1E.

Outputs for Sample Values –

Official: For N = 0, train error = 0.501969611705121, validation error = 0.5073157006190209

Official: For N = 1, train error = 0.4501969611705121, validation error = 0.4805852560495217

Official: For N = 2, train error = 0.32386043894203714, validation error = 0.4113674732695554

Official: For N = 3, train error = 0.0025323579065841305, validation error = 0.31429375351716377

Official: For N = 4, train error = 0.0, validation error = 0.28081035453010694

Official: For N = 5, train error = 0.0, validation error = 0.27096229600450195

Official: For N = 6, train error = 0.0, validation error = 0.2771525042205965

Official: For N = 7, train error = 0.0002813731007315701, validation error = 0.27490151941474394

Official: For N = 8, train error = 0.0005627462014631402, validation error = 0.2923466516601013

Official: For N = 9, train error = 0.0008441193021947102, validation error = 0.31035453010692177

Official: For N = 96, train error = 0.27293190770962295, validation error = 0.5073157006190209

Official: For N = 97, train error = 0.2771525042205965, validation error = 0.5073157006190209

Official: For N = 98, train error = 0.2819358469330332, validation error = 0.5073157006190209

Official: For N = 99, train error = 0.28559369724254363, validation error = 0.5073157006190209

Official: For N = 100, train error = 0.29009566685424876, validation error = 0.507315700619020

Short Paragraph –

I tried the function testValuesOfN(n) with 100 different values of n from 0 – 100. The error rate drastically reduced from 32% at n=2 to 2.5% at n=3. The error rate was 0 for values of n in the range 4 to 6. The error rate started increasing again from n=7 and kept increasing till n=100. Using this data, it would be fair to assume that a value for n between 4 to 6 is the best choice to create an accurate training model.

The lowest error rate for values of n between 4 to 6 might be because in this range, the feature map is able to accurately extract meaningful features from the training data and assign appropriate weights to the extracted features. When the value of n increases from 7 onwards, the features become larger, and it might not accurately understand the patterns of the reviews to perform accurate classification of sentiments.

One Line Review –

“The reason n-grams likely outperforms word function is because it does not care about the format or spellings of the words and focuses more on interpreting the sentiment by purely analyzing the text patterns in the reviews.”

K-means clustering

2]a. $\phi(x_1) = [10, 0]$ $\phi(x_3) = [10, 20]$
$\phi(x_2) = [30, 0]$ $\phi(x_4) = [20, 20]$

i. Initial Centers: $u_1 = [20, 30]$ $u_2 = [20, -10]$

Distance from initial centers –

Point $x_1 \rightarrow$ ~~d~~ $d_{x_1 u_1} = \sqrt{(10-20)^2 + (0-30)^2} = \sqrt{100+900} = 31.62$

~~$d_{x_1 u_2} = \sqrt{(20-20)^2 + (30-(-10))^2} = \sqrt{0+40^2} =$~~

$d_{x_1 u_2} = \sqrt{(10-20)^2 + (0-(-10))^2} = \sqrt{100+100} = 14.14$

Point $x_2 \rightarrow$ $d_{x_2 u_1} = \sqrt{(30-20)^2 + (0-30)^2} = \sqrt{100+900} = 31.62$

$d_{x_2 u_2} = \sqrt{(30-20)^2 + (0-(-10))^2} = \sqrt{100+100} = 14.14$

Point $x_3 \rightarrow$ $d_{x_3 u_1} = \sqrt{(10-20)^2 + (20-30)^2} = \sqrt{100+100} = 14.14$

$d_{x_3 u_2} = \sqrt{(10-20)^2 + (20-(-10))^2} = \sqrt{100+900} = 31.62$

Point $x_4 \rightarrow$ $d_{4u_1} = \sqrt{(20-20)^2 + (20-30)^2} = \sqrt{0+100} = 10$

$d_{4u_2} = \sqrt{(20-20)^2 + (20-(-10))^2} = \sqrt{900} = 30$

Assignments –

Point x_1 and x_2 assigned to cluster 2 (u_2)
Point x_3 and x_4 assigned to cluster 1 (u_1)

Update Cluster Center –

$u_1 = \text{mean}(x_3, x_4) = \left[\frac{10+20}{2}, \frac{20+20}{2}\right] = [15, 20] \rightarrow$ New u_1

$u_2 = \text{mean}(x_1, x_2) = \left[\frac{10+30}{2}, \frac{0+0}{2}\right] = [20, 0] \rightarrow$ New u_2

Distance from new centers - $u_1' = [15, 20]$ $u_2' = [20, 0]$

Point $x_1 \rightarrow d_{x_1 u_1'} = \sqrt{(10-15)^2 + (0-20)^2} = \sqrt{25+400} = 20.62$

$d_{x_1 u_2'} = \sqrt{(10-20)^2 + (0-0)^2} = \sqrt{100} = 10$

Point $x_2 \rightarrow d_{x_2 u_1'} = \sqrt{(30-15)^2 + (0-20)^2} = \sqrt{225+400} = 25$

$d_{x_2 u_2'} = \sqrt{(30-20)^2 + (0-0)^2} = \sqrt{100+0} = 10$

Point $x_3 \rightarrow d_{x_3 u_1'} = \sqrt{(10-15)^2 + (20-20)^2} = \sqrt{25+0} = 5$

$d_{x_3 u_2'} = \sqrt{(10-20)^2 + (20-0)^2} = \sqrt{100+400} = 22.3$

Point $x_4 \rightarrow d_{x_4 u_1'} = \sqrt{(20-15)^2 + (20-20)^2} = \sqrt{25+0} = 5$

$d_{x_4 u_2'} = \sqrt{(20-20)^2 + (20-0)^2} = \sqrt{0+400} = 25$

Assignments -

Point x_1 and x_2 are still assigned to cluster 2 (u_2)
Point x_3 and x_4 are still assigned to cluster 1 (u_1)

Convergence -

Since, the assignments didn't change after the second iteration, we do not need to update the centers

This means, convergence has been reached.

Clusters -

cluster 1	
x_3	x_4

cluster 2	
x_1	x_2

ii] Initial Centers - $u_1 = [10, 20]$ $u_2 = [20, 30]$

distances from initial centers -

Point x_1 - $d_{x_1 u_1} = \sqrt{(10-0)^2 + (0-10)^2} = \sqrt{100+100} = 14.14$
$d_{x_1 u_2} = \sqrt{(10-30)^2 + (0-20)^2} = \sqrt{400+400} = 28.28$

Point x_2 - $d_{x_2 u_1} = \sqrt{(30-0)^2 + (0-10)^2} = \sqrt{900+100} = 31.62$
$d_{x_2 u_2} = \sqrt{(30-30)^2 + (0-20)^2} = \sqrt{0+400} = 20$

Point x_3 - $d_{x_3 u_1} = \sqrt{(10-0)^2 + (20-10)^2} = \sqrt{100+100} = 14.14$
$d_{x_3 u_2} = \sqrt{(10-30)^2 + (20-20)^2} = \sqrt{400+0} = 20.00$

Point x_4 - $d_{x_4 u_1} = \sqrt{(20-0)^2 + (20-10)^2} = \sqrt{400+100} = 22.36$
$d_{x_4 u_2} = \sqrt{(20-30)^2 + (20-20)^2} = \sqrt{100+0} = 10$

Assignments -

Point x_1 and Point x_3 assigned to cluster 1 (u_1)
Point x_2 and Point x_4 assigned to cluster 2 (u_2)

Update Centers -

$u_1' = \text{mean}(x_1, x_3) = \left[\frac{10+10}{2}, \frac{0+20}{2}\right] = [10, 10]$ → New u_1

$u_2' = \text{mean}(x_2, x_4) = \left[\frac{30+20}{2}, \frac{0+20}{2}\right] = [25, 10]$ → New u_2

Distance from new centers - $u_1' = [10, 10]$ $u_2' = [25, 10]$

Point x_1 - $d_{x_1 u_1'} = \sqrt{(10-10)^2 + (0-10)^2} = \sqrt{0+100} = 10$

$d_{x_1 u_2'} = \sqrt{(10-25)^2 + (0-10)^2} = \sqrt{225+100} = 18.03$

Point x_2 - $d_{x_2 u_1'} = \sqrt{(30-10)^2 + (0-10)^2} = \sqrt{400+100} = 22.36$

$d_{x_2 u_2'} = \sqrt{(30-25)^2 + (0-10)^2} = \sqrt{25+100} = 11.18$

Point x_3 - $d_{x_3 u_1'} = \sqrt{(10-10)^2 + (20-10)^2} = \sqrt{0+100} = 10$

$d_{x_3 u_2'} = \sqrt{(10-25)^2 + (20-10)^2} = \sqrt{225+100} = 18.03$

Point x_4 - $d_{x_4 u_1'} = \sqrt{(20-10)^2 + (20-10)^2} = \sqrt{100+100} = 14.14$

$d_{x_4 u_2'} = \sqrt{(20-25)^2 + (20-10)^2} = \sqrt{25+100} = 11.18$

Assignments -

Point x_1 and x_3 are still assigned to cluster 1 (u_1')
Point x_2 and x_4 are still assigned to cluster 2 (u_2')

Convergence -

Since the assignments didn't change after the 2nd iteration, we do not need to update the centers

This means, convergence has been reached.

Clusters -

Cluster 1	Cluster 2
x_1 x_3	x_2 x_4