CS 506 Spring 2020 - HW1

Clustering and Visualization

Due: March 2, 2020

1 Working with the Algorithms

- In this assignment we will be working with the AirBnB dataset, that you can also find here. Our goal is to visualize areas of the NYC with respect to the price of the AirBnb listings in those areas.
  From the detailed *nyc_listings.csv* file, you will use **longitude** and **latitude** to cluster closeness and **price** to cluster for expensiveness.
  Note that spatial coordinates and price are in different units, so **you may need to consider scaling** in order to avoid arbitrary skewed results.

a) [8 pts.] **Find clusters using the 3 different techniques we discussed in class: k-means++, hierarchical, and GMM**. Explain your data representation and how you determined certain parameters (for example, the number of clusters in k-means++).

b) [3 pts.] List a few bullet points describing the pros and cons of the various clustering algorithms.

A few hints:
-Some listings contain missing values. Better strategy for this assignment is to completely ignore those listings.
-Pay attention to the data type of every column when you read a .csv file and convert them to the appropriate types (e.g. float or integer).

2 Data visualization

a) [1pt.] Start by producing a Heatmap using the Folium package (you can install it using pip). You can use the code below to help you (assumes the use of Pandas Dataframes):

```
def generateBaseMap(default_location=[40.693943,
    -73.985880]):
    base_map = folium.Map(location=default_location)
    return base_map
```

```
base_map = generateBaseMap()
HeatMap(data=df[['latitude', 'longitude', 'price']].
    groupby(['latitude', 'longitude']).mean().
    reset_index().values.tolist(), radius=8, max_zoom
    =13).add_to(base_map)
base_map.save('index.html')
```

Is this heatmap useful in order to draw conclusions about the expressiveness of areas within NYC? If not, why?

b) [2pts.] Visualize the clusters by plotting the longitude/latitude of every listing in a scatter plot.

c) [2pts.] For every cluster report the average price of the listings within this cluster.

d) Bonus points [1pt.] if you provide a plot on an actual NYC map! You may use Folium or any other package for this.

e) [1pt.] Are the findings in agreement with what you have in mind about the cost of living for neighborhoods in NYC? If you are unfamiliar with NYC, you can consult the web.

3 Image Manipulation

a) [8 pts.] Download the image found by clicking **here.** For this assignment, you will use the k-means algorithm in the CS506 python package that you built in class to manipulate this image. The goal is to give this image as input, and return the image with like pixels combined into 10 clusters.

A few hints:

-There are a number of useful packages for working with images; we recommend using cv2 (obtained by running pip install opencv). Using this package, you can use the line **img = cv2.imread("file.jpg")** in order to load the image as a numpy array (note that this means you will also need to import numpy).

-If you follow the hint above, your data is no longer being opened from a file inside your k_mean() function so you may need to tweak it a bit.

-To display the image after you have run k-means, you can use the lines

cv2.imshow('Display Window', manipulated_img)

cv2.waitKey(0)

cv2.destroyAllWindows()

-Each pixel is represented by three features: their red, green, and blue values. You only need to tweak your algorithm to find clusters and then replace the pixels with the value of their cluster center.

-The more clusters you work with, the slower this algorithm will run, so for testing it may be useful to work with only 2 clusters.

Here is the starting image:

And here is what your code should return: