Homework 1 Report

Tom Ward

**Question 1.1**

*(20 points) Use k-means with squared-l2 norm as a metric, for GeorgiaTech.bmp and football.bmp and also choose a third picture of your own to work on. We recommend size of 320 × 240 or smaller. Run your k-means implementation with these pictures, with several different k = 2, 4, 8, 16. How long does it take to converge for each k (report the number of iterations, as well as actual running time)? Please write in your report, and also include the resulted compressed pictures for each k.*

Results are shown below. For some of the higher k runs, the number of clusters was reduced due to the presence of empty clusters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Photo** | **K** | **Total Time (seconds)** | **Number of Iterations** | **Compressed Photo** |
| football.bmp | 2 | 4.47 | 6 |  |
| GeorgiaTech.bmp | 2 | 1.84 | 4 |  |
| tom.bmp | 2 | 1.14 | 7 | *A picture containing text  Description automatically generated* |
| football.bmp | 4 | 13.94 | 18 |  |
| GeorgiaTech.bmp | 4 | 17.81 | 29 |  |
| tom.bmp | 4 | 2.52 | 18 |  |
| football.bmp | 8 | 28.11 | 38 |  |
| GeorgiaTech.bmp | 8 | 14.44 | 29 |  |
| tom.bmp | 8 | 1.97 | 18 |  |
| football.bmp | 16 (only ran with k=10 due to empty clusters) | 33.46 | 47 |  |
| GeorgiaTech.bmp | 16 (only ran with k = 11 due to empty clusters) | 14.91 | 32 |  |
| tom.bmp | 16 (only ran with k=10 due to empty clusters) | 2.26 | 19 |  |

**Question 1.2**

*(20 points) Run your k-means implementation (with squared-l2 norm) with different initialization centroids. Please test two initialization strategies, compare the results (output image, running time, iterations) and report: (i) random initialization. Please try multiple time and report the best one (in terms of the image quality). (ii) poor initialization. Please design your own strategy, explain why it qualifies as a poor initialization, try multiple times, and report the results.*

*How does this it affect your final result? (We usually randomize initial location of centroids in general.) Please also explain in the report how you initialize the centroid.*

I chose randomly assigning each centroid a value very close to black (all RGB values between 0 and 5) or white (all RGB values between 250 and 255) as my “poor initialization” strategy. These values ensure that there will be a high initial distance between points and centroids and ensure that clusters will somewhat similar initially.

The results from my tests of 5 runs of each strategy with k=2, k=4, and k=8 are shown below. The final result of the image quality did not change based on the initialization strategy, but as expected, the “poor” strategy took longer to converge, both in total running and number of iterations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Strategy** | **K** | **Average Total Time (seconds)** | **Average Number of Iterations** | **Image Quality** |
| Random | 2 | 13.52 | 8 | Standard |
| Poor | 2 | 26.63 | 18 | Standard |
| Random | 4 | 8.47 | 10 | Standard |
| Poor | 4 | 8.94 | 17 | Standard |
| Random | 8 | 17.27 | 15 | Standard |
| Poor | 8 | 26.67 | 22 | Standard |

**Question 1.3**

*(20 points) Now try your k-means with the Manhattan distance (or l1 distance) and repeat the same steps in Part (1). Please note that the assignment of data point should be based on the Manhattan distance, and the cluster centroid (by minimizing the sum of deviance – as a result o fusing the Manhattan distance) will be taken as the “median” of each cluster. Comment on the difference of image compression results using the two methods.*

Using Manhattan distance instead of Euclidean distance produced slightly faster results on average, both in terms of total time and number of iterations. However, results vary for different runs. In general, the compressed image quality seems to be about the same for the 2 distance methods.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Photo** | **K** | **Total Time (seconds)** | **Number of Iterations** | **Compressed Photo** |
| football.bmp | 2 | 4.95 | 7 |  |
| GeorgiaTech.bmp | 2 | 2.57 | 6 |  |
| tom.bmp | 2 | 0.85 | 8 |  |
| football.bmp | 4 | 13.14 | 19 |  |
| GeorgiaTech.bmp | 4 | 7.89 | 18 |  |
| tom.bmp | 4 | 0.97 | 9 |  |
| football.bmp | 8 | 7.96 | 11 |  |
| GeorgiaTech.bmp | 8 | 7.87 | 18 |  |
| tom.bmp | 8 | 2.36 | 22 |  |
| football.bmp | 16 (only ran with k=15 due to empty clusters) | 23.00 | 32 |  |
| GeorgiaTech.bmp | 16 (only ran with k = 10 due to empty clusters) | 16.86 | 38 |  |
| tom.bmp | 16 (only ran with k=13 due to empty clusters) | 2.65 | 24 |  |

**Question 2.1**

*(10 points) Write down the graph Laplacian matrix and find the eigenvectors associated with the zero eigenvalue. Explain how do you find out the number of disconnected clusters in graph and identify these disconnected clusters using these eigenvectors.*

The graph Laplacian matrix, L:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2 | -1 | -1 | 0 | 0 |
| -1 | 2 | -1 | 0 | 0 |
| -1 | -1 | 2 | 0 | 0 |
| 0 | 0 | 0 | 1 | -1 |
| 0 | 0 | 0 | -1 | 1 |

The 2 eigenvectors associated with the 0 value eigenvalues are

|  |  |
| --- | --- |
| -0.577 | 0 |
| -0.577 | 0 |
| -0.577 | 0 |
| 0 | 0.707 |
| 0 | 0.707 |

The 2 eigenvectors associated with the 0 value eigenvalues show 2 clusters: 1,2,3 and 4,5.

**Question 2.2**

*(10 points) For the graph Laplacian matrix and perform eigedecomposition on it. Plot the eigenvalues (ranked from the largest to the smallest), and based on the plot explain approximately how many clusters you believe there are and why.*

Plot of eigenvalues, largest to smallest:

A picture containing chart

Description automatically generated

Based on the eigenvalues plot, I believe there are 21 clusters. This is the number of eigenvalues that appear higher than 0.4 on a slope.

**Question 2.3**

*(15 points) Now perform spectral clustering, using k = 5, k = 7, k = 10, and your choice of k (based on your answer in Part (2). Report the size of the largest cluster and the smallest cluster based on your result for each k. Report the results for k = 10 by listing the teams and which cluster they belong to (you can either include a table in your solution, or upload a file or spreadsheet that include your result).*

Smallest and largest cluster sizes based on different values of k:

|  |  |  |
| --- | --- | --- |
| **k** | **Smallest Cluster Size** | **Largest Cluster Size** |
| 5 | 36 | 106 |
| 7 | 27 | 104 |
| 10 | 18 | 39 |
| 21 | 11 | 22 |

For a list of the cluster assignments for each team for a run of k=10, see accompanying file question\_2-3\_table.xlsx.

**Question 2.4**

*(5 points) Now run the algorithm a few times for k = 10. You may notice the results are slightly different - please explain why. Also check which clusters that “Georgia Tech”, “Georgia State”, and “Georgia” (UGA) are in. Are they always in the same cluster or not - please explain your reasoning.*

The clusters are slightly different due to the initial randomization of the k-means algorithm. K-means is not deterministic and only converges on a “local” optimum solution. In 3 runs of the k=10 algorithm, Georgia Tech, Georgia, and Georgia State sometimes appeared in the same clusters and sometimes appeared in different clusters. Each run varied. We would not always expect the 3 schools to appear in the same cluster from run to run because k-means clustering is a heuristic algorithm that converges to a local optimum.