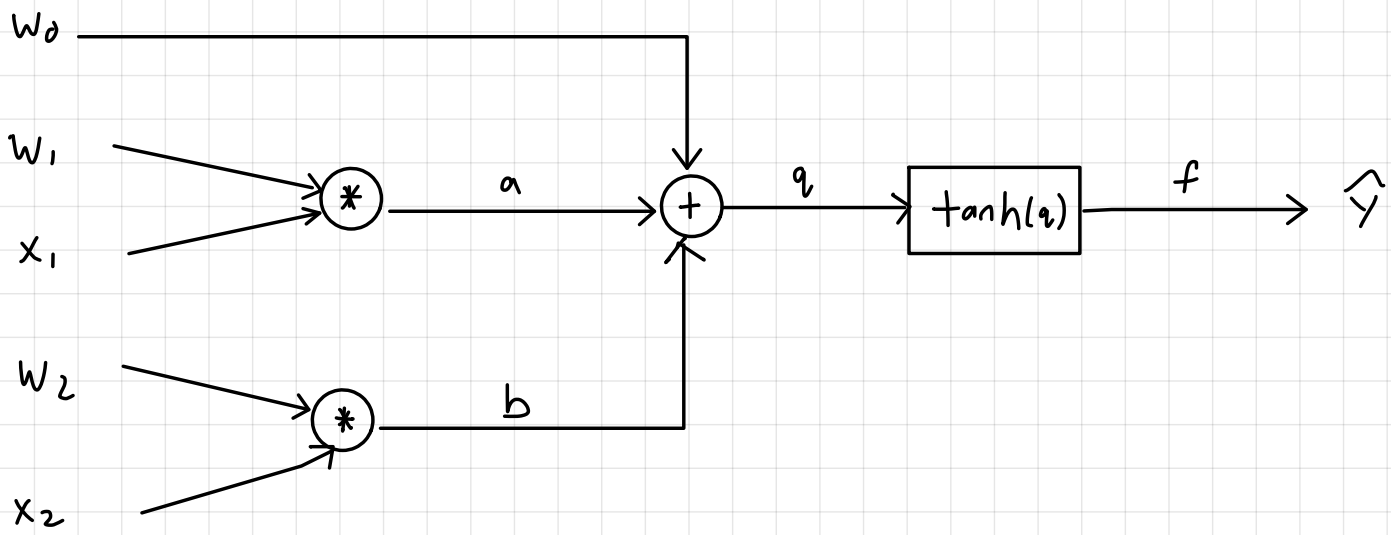
Expected value $> n - \frac{1}{n} > n - 1$, since $n > 0$

Expected value $> \boxed{n - 1}$

Problem 3. (10 points). The tanh function is $\tanh(y) = (e^y - e^{-y})/(e^y + e^{-y})$. Consider the function $\tanh(w_0 + w_1x_1 + w_2x_2)$, with five inputs, and a scalar output.

1. Draw the computational graph of the function (you can use tanh in your computation graph).
2. What is the derivative of $\tanh(y)$ with respect to y.
3. Suppose $(w_0, w_1, w_2, x_1, x_2) = (-2, -3, 1, 2, 3)$. Compute the forward function values, and back-propagation of gradients.

1. $\tanh(y) = \frac{(e^y - e^{-y})}{(e^y + e^{-y})}$, the function is $\tanh(w_0 + w_1x_1 + w_2x_2)$



2. $\frac{d}{dy}\tanh(y) = \frac{d}{dy}\frac{(e^y - e^{-y})}{(e^y + e^{-y})}$

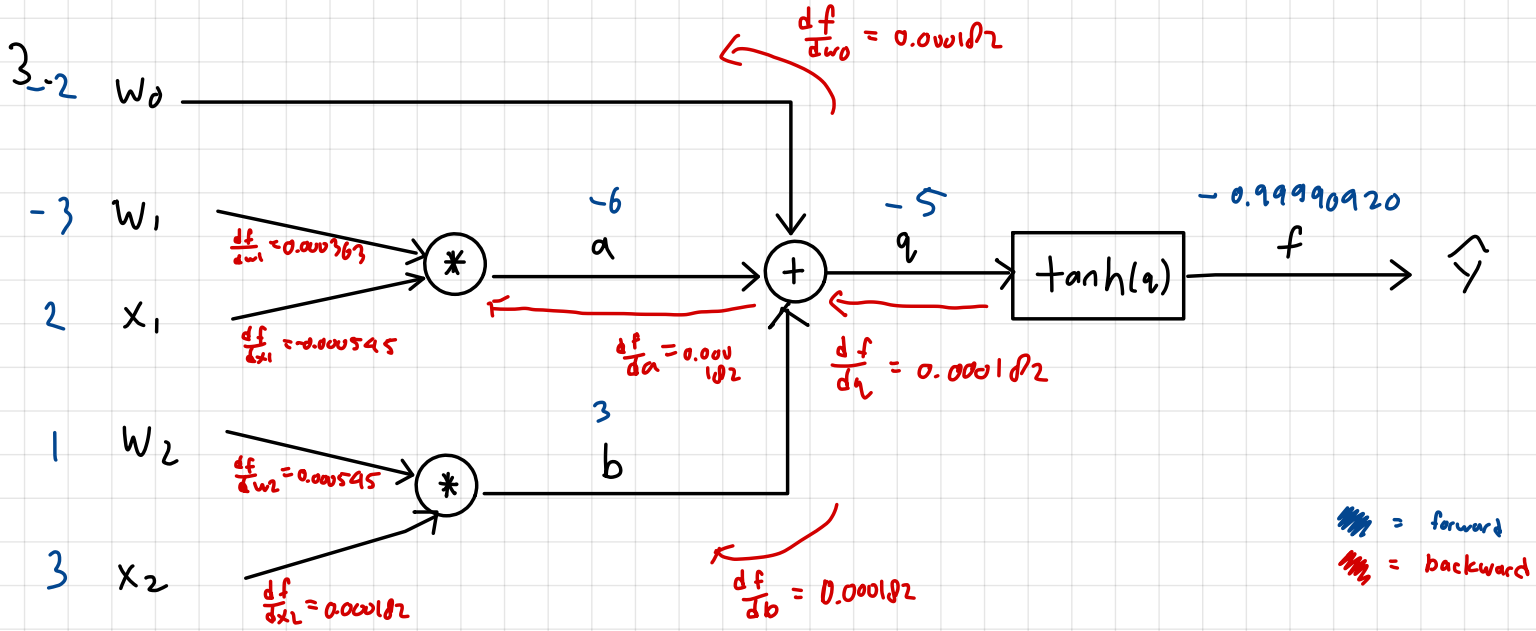
$$= \frac{(e^y + e^{-y})\frac{d}{dy}(e^y - e^{-y}) - \frac{d}{dy}(e^y + e^{-y})(e^y - e^{-y})}{(e^y + e^{-y})^2}$$

$$= \frac{(e^y + e^{-y})(e^y + e^{-y}) - (e^y - e^{-y})(e^y - e^{-y})}{(e^y + e^{-y})^2}$$

$$= \frac{(e^y + e^{-y})^2 - (e^y - e^{-y})^2}{(e^y + e^{-y})^2}$$

$$= \frac{(e^y + e^{-y})^2}{(e^y + e^{-y})^2} - \frac{(e^y - e^{-y})^2}{(e^y + e^{-y})^2}$$

$$= \boxed{1 - (\tanh(y))^2}$$



$q = w_0 + a + b$ $\quad f = \tanh(q)$, $a = w_1 x_1$, $b = w_2 x_2$

We want to find $\frac{\partial f}{\partial w_0}$, $\frac{\partial f}{\partial w_1}$, $\frac{\partial f}{\partial w_2}$, $\frac{\partial f}{\partial x_1}$, $\frac{\partial f}{\partial x_2}$

given $(w_0, w_1, w_2, x_1, x_2) = (-2, -3, 1, 2, 3)$

$q = -2 + (-3)(2) + (1)(3) = -5$

$$\frac{\partial f}{\partial w_0} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial w_0}$$

$$\frac{\partial q}{\partial w_0} = \frac{\partial (w_0 + a + b)}{\partial w_0} = 1$$

$$\frac{\partial f}{\partial q} = \frac{\partial (\tanh(q))}{\partial q} = 1 - (\tanh(q))^2$$

$$= (1 - (\tanh(q))^2) \cdot (1) = (1 - (\tanh(-5))^2) = \underline{0.000182}$$

$$\frac{\partial f}{\partial w_1} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial a} \cdot \frac{\partial a}{\partial w_1}$$

$$= (1 - (\tanh(q))^2) \cdot (1) \cdot (x_1^{\nearrow 2})$$

$$= 0.000182 \times 2 = \underline{0.000363}$$

$$\frac{\partial a}{\partial w_1} = \frac{\partial (w_1 x_1)}{\partial w_1} = x_1$$

$$\frac{\partial q}{\partial a} = \frac{\partial (w_0 + a + b)}{\partial a} = 1$$

$$\frac{\partial f}{\partial q} = 1 - (\tanh(q))^2$$

$$\frac{\partial f}{\partial w_2} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial b} \cdot \frac{\partial b}{\partial w_2}$$

$$= (1 - (\tanh(q))^2) \cdot (1) \cdot (x_2^{\nearrow 3})$$

$$= 0.000182 \times 3 = \underline{0.000545}$$

$$\frac{\partial b}{\partial w} = \frac{\partial (w_2 x_2)}{\partial w_2} = x_2$$

$$\frac{\partial q}{\partial b} = \frac{\partial (w_0 + a + b)}{\partial b} = 1$$

$$\frac{\partial f}{\partial q} = 1 - (\tanh(q))^2$$

$$\frac{\partial f}{\partial x_1} = \frac{df}{dq} \cdot \frac{\partial q}{\partial a} \cdot \frac{\partial a}{\partial x_1}$$

$$= \left(1 - \left(\tanh(q)\right)^2\right) \cdot (1) \cdot (\overrightarrow{w_1}^{-3})$$

$$= 0.0001816 \times -3 = \underline{-0.000545}$$

$$\frac{\partial a}{\partial x_1} = \frac{\partial (w_1 x_1)}{\partial x_1} = w_1$$

$$\frac{\partial q}{\partial a} = \frac{\partial (w_0 + a + b)}{\partial a} = 1$$

$$\frac{df}{dq} = 1 - \left(\tanh(q)\right)^2$$

$$\frac{\partial f}{\partial x_2} = \frac{df}{dq} \cdot \frac{\partial q}{\partial b} \cdot \frac{\partial b}{\partial x_2}$$

$$= \left(1 - \left(\tanh(q)\right)^2\right) \cdot (1) \cdot (\overrightarrow{w_2}^{1})$$

$$= 0.0001816 \times 1 = \underline{0.000182}$$

$$\frac{\partial b}{\partial x_2} = \frac{\partial (w_2 x_2)}{\partial x_2} = w_2$$

$$\frac{\partial q}{\partial b} = \frac{\partial (w_0 + a + b)}{\partial b} = 1$$

$$\frac{df}{dq} = 1 - \left(\tanh(q)\right)^2$$

Forward function:

$w_0 = -2$

$a = w_1 x_1 = -6$

$b = w_2 x_2 = 3$

$q = a + b + w_0 = -5$

$f = \tanh(-5) = -0.9999092$