DS250: Homework 3 (Clustering)

Q1. [10 marks] Let $\{a_1, a_2, \ldots . a_n\}$ be a set of points. Prove that the sum of squared distances of a_i's to a point x is minimized when x is the centroid, namely x = $= \frac{1}{\mathrm{n}} \Sigma_i a_i$
Using this result, prove that the k-means cost function decreases monotonically by the k-means clustering algorithm and that the clustering always converges to a local minimum.

Q2. [10 marks] Show that in 1-dimension, the center of a cluster that minimizes the sum of distances of data points to the center is in general not unique. Suppose we now require the center also to be a data point; then show that it is the median element (not the mean).

Hint: Begin with a cost function for the median as the center and show that if we choose any other point the cost will only increase.

Q3. [10 marks] For the k-median problem, show that there is at most a factor of two ratio between the optimal value when we either require all cluster centers to be data points or allow arbitrary points to be centers.

Hint: Use the triangle inequality to show that the sum of distances from a suitable point is smaller than Twice the Sum of Distances from the optimal cluster center.

Q4. [10 marks] For the k-means problem, show that there is at most a factor of four ratio between the optimal value when we either require all cluster centers to be data points or allow arbitrary points to be centers.

Hint: Use the triangle inequality to show that the sum of square of distances from a suitable point is smaller than Four times the Sum of Distances from the optimal cluster center.

Q5. [10 marks] Consider clustering points in the plane according to the k-median criterion, where cluster centers are required to be data points. Enumerate all possible clustering's and select the one with the minimum cost. The number of possible ways of labeling n points, each with a label from {1, 2, …, k} is k^n; which is prohibitive. Show that we can find the optimal clustering in time at most a constant time n*C (n, k).
Bonus [10 marks]: Can this be reduced further? What is the new time complexity?

Hint: Given a possible clustering, how many computations do we need to compute its cost? Can we use dynamic programming for solving this problem more efficiently?

Q6. [10 marks] Suppose in the previous exercise, we allow any point in space (not necessarily data points) to be cluster centers. Show that the optimal clustering may be found in time at most a constant time n^{2k^2}.

Q7. [10 marks] Suppose S is a finite set of points in space with centroid $\mu(S)$. If a set T of points is added to S, show that the centroid $\mu(S \cup T)$ of $(S \cup T)$ is at distance at most $\frac{|T|}{|S|+|T|}|\mu(T) - \mu(S)|$ from $\mu(S)$.

Q8. [15 marks] Given 10 two-dimensional data points
D= $\big((1,4),(1,5),(2,6),(3,5),(5,2),(8,2),(8,3),(9,1),(9,2),(9,3)\big)$. Please cluster using:

i. K-means clustering: 2 clusters beginning with datapoints, (9,1) and (8,3) as initial centers.
ii. Hierarchical clustering: Use Ward's method where average distance between pairs of points is used as the measure of cluster distances.
iii. Density based clustering: minpts=2; epsilon=$\sqrt{(2)}$

Q9. [15 marks] Three laptops, A, B, and C, have the numerical features listed below:

Feature	A	B	C
Processor Speed (GHz)	3.06	2.68	2.92
Disk	512	256	1024
RAM	8	4	16

We may imagine these values as defining a vector for each laptop; for instance, A's vector is [3.06, 512, 8]. We can compute the cosine distance between any two of the vectors, but if we do not scale the components, then the disk size will dominate and make differences in the other components essentially invisible. Normalize the data in an appropriate way.
Suppose a user have given the following ratings
Consider the following ratings given by users to the laptops.

Ratings	U1	U2	U3	U4
A	5	4		4
B		2	4	5
C	4		3	

[5 marks] Compute user profiles for each user.

[5 marks] Estimate the missing ratings in the data above using content-based method.

[5 marks] Estimate missing ratings using collaborative filtering method.