

UNIVERSITY OF TEXAS AT AUSTIN

Homework Assignment 7K–Means clustering.

Please, provide your **complete solutions** to the following problems. Final answers only, even if correct will earn zero points for those problems.

Problem 7.1. (10 points) Provide an example of when **clustering** would be useful in **actuarial practice**.

Solution: Solutions will vary. One example is grouping of policies into cells.

Problem 7.2. (20 points) As you know from class, in K –means clustering, our objective is to minimize

$$\sum_{k=1}^K \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

The above is a difficult formula to compute with, but there is an alternative called the *centroid formula*:

$$2 \sum_{k=1}^K \sum_{i \in C_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2$$

where

$$\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$$

is the j^{th} component of the centroid \bar{x}_k of the k^{th} cluster.

Prove that the above formula is correct.

Solution: It is sufficient to show that

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2$$

for every $k = 1, \dots, K$.

Let's focus on the summand in the sum on the left-hand side of the equation above. We have

$$\begin{aligned} (x_{ij} - x_{i'j})^2 &= (x_{ij} - \bar{x}_{kj} + \bar{x}_{kj} - x_{i'j})^2 \\ &= (x_{ij} - \bar{x}_{kj})^2 + (\bar{x}_{kj} - x_{i'j})^2 + 2(x_{ij} - \bar{x}_{kj})(\bar{x}_{kj} - x_{i'j}) \end{aligned}$$

The first two terms on the right above will - when we sum across all i which is the same as summing across all i' - yield exactly the sum on the right-hand side of the target equality. As for the "cross-terms", we have

$$\sum_{i \in C_k} \sum_{i' \in C_k} (x_{ij} - \bar{x}_{kj})(\bar{x}_{kj} - x_{i'j}) = \sum_{i \in C_k} (x_{ij} - \bar{x}_{kj}) \sum_{i' \in C_k} (\bar{x}_{kj} - x_{i'j}) = 0$$

by the definition of the centroid.

Problem 7.3. (5 points) A K –means clustering algorithm based on squared Euclidean distance with $K = 2$ produced these clusters:

$$I : (0, 1), (1, 2), (2, 1), (3, 2)$$

$$II : (0, 3), (1, 6), (2, 6)$$

What is the value of the objective function, i.e., the function minimized by the clustering algorithm?

Solution: The centroid of I is

$$\left(\frac{0+1+2+3}{4}, \frac{1+2+1+2}{4} \right) = (1.5, 1.5)$$

The centroid of II is

$$\left(\frac{0+1+2}{3}, \frac{3+6+6}{3} \right) = (1, 5)$$

So, the minimum of our objective function is (and I used **R** as a calculator here)

$$2(5 + 1 + 2 + 6) = 28$$

Problem 7.4. (15 points) *Source: MAS-II, Spring 2019.*

You have decided to perform K -means clustering with $K = 2$ on the following data set and have already randomly assigned the clusters as follows:

Observation	x_1	x_2	Initial Cluster
1	5	5	2
2	4	6	2
3	3	0	1
4	5	3	1
5	5	1	2
6	3	6	1
7	2	5	2

- The centroid of the initial cluster 1 is (3.667, 3).
- The centroid of the initial cluster 2 is (4, 4.25).

Calculate the Euclidean distance of Observation 5 from the final centroid of Cluster 2.

Solution: We start with the squared Euclidean distance from the initial centroids.

Observation	Distance from Centroid 1	Distance from Centroid 2
1	5.7769	1.5625
2	9.1109	3.0625
3	9.4449	19.0625
4	1.7769	2.5625
5	5.7769	11.5625
6	9.4449	4.0625
7	6.7789	4.5625

Comparing the distances, we see that Observations 3, 4, and 5 go to Cluster 1 and the remaining observations go to Cluster 2.

Now, we calculate the new centroids. For Cluster 1, we get (4.3333, 1.3333) and for Cluster 2m we get (3.5, 5.5). Here are the updated distances from the centroids.

Observation	Distance from Centroid 1	Distance from Centroid 2
1	13.8889	2.5
2	21.8889	0.5
3	3.5556	30.5
4	3.2222	8.5
5	0.5556	22.5
6	23.5556	0.5
7	18.8889	2.5

We see that there are no reassignments which means that we have reached the final clustering. We can read from the table above that the final answer is $\sqrt{22.5} = 4.7434$.