**Task 1:**

Please find my notebook with the K-Means Algorithm from scratch along with Euclidean distance, Cosine similarity, and the Generalized Jaccard similarity, here:

https:/

1. After running K-means clustering with Euclidean, Cosine and Jaccard similarity for **K=10**, comparing the SSEs:

|  |  |
| --- | --- |
| **Distance/Similarity** | **SSE** |
| Euclidean means | 435910258473.8313 |
| Cosine K-means | **22624.612885272563** |
| Jaccard K-means | 55504.820261417466 |

**Cosine** is better when we take SSEs into consideration.

1. Labeling each cluster using the majority vote label of the data points in that cluster, let us compare the predictive accuracies:

|  |  |
| --- | --- |
| **Distance/Similarity** | **Predictive accuracy** |
| Euclidean means | 0.6022602260226022 |
| Cosine K-means | **0.6315631563156315** |
| Jaccard K-means | 0.6277627762776278 |

**Cosine** is better when we take the predictive accuracies into consideration.

1. Setting up the same stop criteria:

*when there is no change in centroid position*

*OR*

*when the SSE value increases in the next iteration*

*OR*

*when the maximum preset value (e.g., 500, here ) of iteration is complete,*

for Euclidean-K-means, Cosine-K-means, Jaccard-K-means and comparing the number of iterations and times to converge:

|  |  |  |
| --- | --- | --- |
| **Distance/Similarity** | **Iterations required** | **Time to converge** |
| Euclidean K-means | 2 | 12.04s |
| Cosine K-means | 2 | 29.15s |
| Jaccard K-means | 3 | 29.15s |

1. Comparing the SSEs of Euclidean-K-means, Cosine-K-means, Jaccard-K-means w.r.t the three stop criteria:

* when there is no change in centroid position
* when the SSE value increases in the next iteration
* when the maximum preset value (e.g., 100) of iterations is complete

|  |  |
| --- | --- |
| **Distance/Similarity** | **SSE** |
| * **When there is no change in centroid position** | |
| Euclidean K-means | 438095088569.8631 |
| Cosine K-means | **23233.42737322595** |
| Jaccard K-means | 55504.60587632886 |
| * **when the SSE value increases in the next iteration** | |
| Euclidean K-means | 440691297280.1461 |
| Cosine K-means | **22973.73711875152** |
| Jaccard K-means | 55377.77568657197 |
| * **when the 100 iterations are complete** | |
| Euclidean K-means | 442598304450.5725 |
| Cosine K-means | **23279.368517075323** |
| Jaccard K-means | 55500.95504076842 |

**Task 2:**