CMPUT 307 Lab-7

Kmeans Clustering

Submission Deadline: **March 21**

You are NOT allowed to import any other package to finish this lab. Please only submit your own work and give credit to any material you referenced.

PART 1

Implement the kmeans algorithm, test your algorithm on the given data (available on eclass), and print out the result. Note: use **vectorized operation** for better efficiency.

a) Given n 2D data points (array of size (n, 2)) and m 2D cluster centers (array of size (m, 2)), calculate the distances from those n points to m centers.

```
distance(data, centers): Calculate the distances from points to
cluster centers.
```

b) Finish the implementation of kmeans.

```
kmeans(data, n_centers): Divide data into n_centers clusters, return
the cluster centers and assignments.
```

c) The quality of the clustering result can be calculated using the total sum of squared errors (distortion). You can find the formula for this in page 54 of the point cloud slides. Implement the function for calculating distortion.

```
distortion(data, centers, assignments): Calculate the distortion of
the clustering.
```

PART 2

Answer the following questions.

1. What's the minimum and maximum number of possible clusters when you have n points?
2. Experiment with different numbers of clusters, from minimum to maximum. Record the distortions while you change the parameter for kmeans. Draw a graph to show how distortion changes with different numbers of clusters.
3. What's the lowest possible distortion? When will this happen? Explain your answers.
4. What's the optimal number of clusters? How can we get this number in a program (without any human intervention/interaction)?

Submit the following via eClass (as separate files, **DON'T** zip them):
1. lab7.py: Part 1 Code
2. lab7.pdf: Part 2 Answers
Points will be deducted if the submitted filename and format doesn't match the requirement.

Grading:

- Part 1: 60%
    - Code Quality: 5%
    - Comments: 5%
    - distance(): 10%
    - kmeans(): 30%
    - distortion: 10%
- Part 2: 40%, 5% + 5% + 10% + 20%