CS 550 -- Machine Learning
Homework #3
Due: 23:55, December 21, 2020

PART 1:

Implement the k-means algorithm to cluster the image pixels. Your algorithm should take the pixels of an RGB image, group the pixels into k clusters, and output the clustering vectors as well as the labels of the clustered pixels. In the RGB color space, each pixel is represented with 3 values, which are in between 0 and 255.

To test your algorithm, use the image (sample.jpg) provided on the course website. First run your implementation for the following values of k = {2, 3, 4, 5, 6}. For each of these k values,

- Report the clustering error (averaged over all pixels)
- Report the computational time
- Give the values of the clustering vectors (use only two digits after the decimal point)
- Give the clustered image. To obtain this clustered image, represent each pixel with the RGB values of its clustering vector.

Then, using your implementation, determine the optimum k value for this image. For that, do NOT just consider the k values given above (i.e., do NOT just consider k = {2, 3, 4, 5, 6}). Explain how you determine the optimum value. Moreover, for your optimum k value, report the clustering error and give the clustered image (but not the computational time and the values of the clustering vectros), as explained above.

In your report, give all details of your implementation, such as how to select the initial values and the stopping criterion. Your report should also include the aforementioned outputs.

In this assignment, you may use any programming language you would like. If you do not know how to process images in your preferred programming language, you can employ the following Matlab codes. Of course, if you do not want to use the following codes, that is completely ok as long as you process the image pixels and give the required outputs.

```
% Reads sample.jpg into an im matrix, whose dimensions are N, M, and 3
im = imread('sample.jpg');

% Reshapes the 3D im matrix into a 2D matrix, called im2, whose
% dimensions are NxM and 3. In this 2D matrix, each row corresponds to
% a pixel, and the columns correspond to red, green, and blue channels
% in the RGB color space, respectively
im2 = reshape(im, [size(im,1)*size(im,2) 3]);

% You need to run the k-means algorithm on the im2 matrix. Your
% k-means algorithm should output the clustering vectors (let's call
% them V) and the labels of the clustered pixels (let's call them cmap).
% To implement the k-means algorithm, you can use Matlab. Of course, you
% have to write your own code; you cannot use the kmeans built-in
% function of Matlab. However, if you find Matlab too slow or if you
% prefer using another programming language (but if you do not know how
% to read images in your preferred programming language), you can write
% the contents of im2 into a file and read this file in your program.
% Similarly, you can calculate clustering vectors V and labels cmap in
% your program, write them into files, and read them from Matlab to
% obtain the clustered image.
```

```
% Suppose that your program outputs matrix V, whose dimensions are k
% and 3, and vector cmap, whose dimension is NxM. Also suppose that
% your labels are in between 1 and k. Then you may use the following
% Matlab codes to produce the clustered image
cmap2 = reshape(cmap, [size(im,1) size(im,2)]);
M = V / 255;
clusteredImage = label2rgb(cmap2, M);

% Shows the clustered image in Matlab and writes it into a bitmap file
figure, imshow(clusteredImage)
imwrite(clusteredImage,'output.bmp')
```

PART 2:

Implement an agglomerative hierarchical clustering algorithm to cluster the image pixels. Similarly, your algorithm should take the pixels of an RGB image, group the pixels into k clusters using this agglomerative algorithm, and output the clustering vectors as well as the labels of the clustered pixels.

The computational time of this part could be high due to the number of pixels in the image. Thus, propose a technique to overcome this problem. For example, you may "somehow" form small groups of pixels and consider each of these groups as the initial clusters of your agglomerative algorithm. However, downsampling an image into a lower resolution and running your algorithm on this downsampled image will NOT be accepted as a solution. Do not forget that there is not only one correct solution for this part. Thus, try to consider different alternatives.

Test your algorithm on the same image (sample.jpg) and obtain the results for different values of k. Prepare a report for this second part similar to Part 1, also addressing the same questions given in the first part. Additionally, for this second part, give the details of your technique that you will propose to overcome the computational time problem.

WHAT TO SUBMIT:

This homework asks you to implement the k-means algorithm and an agglomerative hierarchical clustering algorithm by **writing your own codes**. Thus, you are not allowed using any machine learning package. In your implementation, you may use any programming language you would like.

Prepare your report for the first and the second part as explained above. The last part of your report should include a brief comparison of the results that you will obtain in Part 1 and Part 2. Similar to the previous assignments, prepare your report neatly and properly. Your report should be a maximum of 4 pages.

Please email the pdf of your report and the source code of your implementation before the deadline. The subject line of your email should CS 550: HW3.