Data Mining & Machine Learning

Computer Exercise 11 - K-means

Steinarr Hrafn Höskuldsson

October 2022

Section 1.6

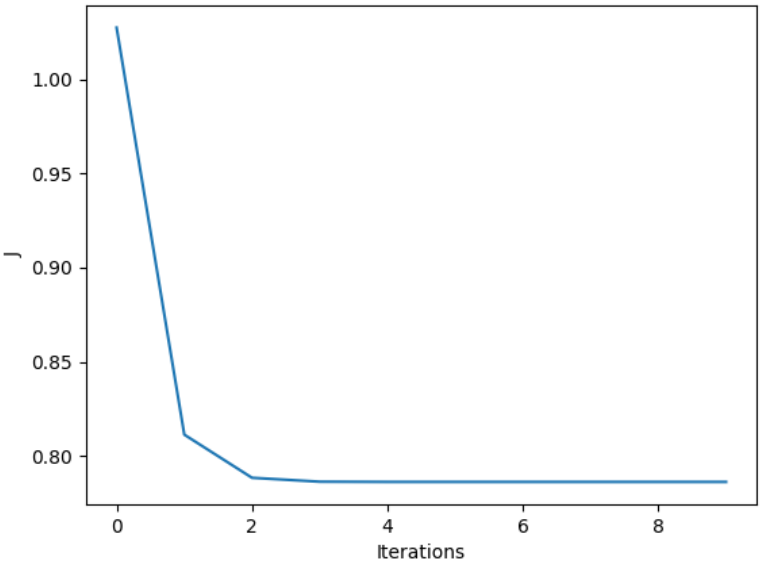

Figure 1: J plotted by iterations

Section 1.7

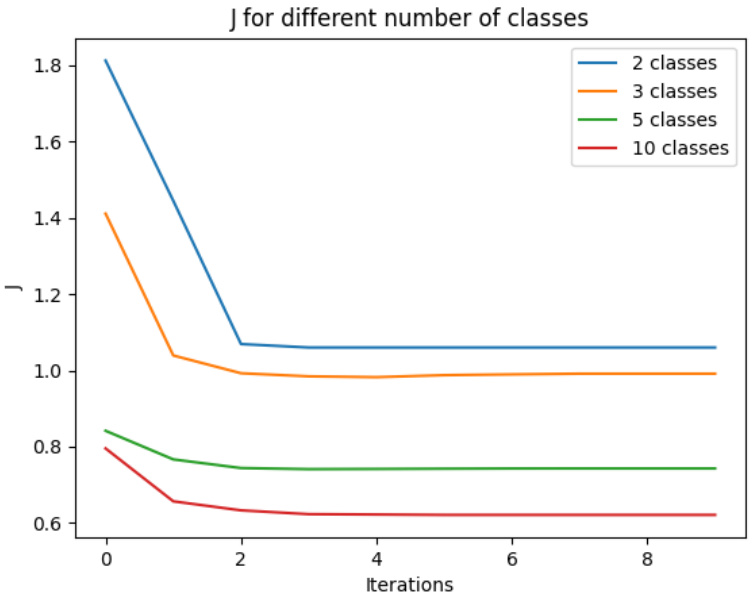

Figure 2: J plotted by iterations for different number of classes

Section 1.8

On Figure 2 we can see that using 10 classes better minimizes the objective function J.

The value of the objective funstion is calculated as the average distance from each datapoint to its classification prototype. And it gets smaller as we increase the classes. That is not because we are getting better results but rather simply because more prototypes occupy the same limited feature space thus decreasing the average distance.

If we were to set $k = n$, each prototype would be initialized as a unique data point and of course none of them would change since they are all closer to themselves rather then any other point. Choosing $k = n$ would however result in $J = 0$ but is not a good strategy since it would result in each point classified in its own class which is, frankly, an utterly useless classification.

Section 1.10

Training the K-means model with $k = 3$ results in Accuracy= 82.67% and Confusion matrix:

$$\text{CM} = \begin{bmatrix} 50 & 0 & 0 \\ 0 & 42 & 8 \\ 0 & 18 & 32 \end{bmatrix}$$

Section 2.1

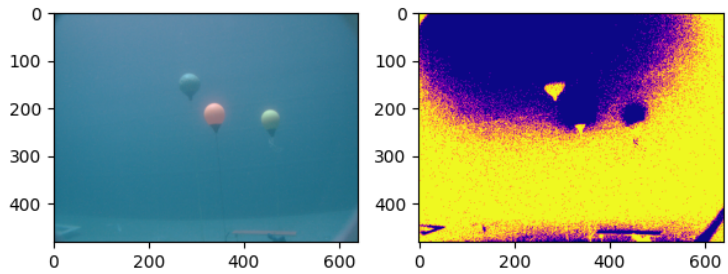

Figure 3: K means applied to cluster an image based on color, using k=2

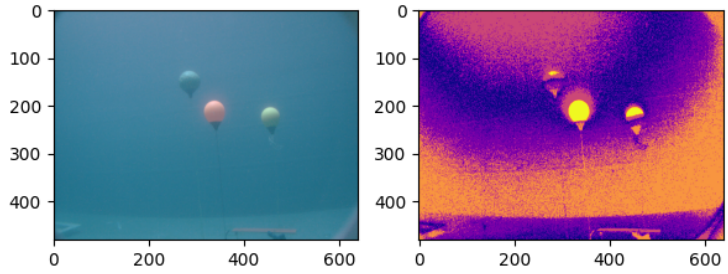

Figure 4: K means applied to cluster an image based on color, using k=5

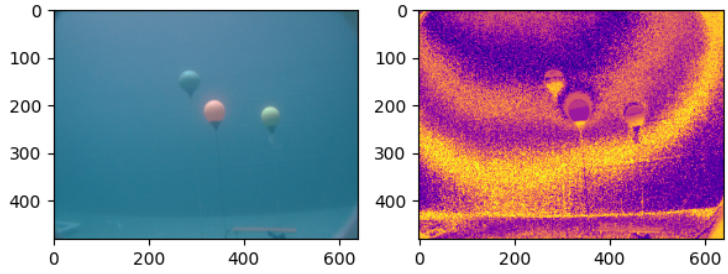

Figure 5: K means applied to cluster an image based on color, using k=10

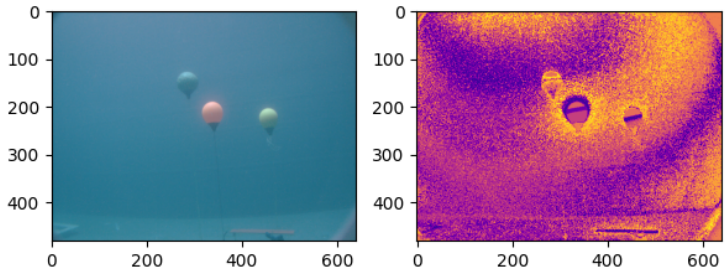

Figure 6: K means applied to cluster an image based on color, using k=20