CS 6375

ASSIGNMENT __3__________

Names of students in your group:

Nimrat Bedi

Net ID: nxb200004

Number of free late days used: _1__________________

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Report : Part I
Theory Questions

Page:
Date:

7.10.20

ASSIGNMENT 3

PART 1

Ans 1. $E_{avg} = \frac{1}{M} \sum_{i=1}^{M} E(\varepsilon_i(x)^2)$

$h_{agg}(x) = \frac{1}{M} \sum_{i=1}^{M} h_i(x)$

Aggregated model $E_{agg}(x) = E\left[\left\{\frac{1}{M} \sum_{i=1}^{M} \epsilon_i(x)\right\}^2\right]$

To prove $E_{agg} = \frac{1}{M} E_{avg}$

$$E_{agg}(x) = E\left[\left\{\frac{1}{M} \sum_{i=1}^{M} \epsilon_i(x)\right\}^2\right]$$

$$= \frac{1}{M^2} E\left[\left\{\sum_{i=1}^{M} \epsilon_i(x)\right\}^2\right]$$

$$= \frac{1}{M^2} E\left[\epsilon_1(x)^2 + \epsilon_2^2(x)^2 + \epsilon_3^2(x)^2 + \dots + \epsilon_M^2(x)^2\right.$$

$$\left. + 2\epsilon_1(x)\epsilon_2(x) + 2\epsilon_1(x)\epsilon_3(x) + \dots\right]$$

$$= \frac{1}{M^2} E\left[\epsilon_1^2(x) + \epsilon_2^2(x) + \dots + \epsilon_m^2(x)\right]$$

as $\epsilon_i(x) \cdot \epsilon_j(x) = 0$

(given in assumption

$$= \frac{1}{M^2}\left[E(\epsilon_1^2(x)) + E(\epsilon_2^2(x)) + \dots\right]$$

$$= \frac{1}{M} \cdot \frac{1}{M} \sum_{i=1}^{M} E(\epsilon_i(x)^2)$$

$$E_{agg} = \frac{1}{M} E_{avg}$$

Hence Proved.

Ans 2. Jensen's inequality states that for any convex function f :-

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i)$$

Using Jensen's inequality, to prove

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

→ From quadratic inequality, we have :-

$$\frac{(x_1 + x_2 + x_3 + \dots + x_n)^2}{n} \leq x_1^2 + x_2^2 + x_3^2 + \dots + x_n^2$$

Using this in our error function :-

$$\frac{1}{M}\{\epsilon_1(x) + \epsilon_2(x) + \dots \dots + \epsilon_m(x)\}^2 \leq \epsilon_1(x)^2 + \epsilon_2(x)^2 + \dots \dots + \epsilon_n(x)^2$$

$$\Rightarrow \frac{1}{M}\left\{\sum_{i=1}^{M} \epsilon_i(x)\right\}^2 \leq \sum_{i=1}^{M} \{\epsilon_i(x)\}^2$$

Putting estimate value equation above :-

$$E\left[\frac{1}{M}\left\{\sum_{i=1}^{M} \epsilon_i(x)\right\}^2\right] \leq E\left[\sum_{i=1}^{M} \{\epsilon_i(x)\}^2\right]$$

Dividing L.H.S & RHS by M

$$\frac{1}{M} \times E\left[\frac{1}{M}\left\{\sum_{i=1}^{M} \epsilon_i(x)\right\}^2\right] \leq \frac{1}{M} \times E\left[\sum_{i=1}^{M} \{\epsilon_i(x)\}^2\right]$$

Page:
Date:

E is a linear function i.e

$$E\left[\sum_{i=1}^{M} x_i\right] = \sum_{i=1}^{M} E(x_i)$$

Therefore :-

$$E\left[\left\{\frac{1}{M}\sum_{i=1}^{M} \epsilon_i(x)\right\}^2\right] \leq \frac{1}{M}\sum_{i=1}^{M} E\left[\left\{\epsilon_i(x)\right\}^2\right]$$

$\because E_{agg} = E\left[\left\{\frac{1}{M}\sum_{i=1}^{M} \epsilon_i(x)\right\}^2\right]$

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

$E_{avg} = \frac{1}{M}\sum_{i=1}^{M} E\left[\left\{\epsilon_i(x)\right\}^2\right]$

Page:
Date:

Ans 3 Deriving the training error for AdaBoost

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

weight for point i at step t+1 is given by:

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

$$D_1 = \frac{1}{N}$$ N = total no. of data points

$$\varepsilon_t = \frac{1}{2} - \gamma_t$$ = total error

→ Step 1 : The first step is to show that

$$D_{T+1}(i) = \frac{1}{N} \frac{e^{-y_i f(x)}}{\prod_t Z_t} \quad \text{— (1)}$$

where $f(x) = \sum_t \alpha_t h_t(x)$

eq (1) can be rewritten as:—

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

Suppose y_i and $h_t(x_i)$ are both in $\{+1, -1\}$ $\{-1, +1\}$

Recursive definition → $\therefore D_{T+1}(i) = D_1(i) \cdot \frac{e^{-\alpha_1 y_i h_1(x_i)}}{Z_1} \cdots\cdots \frac{e^{-\alpha_T y_i h_T(x_i)}}{Z_T}$

Page:
Date:

$$= \frac{1}{N} \cdot \frac{e^{-y_i \Sigma_t \alpha_t h_t (x_i)}}{\Pi_t Z_t}$$

$$= \frac{1}{N} \cdot \frac{e^{-y_i f(x_i)}}{\Pi_t Z_t} \quad \text{as } D_1 = \frac{1}{N}$$

Step 2: Next, we show that the training error of the final classifier H is at most

$$\prod_{t=1}^{T} Z_t$$

Proof:-

$$\text{training error } (H) = \frac{1}{N} \sum_i \begin{cases} 1 & \text{if } y_i \neq H(x_i) \\ 0 & \text{else} \end{cases}$$

by defination of training error

$$= \frac{1}{N} \sum_i \begin{cases} 1 & \text{if } y_i f(x_i) \leq 0 \\ 0 & \text{else} \end{cases}$$

since $H(x) = \text{sign}(f(x))$ ~~and~~ ~~$y_i \in$~~

$$\leq \frac{1}{N} \sum_i e^{(-y_i f(x_i))} \quad \text{since } e^{-z} \geq 1 \text{ if } z \leq 0$$

$$= \sum_t D_{T+1}(i) \prod_t Z_t \quad \text{by step 1 above}$$

$$\text{Training error} = \prod_t^T Z_t \quad \text{since } D_{T+1} \text{ is a distribution}$$

Step 3: The last step is to compute Z_t
we can compute this normalization constant as follows:-

$$Z_t = \sum_t D_t(i) \times \begin{cases} e^{-\alpha_t} & \text{if } h(x_i) = y_i \\ e^{\alpha_t} & \text{if } h(x_i) \neq y_i \end{cases}$$

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

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

$$= e^{-\alpha_t}(1-\varepsilon_t) + e^{\alpha_t}\varepsilon_t$$ (by defination of ε_t)

$$= 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$ (by our choice of α_t → which was chosen to minimize this expression)

$$= \sqrt{1-4\gamma_t^2}$$ plugging in $\varepsilon_t = \frac{1}{2} - \gamma_t$

$$Z_t \leq e^{-2\gamma_t^2}$$ using $1+x \leq e^x$ for all real x

Combining with step 2 gives the claimed upper bound on training error of H

$$\therefore \varepsilon_{training} \leq \prod_t Z_t \leq \exp\left(-2\sum_t \gamma_t^2\right)$$

Report : Part II

Tweets Clustering using k-means

<u>Dataset used : gdnhealthcare</u>

Value of K	SSE	Size of each cluster
5	2284	1 : 171 2 : 358 3 : 1260 4 : 433 5 : 773
10	2186	1 : 93 2 : 168 3 : 296 4 : 106 5 : 354 6 : 257 7 : 795 8 : 231 9 : 213 10 : 482
15	2181	1 : 85 2 : 75 3 : 38 4 : 201 5 : 52 6 : 390 7 : 349 8 : 497 9 : 271 10 : 108 11 : 108 12 : 280 13 : 59 14 : 385 15 : 97
20	2146	1 : 32 2 : 159 3 : 38 4 : 116 5 : 203

		6 : 123 7 : 29 8 : 299 9 : 214 10 : 148 11 : 199 12 : 71 13 : 64 14 : 182 15 : 168 16 : 136 17 : 307 18 : 309 19 : 153 20 : 45
25	2039.0037	1 : 78 2 : 27 3 : 199 4 : 101 5 : 127 6 : 144 7 : 80 8 : 170 9 : 114 10 : 106 11 : 69 12 : 123 13 : 140 14 : 144 15 : 34 16 : 125 17 : 30 18 : 57 19 : 231 20 : 181 21 : 470 22 : 76 23 : 88 24 : 61 25 : 20