COSC 425: Intro to Machine Learning
Lab 2: K-Means

Due: October 14, 2022, 11:59 PM

Introduction

In this lab, you will be practicing using K-means clustering and applying it to a dataset. Here, the data will a variety of images, all of which are available for download on Canvas. We will be working with JPG images. Think of each JPG image as a set of data points with three features each (the RGB values of each pixel). For each image, you will use K-means clustering to cluster the pixels to some number of cluster points which will be used as the best set of RGB values to represent the image.

You will create a Jupyter notebook that you will submit where you will put the code that you wrote for this lab. You will also submit a separate written report in PDF form in which you include the answers to the questions and/or plots required by the question.

You will need to install and use cv2 to read the jpg image from the files. You should use "pip install opencv-python" to install.

Additionally, you can use the following code to read in and visualize the data:

```
# Read in the image
image = cv2.imread('image_name.jpeg')

# Change color to RGB (from BGR)
image = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)

plt.imshow(image)
```

There are four images that are provided: checkerneyland.jpeg, ayreshall.jpeg, minkao.jpeg, and smokey.jpeg.

Questions

Question 1 (40 points): Create a function that takes as input the image file name and the number of colors to find (clusters to find) and returns the updated image based on the new clusters. (Note: Your function can also return other elements as needed for plotting.) This function will use the scikit-learn KMeans function to find the cluster centers and which pixels belong to which clusters. You will then create an updated image that replaces each

pixel value with its corresponding cluster center value. Note that the scikit-learn cluster centers may not be integers. You should convert the cluster centers to integers using floor(). We recommend setting the random_state option in K-Means function to a fixed value for repeatable results.

Question 2 (40 points – 10 for each image): For each of the four images, try K=4, K=8, and K=16 and visualize the updated image with the new colors based on the cluster centers. You should include a 2x2 figure for each image that shows the original image and the three updated images for the different values of K. Include a brief discussion of which images look acceptable (capturing the details of the original image) and which ones do not. For those that do not look acceptable, comment in the report on why they do not look acceptable (based on which colors were selected as cluster centers).

Question 3 (20 points – 5 for each image): For each image, create three histogram plots (one for each value of K) showing the distribution of pixel values across the different cluster centers (how many pixels in the image below to each cluster).

Extra Credit (5 points): Color the bars of the histograms for each cluster with the associated color for that cluster.

Submission Checklist

In this lab, you will submit on Canvas:

- Jupyter notebook
- PDF of your report

Late Penalty: 20 points off per day late.