CS 6375

ASSIGNMENT 2 K-Means Clustering

Names of students in your group:

Naxel Santiago NAS180011

Danny Bao DXB180034

Number of free late days used: 0

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.

Naxel Santiago NAS180011
Danny Bao DXB180034

CS 6375 Assignment 2

1) **Prove that $E_{agg} = \frac{1}{M} E_{avg}$ provided you make the following assumptions:**
   - $E(\varepsilon_i(x)) = 0 \; for \; all \; i$
   - $E\left(\varepsilon_i(x)\varepsilon_j(x)\right) = 0 \; for \; all \; i \neq j$

   Consider the formula for the error using the aggregated model:

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

   We can move the $\frac{1}{M^2}$ outside of the brackets since it is a constant:

$$E_{agg}(x) = \frac{1}{M^2} E\left[\left\{\sum_{i=1}^{M} \varepsilon_i(x)\right\}^2\right] = \frac{1}{M^2} E\left[\sum_{i=1}^{M} \varepsilon_i(x)^2 + 2\sum_{j=1}^{M}\sum_{i=1}^{j-1} \varepsilon_i(x)\varepsilon_j(x)\right]$$

   Using the Linearity of Expectation property, we can rewrite this as:

$$E_{agg}(x) = \frac{1}{M^2}\left(E\left[\sum_{i=1}^{M} \varepsilon_i(x)^2\right] + E\left[2\sum_{j=1}^{M}\sum_{i=1}^{j-1} \varepsilon_i(x)\varepsilon_j(x)\right]\right)$$

   Again, using Linearity of Expectation, we can move the E inside the summation:

$$E_{agg}(x) = \frac{1}{M^2}\left(\sum_{i=1}^{M} E[\varepsilon_i(x)^2] + 2\sum_{j=1}^{M}\sum_{i=1}^{j-1} E\left[\varepsilon_i(x)\varepsilon_j(x)\right]\right)$$

   Since $E\left(\varepsilon_i(x)\varepsilon_j(x)\right) = 0 \; for \; all \; i \neq j$ (assumption 2), this can be simplified to:

$$E_{agg}(x) = \frac{1}{M^2}\left(\sum_{i=1}^{M} E[\varepsilon_i(x)^2] + 2(0)\right) = \frac{1}{M^2}\left(\sum_{i=1}^{M} E[\varepsilon_i(x)^2]\right)$$

   This can be rewritten as:

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

   Since $E_{avg} = \frac{1}{M}\sum_{i=1}^{M} E[\varepsilon_i(x)^2]$, we can replace that with E_{avg}:

$$E_{agg}(x) = \frac{1}{M} \times E_{avg}$$

   Thus, we have proven that $E_{agg} = \frac{1}{M} E_{avg}$.

Naxel Santiago NAS180011
Danny Bao DXB180034

CS 6375 Assignment 2

2) Show that using Jensen's inequality, it is still possible to prove that:

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

Recall the formula $E_{agg} = E\left[\left\{\frac{1}{M}\sum_{i=1}^{M}\epsilon_i(x)\right\}^2\right]$. Since $f(x) = x^2$ is a convex function, we can apply Jensen's rule to the section of E_{agg} inside the brackets:

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

Take expectation of both sides:

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

$$E\left[\left\{\frac{1}{M}\sum_{i=1}^{M}\epsilon_i(x)\right\}^2\right] \leq \frac{1}{M}E\left[\sum_{i=1}^{M}\epsilon_i(x)^2\right] \qquad : \text{since } \frac{1}{M} \text{ is a constant}$$

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

$$E_{agg} \leq \frac{1}{M}\sum_{i=1}^{M}E[\epsilon_i(x)^2] \qquad : \text{Linearity of Expectation } \left(E\left[\sum_{i=1}^{N}X_i\right] = \sum_{i=1}^{N}E[X_i]\right)$$

$$E_{agg} \leq E_{avg} \qquad : \text{since } E_{avg} = \frac{1}{M}\sum_{i=1}^{M}E[\epsilon_i(x)^2]$$

Thus, we have proven that the inequality $E_{agg} \leq E_{avg}$ can still hold true even when the errors are correlated.

3) Prove that at the end of *T* steps, the overall training error will be bounded by:

$$\boldsymbol{e^{-2\sum_{t=1}^{T}\gamma_t^2}}$$

Recall the form$ula\ for\ D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times e^{-\alpha_t h_t(i) y(i)}$. This can be rewritten as:

$$D_{t+1}(i) = D_1(i) \times \frac{e^{-\alpha_1 h_1(i) y(i)}}{Z_1} \times \ \dots \times \frac{e^{-\alpha_T h_T(i) y(i)}}{Z_T} = D_1(i) \times \frac{e^{-y(i)\sum_{t=1}^{T}\alpha_t h_t(i)}}{\prod_t Z_t}$$

Since $D_1(i) = \frac{1}{N}$:

Naxel Santiago NAS180011
Danny Bao DXB180034

CS 6375 Assignment 2

$$D_{t+1}(i) = \frac{1}{N} * \frac{e^{-y(i)\sum_{t=1}^{T}\alpha_t h_t(i)}}{\prod_t Z_t}$$

Since $D_{t+1}(i)$ is a distribution, $D_{t+1}(i) = 1$ if we sum both sides from *1* to *N*:

$$1 = \frac{1}{N} * \frac{\sum_{i=1}^{N} e^{-y(i)\sum_{t=1}^{T}\alpha_t h_t(i)}}{\prod_t Z_t} \quad \rightarrow \quad \prod_t Z_t = \frac{1}{N}\sum_{i=1}^{N} e^{-y(i)\sum_{t=1}^{T}\alpha_t h_t(i)}$$

Now, consider the following formula for the training error of hypothesis H :

$$Err_{train} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 & if\ H(i) \neq y(i) \\ 0 & otherwise\end{cases}$$

Given $H(i) = sign(\sum_{t=1}\alpha_t h_t(i))$, we can rewrite this as:

$$Err_{train} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 & if\ y(i) * \sum_{t=1}^{T}\alpha_t h_t(i) \leq 0 \\ 0 & otherwise\end{cases}$$

Recall the equation previously calculated for $\prod_t Z_t$, and that $e^{-x} \geq 1$ if $x \leq 0$ and $e^{-x} > 0$ if $x > 0$, where $x = y(i) * \sum_{t=1}^{T}\alpha_t h_t(i)$ in this case. Thus, we can say that:

$$\frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 & if\ y(i) * \sum_{t=1}^{T}\alpha_t h_t(i) \leq 0 \\ 0 & otherwise\end{cases} \quad \leq \quad \frac{1}{N}\sum_{i=1}^{N} e^{-y(i)\sum_{t=1}^{T}\alpha_t h_t(i)}$$

Which is equivalent to:

$$Err_{train} \leq \prod_t Z_t$$

Now solve for Z_t :

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

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

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

$$Z_t \times \sum_{i=1}^{N} D_{t+1}(i) = \sum_{i=1}^{N} D_t(i) e^{-\alpha_t h_t(i) y(i)} \qquad : \text{since } Z_t \text{ is a constant}$$

Naxel Santiago NAS180011
Danny Bao DXB180034

CS 6375 Assignment 2

$$Z_t = \sum_{i=1}^{N} D_t(i) e^{-\alpha_t h_t(i) y(i)} \qquad : \text{since } D_{t+1} \text{ is a distribution, } \sum_{i=1}^{N} D_{t+1}(i) = 1$$

$$Z_t = \sum_{i=1}^{N} D_t(i) e^{-\alpha_t} I[h_t(i) = y(i)] + \sum_{i=1}^{N} D_t(i) e^{\alpha_t} I[h_t(i) \neq y(i)]$$

Since $I[h_t(i) = y(i)] = 1 - I[h_t(i) \neq y(i)]$, we can rewrite this as:

$$Z_t = \sum_{i=1}^{N} D_t(i) e^{-\alpha_t} (1 - I[h_t(i) \neq y(i)]) + \sum_{i=1}^{N} D_t(i) e^{\alpha_t} I[h_t(i) \neq y(i)]$$

$$= e^{-\alpha_t} \sum_{i=1}^{N} D_t(i) (1 - I[h_t(i) \neq y(i)]) + e^{\alpha_t} \sum_{i=1}^{N} D_t(i) I[h_t(i) \neq y(i)]$$

$$Z_t = e^{-\alpha_t} \left(\sum_{i=1}^{N} D_t(i) - \sum_{i=1}^{N} D_t(i) I[h_t(i) \neq y(i)] \right) + e^{\alpha_t} \sum_{i=1}^{N} D_t(i) I[h_t(i) \neq y(i)]$$

Recall the following formula for the training error ϵ_t of h_t (from the lecture slides):

$$\epsilon_t = \sum_{i=1}^{N} D_t(i) I[h_t(i) \neq y(i)]$$

Using this, we can rewrite Z_t as:

$$Z_t = e^{-\alpha_t}(1 - \epsilon_t) + e^{\alpha_t} \epsilon_t$$

Since $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$:

$$Z_t = e^{-\frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}} (1 - \epsilon_t) + e^{\frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}} \epsilon_t$$

$$= e^{-\ln \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}} (1 - \epsilon_t) + e^{\ln \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}} \epsilon_t$$

$$= \frac{1}{\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}} (1 - \epsilon_t) + \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \epsilon_t$$

$$= \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} (1 - \epsilon_t) + \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \epsilon_t$$

$$= \sqrt{\frac{\epsilon_t (1-\epsilon_t)^2}{1-\epsilon_t}} + \sqrt{\frac{(1-\epsilon_t)(\epsilon_t)^2}{\epsilon_t}}$$

$$= \sqrt{\epsilon_t (1 - \epsilon_t)} + \sqrt{(1 - \epsilon_t) \epsilon_t}$$

$$Z_t = 2\sqrt{\epsilon_t (1 - \epsilon_t)}$$

Naxel Santiago NAS180011
Danny Bao DXB180034

CS 6375 Assignment 2

Since $\epsilon_t = \frac{1}{2} - \gamma_t$ for the total error of h_t :

$$Z_t = 2\sqrt{\left(\frac{1}{2} - \gamma_t\right)\left(1 - \left(\frac{1}{2} - \gamma_t\right)\right)}$$

$$= \sqrt{4\left(\left(\frac{1}{2} - \gamma_t\right)\left(\frac{1}{2} + \gamma_t\right)\right)}$$

$$Z_t = \sqrt{1 - 4\gamma_t^2}$$

Using the inequality $1 - x \leq e^{-x}$, we can conclude that:

$$Z_t = \sqrt{1 - 4\gamma_t^2} \leq \sqrt{e^{-4\gamma_t^2}} \quad \rightarrow \quad e^{\frac{1}{2}(-4\gamma_t^2)} \quad \rightarrow \quad e^{-2\gamma_t^2}$$

$$\prod_t Z_t \leq \prod_t e^{-2\gamma_t^2}$$

$$\prod_t Z_t \leq e^{-2\sum_{t=1}^{T}\gamma_t^2}$$

Earlier we showed that $Err_{train} \leq \prod_t Z_t$. Thus, we now have:

$$Err_{train} \leq \prod_t Z_t \leq e^{-2\sum_{t=1}^{T}\gamma_t^2}$$

Using the transitive law (which states that if $a \leq b \leq c,$ then $a \leq c$), we can conclude that:

$$Err_{train} \leq e^{-2\sum_{t=1}^{T}\gamma_t^2}$$

Thus, we have proven that the overall training error of the hypothesis H will be less than or equal to $e^{-2\sum_{t=1}^{T}\gamma_t^2}$.