caige middaugh
11/14/21
machine learning

Part I

1.) Pf: $E_{agg}(x) = E[\{\frac{1}{M}\sum_{i=1}^{M} h_i(x) - f(x)\}^2]$

$= E[\frac{1}{M^2}[\sum_{i=1}^{M} h_i(x) - f(x)]^2]$

since $E[ax+b] = aE[x]+b$ we can say:

$= \frac{1}{M^2} \cdot E[[\sum_{i=1}^{M}(h_i(x) - f(x))]^2]$

Now since $E[x_1+x_2+\dots+x_m] = E[x_1]+E[x_2]+\dots+E[x_m]$
$= \sum_{i=1}^{M} E[x_i]$

we can say:

$= \frac{1}{M^2} \cdot \sum_{i=1}^{M} E[-(f(x)-h_i(x))]^2$ and due to given assumptions.

$= \frac{1}{M^2} \cdot \sum_{i=1}^{M} E[f(x)-h_i(x))^2] \rightarrow \epsilon_i = f(x) - h_i$

$= \frac{1}{M^2} \cdot \sum_{i=1}^{M} E[\epsilon_i(x)^2] = \frac{1}{M} \cdot \boxed{\frac{1}{M} \cdot \sum_{i=1}^{M} E[\epsilon_i(x)^2]}$ notice dotted box

the dotted box is E_{avg}, substitute to get

$E_{agg}(x) = \frac{1}{M} E_{avg}$ ∎

2.) If "f" convex function on (a,b) and x is r.v. then

$$f(E(x)) \leq E[f(x)]$$

Then using Jensen's Inequality we get:

$$E[f(x)] = \lambda(x_1)\, f(x_1) + \sum_{i=2}^{m} \lambda(x_i)\, f(x_i)$$

$$= \lambda_i(x_1) f(x_1) + (1-\lambda(x_1)) \cdot \frac{\sum_{i=2}^{m} \lambda(x_i) f(x_i)}{(1-\lambda(x_1))}$$

$$\leq \lambda(x_1) f(x_1) + (1-\lambda(x_1))\, f\left(\frac{\sum_{i=2}^{m} \lambda(x_i)\, x_i}{1-P(x_1)}\right)$$

$$\leq f\left(\lambda(x_1)\, x_1 + (1-P(x_1))\left(\frac{\sum_{i=2}^{m} \lambda x_i\, x_i}{(1-\lambda(x_1))}\right)\right)$$

Hence

$E[f(x)] \leq f(E[x])$, notice:

$E_{agg} \geq E[f(x)]$, $E_{avg} \geq f(E[x])$ so we get

$E_{agg} \leq E_{avg}$

3.)

Hypothesis for Boolean classifier:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \quad (1)$$

weight for point i at $t+1$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \cdot e^{-\alpha_t h_t(i) y(i)} \quad (2)$$

Base case of recurrence:

$$D_1 = \frac{1}{N} \ \forall i \quad (3)$$

ϵ_t, error for adaboost can be measured with relation to D_t

$$\epsilon_t = \sum_{i: h_t(i) \neq y(i)} D_t(i) \quad (4)$$

meaning error at t is sum of weights corresponding to all points i which are misclassified meets the $h_t(i) \neq y(i)$

we start with (2)

$$D_{t+1}(i) = \frac{D_t(i)\, e^{-\alpha_t h_t(i) y(i)}}{Z_t}$$

Now notice we can say $h_t(i)$ and $y(i)$ is in $\{1, -1\}$ because this is a boolean classifier

we then expand the recurrence from (2) to its base case (3)

3 continued)

$$D_{t+1}(i) = D_1(i) \cdot \frac{e^{-\alpha_1 h_1(i) y(i)}}{Z_1} \cdot \frac{e^{-\alpha_2 h_2(i) y(i)}}{Z_2} \cdot \dots \cdot \frac{e^{-\alpha_t h_t(i) y(i)}}{Z_t}$$

from (3) $\rightarrow = \frac{1}{N} \cdot \frac{e^{-\alpha_1 h_1(i) y(i)}}{Z_1} \cdot \dots \cdot \frac{e^{-\alpha_t h_t(i) y(i)}}{Z_t}$

(5) $$= \frac{1}{N} \frac{e^{-y(i) f_t(i)}}{\prod_{j=1}^{t} Z_j} \quad \text{where } f_t(i) = -\sum_{j=1}^{t} \alpha_j h_j(i)$$

Lets look at T_H, the total training error for H.

$$T_H = \frac{1}{N} \sum_{(i: H(i) \neq y(i))} 1$$

This is average of mis-classified points. Another way to say it is accuracy loss.

$H(i) = \text{sign}(f(i))$, we can say:

$T_H = \frac{1}{N} \sum_{(i: y(i) f(i) \leq 0)}$ because $y(i)$ and $f(i)$ will have opposite signs when mis-classified.

Hence $y(i) \cdot f(i) \leq 0$,

with this T_H we can say

$$T_H = \frac{1}{N} \sum_{(i: y(i) f(i) \leq 0)} 1 \leq \frac{1}{N} \sum_i e^{-y(i) f(i)}$$

because $e^{-z} \geq 1$ when $z \leq 0$

Hence, $T_H \leq \frac{1}{N} \sum_i e^{-y(i) f(i)}$

3. (continued)

from ⑤ we substitute for

$$T_H \le \frac{1}{N} (m \prod_t Z_t) \sum D_{t+1}(i)$$

$$T_H \le (\prod_t Z_t) \underbrace{(\sum_i D_{t+1}(i))}_{\text{will be 1 because it's a probability distribution.}}$$

$$T_H \le \prod_t Z_t \quad ⑥$$

now $Z_t = \sum_i D_t(i) e^{-\alpha_t h_t(i) y(i)}$

$$= \sum_{i: h_t(i) = y(i)} D_t(i) e^{-\alpha_t} + \sum_{i: h_t(i) \neq y(i)} D_t(i) e^{\alpha_t}$$

we do this because of a $\{-1, 1\}$ range. $h_t(i) = y(i)$ is for 1 and $h_t(i) \neq y(i)$ is for -1.

$$Z_t = e^{-\alpha_t} \sum_{i: h_t(i) = y(i)} D_t(i) + e^{\alpha_t} \sum_{i: h_t(i) \neq y(i)} D_t(i)$$

from ④

$Z_t = e^{-\alpha_t}(1 - \epsilon_t) + e^{\alpha_t} \epsilon_t$, to minimize we must evaluate α_t roughly comes out to be $\frac{1 - \epsilon_t}{\epsilon_t}$

substitute this and simplify. Due to lengthyness I will provide the simplified answer of:

$Z_t = 2\sqrt{\epsilon_t (1 - \epsilon_t)}$, consider the given

$$\epsilon_t = \frac{1}{2} - \gamma_t$$

3 continued)

$$Z_t = 2\sqrt{(\tfrac{1}{2} - Y_t)(\tfrac{1}{2} + Y_t)}$$, multiply inside sqrt and simplify

$$= \cancel{2}\frac{\sqrt{1-4Y_t^2}}{\cancel{2}} = \sqrt{1-4Y_t^2}$$

consider $1+x \leq e^x \;\; \forall x \in \mathbb{R}$

thus $1-4Y_t^2 \leq e^{-4Y_t^2} \;\; \forall \{-4Y_t^2\} \in \mathbb{R}$

notice this inequality gives us:

$$Z_t \leq \sqrt{e^{-4Y_t^2}} = e^{-2Y_t^2}$$

Now substitute this Z_t into 6

$$T_H \leq \prod_t Z_t$$

$$T_H \leq \prod_t e^{-2Y_t^2} = e^{-2\sum_{t=1}^{T} Y_t^2}$$

Hence

$$T_H \leq e^{-2\sum_{t=1}^{T} Y_t^2} \quad \blacksquare$$