

This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset L^AT_EX solutions.

1.e

The best value of n I found is 7.

It had a validation error of approximately 0.270 compared to the word extractor with a validation error of approximately 0.275. I believe this is because the average length of words used in reviews is 7, so it mimics the word extractor. Small values of n (e.g., 2) and large values (e.g., 15) had similar performance. This supports my hypothesis that n must be near the average word length to achieve a validation error similar to the word feature extractor.

A movie review where n -grams would likely outperform word features would be:

"I didn't enjoy this movie! I would not recommend this movie. It is not good."

n -grams outperform word features here because they capture context words like "didn't" and "not" that negate the positive words following them.

2.a

Points: [10, 0], [30, 0], [10, 20], [20, 20]

(i)

Iteration 1: Centroids $\mu_1 = [20, 30]$, $\mu_2 = [20, -10]$

- $[10, 0] \rightarrow 2$

$$d_1 = \sqrt{(20-10)^2 + (30-0)^2} \approx 31.62, \quad d_2 = \sqrt{(20-10)^2 + (-10-0)^2} \approx 14.14$$

- $[30, 0] \rightarrow 2$

$$d_1 = \sqrt{(20-30)^2 + (30-0)^2} \approx 31.62, \quad d_2 = \sqrt{(20-30)^2 + (-10-0)^2} \approx 14.14$$

- $[10, 20] \rightarrow 1$

$$d_1 = \sqrt{(20-10)^2 + (30-20)^2} \approx 14.14, \quad d_2 = \sqrt{(20-10)^2 + (-10-20)^2} \approx 31.62$$

- $[20, 20] \rightarrow 1$

$$d_1 = \sqrt{(20-20)^2 + (30-20)^2} = 10, \quad d_2 = \sqrt{(20-20)^2 + (-10-20)^2} = 30$$

Updated centroids:

$$\mu_1 = \left[\frac{10+20}{2}, \frac{20+20}{2} \right] = [15, 20], \quad \mu_2 = \left[\frac{30+10}{2}, \frac{0+0}{2} \right] = [20, 0]$$

Loss:

$$14.14^2 + 14.14^2 + 14.14^2 + 10^2 \approx 699.8188$$

Iteration 2: Centroids $\mu_1 = [15, 20]$, $\mu_2 = [20, 0]$

- $[10, 0] \rightarrow 2$

$$d_1 = \sqrt{(15-10)^2 + (20-0)^2} \approx 20.61, \quad d_2 = \sqrt{(20-10)^2 + (0-0)^2} = 10$$

- $[30, 0] \rightarrow 2$

$$d_1 = \sqrt{(15-30)^2 + (20-0)^2} = 25, \quad d_2 = \sqrt{(20-30)^2 + (0-0)^2} = 10$$

- $[10, 20] \rightarrow 1$

$$d_1 = \sqrt{(15-10)^2 + (20-20)^2} = 5, \quad d_2 = \sqrt{(20-10)^2 + (0-20)^2} \approx 22.36$$

- $[20, 20] \rightarrow 1$

$$d_1 = \sqrt{(15-20)^2 + (20-20)^2} = 5, \quad d_2 = \sqrt{(20-20)^2 + (0-20)^2} = 20$$

Updated centroids:

$$\mu_1 = [15, 20], \quad \mu_2 = [20, 0]$$

Loss:

$$10^2 + 10^2 + 5^2 + 5^2 = 250$$

Iteration 3: Centroids unchanged $\mu_1 = [15, 20]$, $\mu_2 = [20, 0]$

Distances and assignments remain the same.

Loss:

$$10^2 + 10^2 + 5^2 + 5^2 = 250$$

(ii)

Iteration 1: Initial centroids $\mu_1 = [0, 10]$ and $\mu_2 = [30, 20]$

- $[10, 0] \rightarrow 1$

$$d_1 = \sqrt{(0 - 10)^2 + (10 - 0)^2} = 14.14, \quad d_2 = \sqrt{(30 - 10)^2 + (20 - 0)^2} \approx 28.28$$

- $[30, 0] \rightarrow 2$

$$d_1 = \sqrt{(0 - 30)^2 + (10 - 0)^2} \approx 31.62, \quad d_2 = \sqrt{(30 - 30)^2 + (20 - 0)^2} = 20$$

- $[10, 20] \rightarrow 1$

$$d_1 = \sqrt{(0 - 10)^2 + (10 - 20)^2} \approx 14.14, \quad d_2 = \sqrt{(30 - 10)^2 + (20 - 20)^2} = 20$$

- $[20, 20] \rightarrow 2$

$$d_1 = \sqrt{(0 - 20)^2 + (10 - 20)^2} \approx 22.36, \quad d_2 = \sqrt{(30 - 20)^2 + (20 - 20)^2} = 10$$

Updated centroids:

$$\mu_1 = \left[\frac{10 + 10}{2}, \frac{0 + 20}{2} \right] = [10, 10]$$

$$\mu_2 = \left[\frac{30 + 20}{2}, \frac{0 + 20}{2} \right] = [25, 10]$$

Loss:

$$14.14^2 + 20^2 + 14.14^2 + 10^2 \approx 899.8792$$

Iteration 2: Centroids $\mu_1 = [10, 10]$, $\mu_2 = [25, 10]$

- $[10, 0] \rightarrow 1$

$$d_1 = \sqrt{(10 - 10)^2 + (10 - 0)^2} = 10, \quad d_2 = \sqrt{(25 - 10)^2 + (10 - 0)^2} \approx 18.03$$

- $[30, 0] \rightarrow 2$

$$d_1 = \sqrt{(10 - 30)^2 + (10 - 0)^2} \approx 22.36, \quad d_2 = \sqrt{(25 - 30)^2 + (10 - 0)^2} \approx 11.18$$

- $[10, 20] \rightarrow 1$

$$d_1 = \sqrt{(10 - 10)^2 + (10 - 20)^2} = 10, \quad d_2 = \sqrt{(25 - 10)^2 + (10 - 20)^2} \approx 18.03$$

- $[20, 20] \rightarrow 2$

$$d_1 = \sqrt{(10 - 20)^2 + (10 - 20)^2} \approx 14.14, \quad d_2 = \sqrt{(25 - 20)^2 + (10 - 20)^2} \approx 11.18$$

Updated centroids:

$$\mu_1 = \left[\frac{10 + 10}{2}, \frac{0 + 20}{2} \right] = [10, 10]$$

$$\mu_2 = \left[\frac{30 + 20}{2}, \frac{0 + 20}{2} \right] = [25, 10]$$

Loss:

$$10^2 + 11.18^2 + 10^2 + 11.18^2 \approx 449.9848$$

Iteration 3: Same calculations as Iteration 2.**Loss remains:**

$$449.9848$$