**Tamas Palfi**

**5/3/2019**

**Homework 5: Unsupervised Learning**

The elbow rule is one of the ways to find the “optimal number” of clusters for K-means unsupervised learning algorithm. It works by visually plotting the result of the objective function for K-means as a function with differing k as inputs. The objective metric for K-means is the RSS or average Euclidean distance between each point and centroid of the cluster, and the resulting graph is a plot of number of clusters versus a percentage metric of variance. The rule works by looking at this resulting plot and taking the k-value where there is a visual “elbow” or bend in the curve in which adding more value to k does not decrease variance significantly.

An issue with the k-means algorithm is that it can often lead to “poor” clusters due to the random initialization of centroids. K-means++ fixes this by only allocating the first centroid to be selected randomly. From there, it improves accuracy by actively selecting the next best centroid due to a metric which usually is based on computing the distances between each point and the first centroid, so some relationship is established and isn’t purely random selection. This process is repeated until k centroids have been made which will later allow for k clusters. This is more effective at reducing error due to how purely random selection hasn’t been allocated for each centroid but rather we are relying on a metric based on point distances.

**Pictures Representation of Results with different k values:**

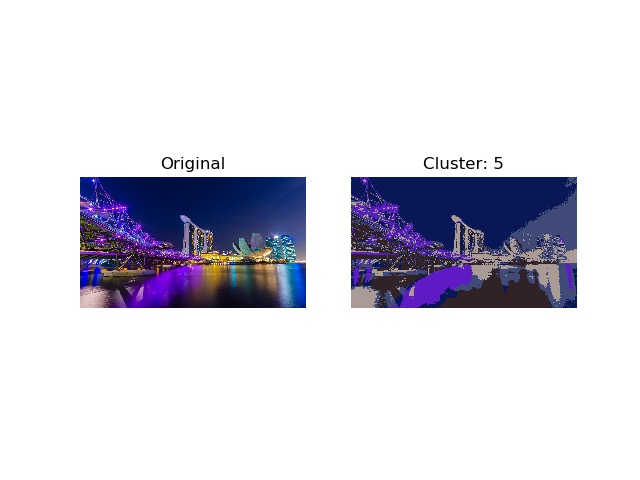
Original:



K = 2



K = 5



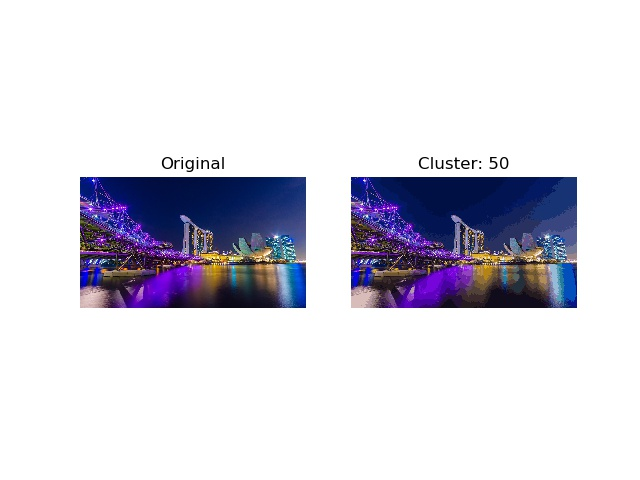
K = 10



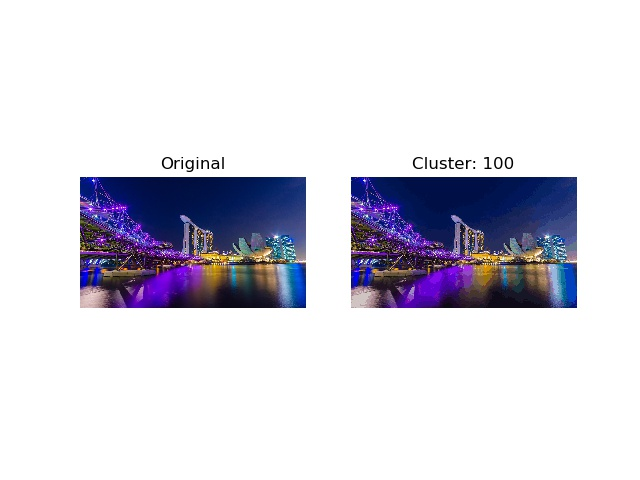
K = 25



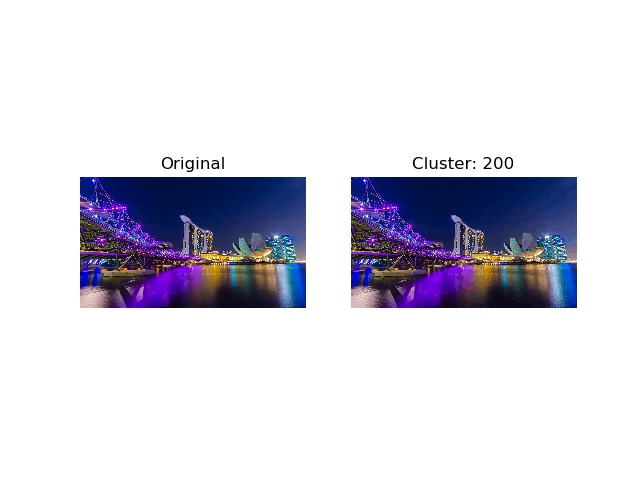
K = 50



K = 100



K = 200



**Table on Relationship between K value and Reconstruction Error (MSE)**

|  |  |
| --- | --- |
| **K** | **Reconstruction Error (MSE)** |
| 2 | 57.88496184277203 |
| 5 | 40.97827233218619 |
| 10 | 30.066489389552572 |
| 25 | 20.620469597743412 |
| 50 | 15.79494050368481 |
| 100 | 11.901492503658112 |
| 200 | 9.24072866805806 |

**Table on Relationship between K value and Compression Rate**

|  |  |
| --- | --- |
| **K** | **Compression Rate** |
| 2 | 95.83183990442055 |
| 5 | 90.32156603235403 |
| 10 | 86.1511657934054 |
| 25 | 80.63193134786218 |
| 50 | 76.44659681978571 |
| 100 | 72.24259443029946 |
| 200 | 68.00125631799361 |