**Name (netid):** Minghao Shi (mshi18)

**CS 441 - HW 4: Dealing with Data**

Complete the claimed points and sections below.

**Total Points Claimed [115 ] / 142**

1. Clustering and Fast Retrieval
   1. Test Kmeans Purity & Centroids [ 15 ] / 15
   2. Questions [ 10 ] / 10
   3. Fast 1-NN Retrieval [ 15 ] / 15
2. Estimating PDFs
   1. Histograms [10 ] / 10
   2. Clustering [10 ] / 10
   3. Gaussian Mixture Model [ 15 ] / 15
3. PCA and Data Compression
   1. Display Principal Components [5 ] / 5
   2. Scatter Plot [ 5 ] / 5
   3. Plot cumulative explained variance [ 5] / 5
   4. Time & Accuracy [ 10] / 10
4. Stretch Goals
   1. Rotate Using PCA and comparison

To original approach [ 15] / 15

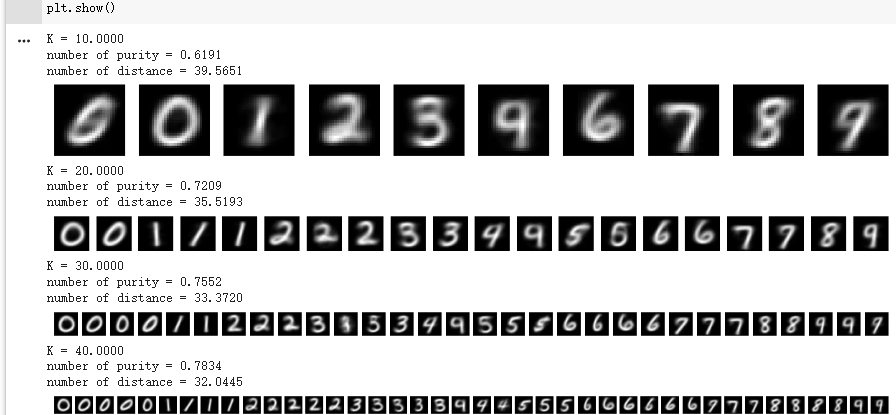
* 1. Try Part 2 with your own images [ ] / 10
  2. Plot Using t-SNE and MDS [ ] / 15
  3. Completed HW3 survey by DATE [ ] / 2

**1. Clustering & Fast Retrieval**

1. Test KMeans Purity [15]

|  |  |
| --- | --- |
| K Vs Purity Plot | K Vs Mean\_Distance Plot |
|  |  |

Paste images of centroids for K = 10, K = 20, and K=30 below (three rows).

****

1. Questions [10]
   1. As you increase K, do you expect the purity to increase? Why or why not?

**When I increase the K from 10 to 90, I observe that the purity increases. However, when K tries to increase from 90 to 100, because of the overfitting, the purity will be slightly decreasing instead of increasing.**

* 1. In a given run, is the average distance of a sample to centroid guaranteed to monotonically decrease with each iteration (i.e. cannot increase)? Why or why not?

**Looking at the result from my code, when K increases from 10 to 100, the number of distances will decrease without increasing. Therefore, based on my code and output, the average distance of a sample to the centroid will be guaranteed to monotonically decrease with each iteration.**

**However, in the general case, since it is possible to converge to a local minimum instead of a global minimum, then it is not guaranteed to monotonically decrease.**

* 1. If you do enough iterations, is Kmeans guaranteed to give you the optimal clustering that minimizes the sum of distances between each sample and its center? Why or why not?

**Doing enough iterations will not guarantee to give you the optimal clustering that minimizes the sum of distances between each sample and its center. The initialization of the centroids will also impact for finding the solution and we need to try different random initialization instead of just doing enough iterations.**

* 1. Does improving the Kmeans objective (i.e. achieving lower mean squared error) necessarily improve expected purity? Why or why not?

**It is not necessarily improving the expected purity since the purity measures how well the clustering matches the true labels and the mean squared error measures the distance between data and their assigned centroids.**

1. Fast Retrievals
   1. Brute Force: [7]

|  |  |  |
| --- | --- | --- |
| Test Error | Time to Add | Time to Search |
| 0.0331 | 0.2531 | 13.8249 |

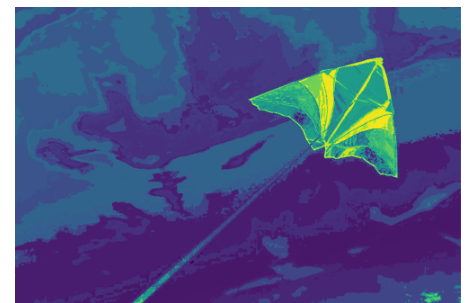
* 1. LSH: [8]

|  |  |  |  |
| --- | --- | --- | --- |
| Test Error | Time to Add | Time to Search | Nbits parameter |
| 0.0318 | 2.7257 | 18.2927 | 1568 |

**2. Estimating PDFs**

Include the generated images (score map, thresholded score map, thresholded RGB) from the display code.

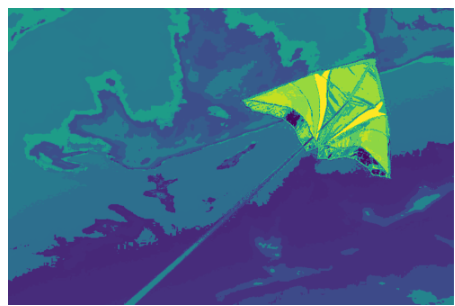
1. Histogram: [10]

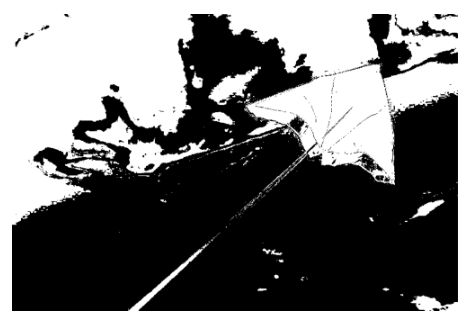


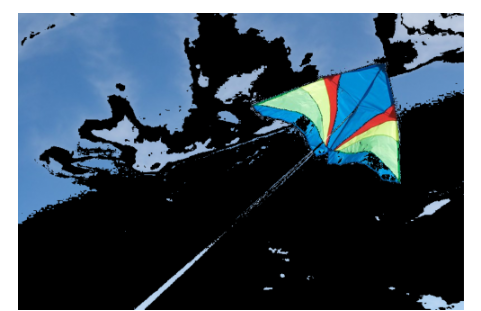




1. Clustering: [10]

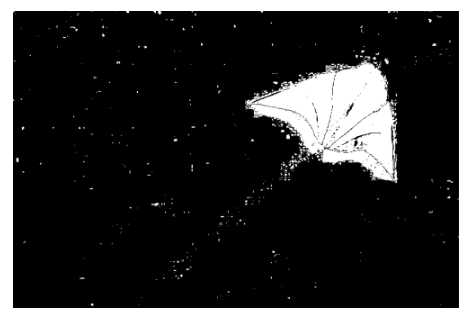


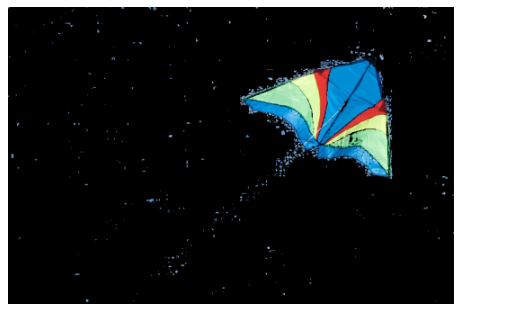




1. Gaussian mixture Model: [15]







**3. PCA and Data Compression**

1. First 10 principal components [5]

|  |
| --- |
| Visualization |
|  |

1. Scatterplot [5]

|  |
| --- |
| PLOT |
|  |

1. Cumulative explained Variance [5]

|  |
| --- |
| PLOT |
|  |

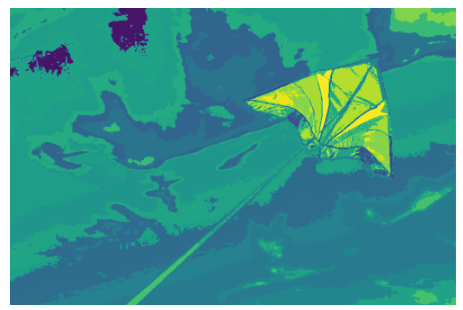
1. Faiss [10]

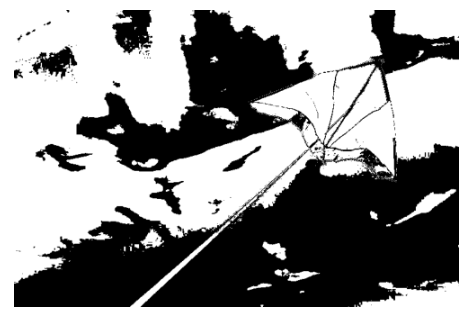
|  |  |  |  |
| --- | --- | --- | --- |
|  | Total Time | Test Error | Dimensions |
| Brute Force (PCA) | 1.902 | 0.0291 | 88 |
| Brute Force | 14.078 | 0.0331 | 784 |
| LSH | 21.0184 | 0.0318 | 784 |

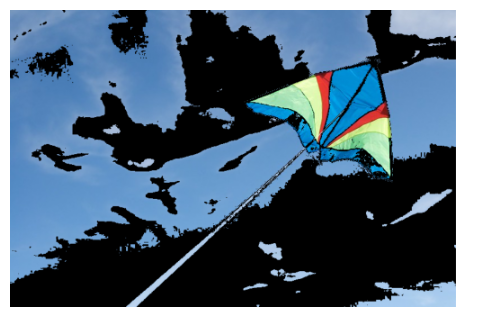
Note: the last two rows are copied from 1.c for reference.

**4. Stretch Goals**

1. PDFS after using PCA to rotate your data [15]







1. Apply Part 2 to your own choice of image, with the same deliverables [10]
2. Scatterplots using at least two of t-SNE, MDS, and Linear Discriminant Analysis [15]

**Acknowledgments / Attribution**

List any outside sources for code or improvement ideas or “None”.