CS 4700, Foundations of Artificial Intelligence
Spring 2020
Solutions to Quiz 10

1. There were two possible questions, but they both have the same basic answer:
   a. If there are N data points and k=N then k-means clustering will end after its first iteration. Assume that the data are all points in R^n, distance is traditional Euclidean distance, and each point is distinct - there are no duplicates.
   b. If there are N data points and k=N then k-means clustering will end with clusters that each contain exactly one distinct data point. Assume that the data are all points in R^n, distance is traditional Euclidean distance, and each point is distinct - there are no duplicates.

   True. Each initial centroid will be a different one of the N examples. The set corresponding to each will just contain that example. After the first iteration this won't have changed, so the algorithm halts.

2. There were two possible questions:
   a. Consider a problem where data are three-character strings over the uppercase letters A-Z - for example, "CAT" and "DOG". The distance between two strings is the sum of their letter-by-letter distance in the alphabet. In other words, "CAT" can be thought of as "3-1-20" - C being the third letter of the alphabet, A the first, and T the 20th - and similarly "DOG" would be "4-15-7" and their distance apart is |3-4|+|1-15|+|20-7|=1+14+13=28|3−4|+|1−15|+|20−7|=1+14+13=28. You are given a dataset with 4 such strings:
      A. DOG
      B. CAT
      C. RAT
      D. PIG

      You are to do k-means clustering on these 4 examples, where k=2 and the initial two centroids are B (CAT) and C (RAT). Which examples wind up in each of the resulting clusters? Please write your answer in the following form:
      - List the labels (A, B, C, or D) for the points in each cluster in alphabetically increasing order, without any spaces or punctuation. For example, if a cluster were to have all four data points A, B, C, and D in it you would write it as the string ABCD - listing the example labels in alphabetically increasing order, concatenated with no spaces or punctuation marks.
      - List the clusters in alphabetically increasing order, each separated by a single space, with no extra spaces or punctuation. For example, if you were doing clustering with k=3 and the resulting clusters contained {A}, {B,C}, and {D} you would write this as "A BC D" where there is a single space between A and BC and another single space between BC and D.
      - Only refer to each data point by its label (A, B, C, or D) and not by the string it represents. Answers should thus be comprised only of the letters A-D and the necessary number of spaces.

      The final clusters are: AB CD

      To make this easier let's switch the four examples from letters to their position in the alphabet:
      A. 4-14-7
      B. 3-1-20
      C. 18-1-20
      D. 16-9-7

      If the initial centroids are B and C, we need to know which cluster A and D wind up in.
      - A: Distance from B = 1+13+13 = 27 Distance from C = 14+13+13 = 40 So it goes with B
      - D: Distance from B = 13+8+13 = 34 Distance from C = 2+8+13 = 23 So it goes with C

      This means we have a cluster with A and B and another with C and D. We compute the new centroids:
      - Centroid for cluster 1 (containing A&B): 1.5-7.5-13.5
      - Centroid for cluster 2 (containing C&D): 17-5-13.5

We now assign each example to the two clusters once again (closer centroid shown in red):

Example	Distance to 3.5-7.5-13.5	Distance to 17-5-13.5
A	13.5	28.5
B	13.5	24.5
C	27.5	11.5
D	20.5	11.5

As a result A and B stay in the first cluster and C and D stay in the second cluster. On the next iteration the centroids stay unchanged so the algorithm halts.

b. Consider a problem where data are three-character strings over the uppercase letters A-Z - for example, "ADA" and "HAL". The distance between two strings is the sum of their letter-by-letter distance in the alphabet. In other words, "ADA" can be thought of as "1-4-1" - A being the first letter of the alphabet, D the fourth, and A the first - and similarly "HAL" would be "8-1-12" and their distance apart is |1-8|+|4-1|+|1-12|=7+3+11=21|1−8|+|4−1|+|1−12|=7+3+11=21. You are given a dataset with 4 such strings:

A. ADA
B. HAL
C. VAL
D. TED

You are to do k-means clustering on these 4 examples, where k=2 and the initial two centroids are B (HAL) and C (VAL). Which examples wind up in each of the resulting clusters? Please write your answer in the following form:

- List the labels (A, B, C, or D) for the points in each cluster in alphabetically increasing order, without any spaces or punctuation. For example, if a cluster were to have all four data points A, B, C, and D in it you would write it as the string ABCD - listing the example labels in alphabetically increasing order, concatenated with no spaces or punctuation marks.
- List the clusters in alphabetically increasing order, each separated by a single space, with no extra spaces or punctuation. For example, if you were doing clustering with k=3 and the resulting clusters contained {A}, {B,C}, and {D} you would write this as "A BC D" where there is a single space between A and BC and another single space between BC and D.
- Only refer to each data point by its label (A, B, C, or D) and not by the string it represents. Answers should thus be comprised only of the letters A-D and the necessary number of spaces.

The final clusters are: AB CD

To make this easier let's switch the four examples from letters to their position in the alphabet:

A. 1-4-1
B. 8-1-12
C. 22-1-12
D. 20-5-4

If the initial centroids are B and C, we need to know which cluster A and D wind up in.

- A: Distance from B = 7+3+11 = 21 Distance from C = 21+3+11 = 35 So it goes with B
- D: Distance from B = 12+4+8 = 24 Distance from C = 2+4+8 = 14 So it goes with C

This means we have a cluster with A and B and another with C and D. We compute the new centroids:

- Centroid for cluster 1 (containing A&B): 4.5-2.5-6.5
- Centroid for cluster 2 (containing C&D): 21-3-8

We now assign each example to the two clusters once again (closer centroid shown in red):

Example	Distance to 4.5-2.5-6.5	Distance to 21-3-8
A	10.5	28

B	10.5	19
C	24.5	7
D	20.5	7

As a result A and B stay in the first cluster and C and D stay in the second cluster. On the next iteration the centroids stay unchanged so the algorithm halts.